Next generation sequencing demonstrates minor variant HIV drug resistance mutations

by Randall Fisher

Dissertation presented for the Degree of Doctor of Philosophy in the Faculty of Health Sciences, at Stellenbosch University

Supervisor: Prof. Gert van Zyl

Co-supervisors: Prof. Susan Engelbrecht and Prof. Wolfgang Preiser

March 2016

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own original work, that I am the authorship owner thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Randall Graeme Fisher

March 2016

Abstract

Introduction: The South African public sector antiretroviral therapy (ART) roll-out has been associated with emergence of HIV drug resistance (HIVDR) in therapy-naïve and treated patients. These drug-resistant viral strains are archived in viral reservoirs and may persist as minority variants when outgrown by fitter wild-type strains, in the absence of sufficient drug pressure. Although minority drug resistant variants may predict failure, they are not readily detectable with PCR and Sanger sequencing when constituting less than 20% of the viral population.

Next generation sequencing (NGS) allows the detection of minor drug resistant variants and investigation of linkage of HIVDR mutations when sufficient read-length is obtained. Various NGS platforms, available in South Africa, offer per-sample cost reductions when sufficient numbers of samples are pooled. However, targeted resequencing on NGS platforms requires template enrichment by PCR that can result in PCR-induced error and template re-sampling error. Resampling error is a result of random error or a primer selection bias and is more pronounced when using long fusion primers for NGS amplicon sequencing.

Methods: As data were limited on minor HIVDR variants in infants who became infected despite the Western Cape Prevention of Mother-to-Child HIV Transmission (PMTCT) regimen and in adults with virological failure without major protease inhibitor (PI)-resistance mutations on a PI-based regimen, we performed NGS in these populations.

In the first cross-sectional study we sample minority variants from seven adults who failed PI-based therapy due to poor adherence. Samples were enriched for NGS with a nested fusion primer PCR and minor variants were identified on the 454 Life Sciences, FLX Titanium platform (Connecticut, USA).

In the second cross-sectional study we sampled viral species from 15 PMTCT-failed infants. Nested PCR amplicons were size-fragmented and ligated to platform-specific sequencing adaptors. The sequence library was sequenced in parallel on the Ion Torrent Personal Genome Machine (PGM) (Life Technologies, California, USA) and the Illumina MiSeq (California, USA) and was validated by clonal sequencing.

In study three, we attempted to characterise resampling error when using fusion primers and correct it by partitioning PCR reactions in emulsion. We compared the 454 NGS sequence yields when using “open” or emulsified PCR for template enrichment and investigated the effect of varying degrees of primer-template mismatches when the sampled population consisted of mixtures of plasmids.

Findings: In the first study, NGS improved the drug resistance mutation detection in five of the seven patients although no majority variants with PI-resistance were identified. We concluded that limited or intermittent drug pressure resulted in insufficient selection for major drug-resistant variants to emerge. In study two, NGS conducted on the PGM and MiSeq improved the drug resistance detection in 15 PMTCT-failed infants. Although amplicon sequencing using fusion primers allows the most efficient use of NGS coverage, it is prone to PCR resampling error depending on the degree of template-to-primer mismatches. In study three, emulsion PCR was able to reduce but not correct this resampling error.

Opsomming

Inleiding: Die Suid Afrikaanse publieke sektor antiretrovirale terapie (ART) uitrol gaan gepaard met die tevoorskyning van MIV middelweerstandigheid (MIVMW) in terapie-naïewe en behandelde pasiënte. Hierdie middel-weerstandige virusstamme word in die virus reservoir opgeneem en kan voortbestaan as minderheid-variante wanneer dit oorgroei word deur wilde tipe variante, in die afwesigheid van voldoende geneesmiddel-druk. Alhoewel minderheid-middelweerstandige variante terapie faling mag voorspel, kan dit nie geredelik met PKR en Sanger nukleinsuurbasispaaropeenvolgingbepaling (NSOB) aangetoon word wanneer dit minder as 20% van die viruspopulasie uitmaak nie.

Nuwe generasie NSOB (NGB) laat die aantoning van minderheid middel weerstandige variante toe en die ondersoek van gepaartgaande MIVMW mutasies wanneer die lees-lengte voldoende is. Verskillende NGB platforms, beskikbaar in Suid Afrika, bied kostebesparing per monster wanneer daar voldoende monsters saamgepoel word. Nietemin vereis geteikende NGB die verryking van die templaat deur PKR wat kan lei tot PKR-geïnduseerde ewekansige foute of inleier-seleksie-sydigheid en is meer uitgeproke wanneer lang fusie-inleiers gebruik word vir NGB amplikon analise.

Metodes: Aangesien data beperk is oor minderheids-MIVMW variante in babas wat geïnfekteer is, ten spyte van die Wes-Kaap Voorkoming van Moeder-na-Kind MIV Oordrag (VMNKO) behandelingsplan en in volwassenes met virologiese faling sonder major protease inhibitor (PI)-weerstandigheid mutasies, op ‘n PI-gebaseerde middelkombinasie, het ons NGB in hierdie populasies uitgevoer.

In die eerste deursnit studie het ons monsters geneem van minderheidsvariante van sewe volwassenes wat faling van PI-gebaseerde terapie gehad het weens swak behandelingsgetrouheid. Monsters is verryk vir NGB met ‘n geneste fusie-inleier PKR en minderheidsvariante is geïdentifiseer op die 454 Life Sciences, FLX Titanium platform (Connecticut, VSA).

In die tweede deursnitstudie het ons monsters van virus variante van 15 VMNKO –gefaalde babas geneem. Geneste PKR amplikons is grootte-gefragmenteer en ligeer aan platform-spesifieke NGB-aanpassers. Die NGB biblioteek is in parallel op die Ion Torren Personal Genome Machine (PGM) (Life Technologies, California, USA) en die Illumina MiSeq (California, VSA) getoets en bevestig deur klonale NSOB.

In die derde studie, het ons gepoog om die herproefneming fout wat voorkom wanneer fusie-inleiers gebruik word the karakteriseer en te korrigeer deur PKR reaksies “af te sper” in emulsie. Ons het die 454 NGB opbrengs vergelyk tussen “oop” PKR en geëmulsifiseerde PKR reaksies vir templaat verryking en vir die effek van variërende inleier templaat mispassings wanneer die populasie-steekproef opgemaak word deur plasmiedmengsels.

Bevindings: In die eerste studie het NGB die aantoning van MIVMW mutasies in vyf van sewe pasiënte verbeter, alhoewel geen meerderheid-variante met PI-weerstandigheid geïdentifiseer is nie. Ons het tot die gevolgtrekking gekom dat beperkte of afwisselende middel-druk gelei het tot onvoldoende seleksie om meerderheid middel-weerstandheid variante te laat verskyn. In die tweede studie het NGB, uitgevoer op die PGM en MiSeq, middel weerstandigheid-aantoning in 15 VMNKO –gefaalde babas, verbeter. Alhoewel amplikon NSOB met fusie-inleiers die effektiefste gebruik van NGB dekking bewerkstellig, is dit geneig tot PKR herproefneming foute, afhangend van die graad van templaat-inleier mispassing. In studie drie, was emulsie PKR instaat om hierdie fout te verminder maar nie te korrigeer nie.

Acknowledgements

There are many people without whom my PhD would have not been possible.

To my wife, Farzana, words cannot express how grateful I am for all your love and support throughout my PhD. You've been more than a blessing and I thank you.

To my supervisor, Prof van Zyl, your support and guidance have been invaluable. I thank you for always believing that I could overcome all the challenges that we faced together and for the opportunity to grow into a critical researcher.

To my co-supervisors, Prof Engelbrecht and Prof Preiser, thank you for encouraging me to keep on keeping on. You've inspired me to inspire others.

To my colleagues at the Division of Medical Virology, thank you for your tear-soaked shoulders and the all the loaned reagents. You've made the journey as valuable as the destination.

To all the NHLS staff, particularly Mathilda, Constance and Amanda, thank you for all your technical assistance and your willingness to help.

To my parents, Robin and Vanessa, thank you for encouraging me to pursue my dreams and for setting me on the path. The journey is always smoother when someone has walked it before you.

To my in-laws, Lenny and Sandra, thank you for being an unfaltering support to both Farzana and I throughout our studies. We are blessed to have you in our lives and we thank God for you daily.

To my siblings, Nicole, Nadia, Sean and Stuart, thank you for encouraging me, and for pretending to understand when I explain the "founder effect" and biased sampling of a diverse population.

To my grandma, Margret, thank you for always believing in me and for you daily messages of encouragement. I pray that you will be blessed with a long and prosperous life.

To the National Research Foundation (NRF), the German Research Foundation (DFG) and the Poliomyelitis Research Foundation (PRF), thank you for your financial support.

To my collaborators, Simon Travers, Sergei Pond, Ben Murrell, Konrad Scheffler, David Smith, Ruhan Slabbert, Bronwyn Kirby, Clair Edson, Mark Cotton and Richard Haubrich, thank you for your valuable inputs and expert advice.

To my friends, Dr Govender, the Cummings', the Sickles', the Lillys', the Fraansmans', the Davids' and the Fieldings', thank you for the welcomed distractions and for you words of motivation.

To the University of Stellenbosch, Tygerberg Medical Campus and the Division of Medical Virology, thank you for hosting my research.

Finally, to my saviour and friend Jesus Christ, who carries me through the good times and the bad, whose heart is always moved by my tears, whose arms always pick me up when I fall and who holds my world in the palm of His hands…All praise and glory to you Jesus!

"For He has been counted worthy of more glory than Moses, by just so much as the builder of the house has more honour than the house." *Hebrews 3:3*

Research outputs

Peer reviewed publications:

Fisher, R.G., van Zyl, G. U., Travers, S. A. A., Kosakovsky Pond, S. L., Engelbrecht, S., Murrell, B., Scheffler, K., Smith, D. (2012). Deep Sequencing Reveals Minor Protease Resistance Mutations in Patients Failing a Protease Inhibitor Regimen. *Journal of Virology*, 86(11), 6231–6237. http://doi.org/10.1128/JVI.06541-11 - A summary from thesis chapter 2.

Fisher, R. G., Smith, D., Murrell, B., Slabbert, R., Kirby, B. M., Edson, C., Cotton, M. F., Haubrich, R. H., Kosakovsky Pond, S. L., van Zyl, G. U. (2015). Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure. *Journal of Clinical Virology*, 63, 48-53. http://dx.doi.org/10.1016/j.jcv.2014.11.014 - A summary from thesis chapter 3.

International conference presentations:

Fisher, R.G., Engelbrecht, S., Preiser, W., van Zyl, G. U. (2011). Massive parallel sequencing for cost-effective detection and quantification of HIV-1 minority resistant variants. Oral presentation at the International research training group 1522 annual symposium held in Cape Town, South Africa. – A summary of thesis chapter 1.

Fisher, R.G., van Zyl, G. U., Travers, S. A. A., Kosakovsky Pond, S. L., Engelbrecht, S., Murrell, B., Scheffler, K., Smith, D. (December 2011). Ultra Deep Pyrosequencing reveals minor variant evolution in HIV patients on therapy. Oral presentation at the Virology Africa conference held in Cape Town, South Africa. – A summary of thesis chapter 2.

Fisher, R.G., van Zyl, G. U., Travers, S. A. A., Kosakovsky Pond, S. L., Engelbrecht, S., Murrell, B., Scheffler, K., Smith, D. (June 2012). Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. Oral and poster presentations at the International research training group 1522 annual symposium held in Würzburg, Germany. – A summary of thesis chapter 2.

Fisher, R.G., Engelbrecht, S., Preiser, W., van Zyl, G.U. (February 2013). Template composition and random primer Tm determine RT initiation site. Oral presentation at the International research training group 1522 annual symposium held in Cape Town, South Africa. – A summary of thesis chapter 3.

Fisher, R.G., van Zyl, G. U., Travers, S. A. A., Kosakovsky Pond, S. L., Engelbrecht, S., Murrell, B., Scheffler, K., Smith, D. (April 2014). Ion PGM deep sequencing improves mutation detection in infants who failed PMTCT. Poster presentation at the International conference on infectious diseases held in Cape Town, South Africa. – A summary of thesis chapter 3.

For my loving wife Farzana, for her unwavering support through my many many many post graduate years.

Table of Contents

Title **Page**

List of Figures

Title **Page**

List of Tables

Title **Page**

List of Abbreviations

Abbreviation	Full text
$ or USD	American dollar
%	Percent
.csv	Comma separated format
.sff	standard flowgram format
/r	Low dose of ritonavir
~	Approximately
<	Less than
>	Greater than
∴	Therefore
≤	Less than or equal to
≥	Greater than or equal to
®	Registered trademark
°C	Degrees Celsius
µg	Microgram
µl	Microlitre
µM	Micromolar
µm	Micrometre
1D	One dimensional
3TC	Lamivudine
ABC	Abacavir
ABI	Applied Biosystems (Massachusetts, USA)
ACTG	AIDS clinical trials group
ADP	Adenosine diphosphate
AIDS	Acquired immunodeficiency syndrome
AMV	Avian myeloblastosis virus
ART	Antiretroviral therapy
ARV	Antiretroviral
AS-PCR	Allele-specific PCR
ATP	Adenosine triphosphate
AT:GC	Ratio of adenine and thymine to guanine and cytosine
AZT	Azidothymidine or Zidovudine
bp	Base pairs
BSA	Bovine serum albumin
c/µl	Copies per microlitre
c/ml	Copies per millilitre
CA	Capsid
cART	Combination antiretroviral therapy
CCD	Charge coupled device
cDNA	Complementary DNA
CHAVI	Center for HIV/AIDS vaccine immunology
CI	Confidence interval
CMOS	complementary metal–oxide–semiconductor

Abbreviation	**Full text**
Conc.	Concentration
CPR	Calibrated population resistance
CRT	Cycling reversible terminator
Ct	Cycle threshold
D4T	Stavudine
DDI	Didanosine
ddPCR	Digital droplet PCR
DMSO	Dimethyl sulphoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DRM	Drug resistance mutation
DRT	Drug resistance testing
dsDNA	Double-stranded DNA
DTT	Dithiothreitol
E. coli	Escherichia coli
e.g.	Example
EB	Elution buffer
EDTA	Ethylenediaminetetraacetate
EFV	Efavirenz
em-PCR	Emulsion PCR
et al	And others
ETV	Etravirine
EVG	Elvitegravir
FAU	Fluorescence amplitude units
FDA	Food and Drug Administration
FI	Fusion inhibitor
FTC	Emtricitabine
Fwd	Forward
GA	Genome analyzer
Gb	Gigabases
GS	Genome sequencer
HAART	Highly active antiretroviral therapy
HIV	Human immunodeficiency virus
HIVDR	HIV drug resistance
hrs	Hours
i.e.	Namely
ID	Identification
IMBM	Institute for microbial biotechnology and metagenomics
INSTI	Integrase strand transfer inhibitor
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IQR	Interquartile range
kb	Kilobases
LB media	Luria-Bertani media
LD-PCR	Limiting dilution PCR

Abbreviation	**Full text**
LPV	Lopinavir
LPV/r	Lopinavir boosted with low dose ritonavir
MA	Matrix
Mb	Megabases
mg	Milligram
$MgCl_2$	Magnesium chloride
$MgSO_4$	Magnesium sulfate
MID	Molecular identifier
min	Minutes
ml	Millilitre
$N.F.H_2O$	Nuclease free water
NA	Not available
NC	Nucleocapsid
ng	Nanogram
NGS	Next generation sequencing
NHLS	National Health Laboratory Service (South Africa)
nM	Nanomolar
nm	Nanometres
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleos(t)ide reverse transcriptase inhibitor
NTC	No-template control
NVP	Nevirapine
OH	Hydroxyl
OLA	Oligonucleotide ligation assay
o-PCR	Open PCR
PCR	Polymerase chain reaction
PGM	Personal genome machine
pH	Potential of hydrogen
PHRED	Phil's read editor
PI	Protease inhibitor
pM	Picomolar
PMTCT	Prevention of mother-to-child transmission of HIV
PN-PCR	Pre-nested PCR
qPCR	Quantitative PCR
Q-score	Quality score
RAMICS	Rapid amplicon mapping in codon space
RC	Reverse complement
RCT	Reversible chain terminating
Rev	Reverse
RNA	Ribonucleic acid
RT-PCR	Reverse transcription and PCR
RTV	Ritonavir
SANBI	South African national bioinformatics institute
SBL	Sequencing-by-ligation

Abbreviation	Full text
SBS	Sequencing-by-synthesis
sdNVP	Single dose nevirapine
SDRM	Surveillance drug resistance mutations list
sec	Seconds
SGA	Single genome amplification
SGS	Single genome sequencing
SOC media	Super-optimal broth with catabolite repression media
SOLiD	Sequencing by oligonucleotide ligation and detection
SOP	Standard operating procedure
TAM	Thymidine associated mutation
TDF	Tenofovir
TE buffer	Tris-EDTA buffer
Tm	Melting temperature
™	Trademark
U	Enzyme units
UCSD	University of California, San Diego
USA	United States of America
UV	Ultraviolet
UWC	University of the Western Cape
V	Volts
VF	Virologic failure
VL	Viral load (with reference to HIV-1 RNA load)
Vol.	Volume
WHO	World Health Organisation
xg	Times gravity
X-gal	5-bromo-4-chloro-3-indolyl-beta-D-galacto-pyranoside

1 Introduction

1.1 HIV in South Africa

There are currently 25.8 million people estimated to be infected with human immunodeficiency virus (HIV) in sub-Saharan Africa (World Health Organisation 2015), 6.19 million of which are estimated to be living in South Africa (Statistics South Africa 2014). The Human Sciences Research Council estimates that approximately 5.5% of the population that resides in the Western Cape Province is HIV positive (Shisana et al. 2014). As antiretroviral therapy (ART) is the most effective way to improve patient health by suppressing viral replication and allowing immune recovery, the national government supplies antiretrovirals (ARVs) through a public sector therapy roll-out. Combination antiretroviral therapy has been shown to drastically limit or completely halt viral evolution in patients who are adherent to therapy (Kearney et al. 2014), with the incident risk of therapy failure decreasing with increased duration of therapy (Rosenblum et al. 2009). However, patients who are infected with a viral strain that already harbours HIV drug resistance mutations (transmitted drug resistance) are at a high risk of therapy failure (Cambiano et al. 2013) and in patients with inadequate therapy adherence or pre-existing drug resistance from prior regimen failure, drug resistance could emerge and evolve if therapy continues in the presence of drug resistance (Nachega et al. 2011).

1.2 Anti-retroviral therapy (ART)

ART began with zidovudine (AZT), which was approved by the USA Food and Drug Administration (FDA) in 1987 (Brook 1987; Ezzell 1987). When administered to HIV-infected individuals, clinicians observed a decrease in their blood-HIV load (at that stage measured by a p24 antigenaemia assay) for a limited period before their viral load increased again (Hirsch 1988). These patients then progressed to therapy failure, acquired immunodeficiency syndrome (AIDS) and eventually death (Molina et al. 1994). In the years that followed the FDA approved additional ARVs which were sequentially administered as ART, however, this mono-therapeutic approach did not prevent the onset of AIDS.

ARVs are divided into five broad categories based on the stage of viral replication in which they are active:

- Attachment and fusion inhibitors (FIs) – Either prevent the virus from binding to the target host cell (attachment inhibitors e.g. Maraviroc) or prevent the virus from fusing with the host cell membrane (e.g. Enfuvirtide) (Lalezari et al. 2003).
- Nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs) – Nucleoside reverse transcriptase inhibitors are not phosphorylated and require triphosphorylation, whereas nucleotide reverse transcriptase inhibitors are monophosphorylated and require only two

phosphorylation steps. This class of ARVs prevents the reverse transcription of viral RNA into cDNA by the drug triphosphate competing with the native nucleotide triphosphate and resulting in chain termination (since they lack 3' OH groups). Common examples of NRTIs include zidovudine (AZT), abacavir (ABC), lamivudine (3TC), emtricitabine (FTC), and tenofovir (TDF) (De Clercq 2004).

- Non-nucleoside reverse transcriptase inhibitors (NNRTIs) – NNRTIs inhibit reverse transcription by binding to a pocket next to the active site, thereby changing the conformation of the viral reverse transcriptase enzyme and preventing nucleotide incorporation into the viral cDNA strand. Examples of this class of inhibitor include nevirapine, Efavirenz (EFV), etravirine and rilpivirine (De Clercq 2009).
- Protease inhibitors (PIs) – These inhibitors prevent the viral aspartyl protease enzyme from by binding to the enzyme's active site and competing with the natural substrate. Typical PIs are lopinavir (LPV), ritonavir, indinavir, darunavir and atazanavir (De Clercq 2009).
- Integrase nuclear strand transfer inhibitors (INSTIs) – This class of inhibitors prevent the viral integrase enzyme from incorporating the viral DNA into the host genome. This class includes raltegravir, elvitegravir (EVG) and dolutegravir (De Clercq 2012).

In 1997, two research groups publicised improved therapy outcomes seen when combining two NRTIs with a PI (Hammer et al. 1997; Gulick et al. 1997), while other studies observed similar benefit when two NRTIs were combined with one NNRTI (Williams & De Cock 1996; D'Aquila et al. 1996). This combination antiretroviral therapy (cART) marked the birth of highly active antiretroviral therapy (HAART) which was capable of durable viral suppression (Gulick et al. 1997). cART was quickly implemented clinically and has been used for treatment-naïve (first-line regimens) and treatment-experienced patients (second-line, third-line or salvage regimens).

1.2.1 First-line cART

First-line cART has developed from the initial combinations consisting of two NRTIs and one PI (Gulick et al. 1997) or two NRTIs and one NNRTI (D'Aquila et al. 1996; Montaner et al. 1998). In resource-rich settings, fist-line cART now consists of two NRTIs with either a boosted PI, NNRTI or an INSTI (European AIDS Clinical Society 2014) while two NRTIs combined with one NNRTI are used as first-line therapy in resource-limited settings (World Health Organisation 2013). To improve therapy adherence fixed-dose combination tablets are now available and reduce the pill burden, that is the number of pills taken per day: these include the combination of EFV, TDF and FTC (Nachega et al. 2014) or FTC, TDF, EVG and cobicistat (marketed as STRIBILD) (Manzardo & Gatell 2014). TDF, FTC and EFV is the most widely used first-line regimen in developing countries and this and other NNRTI-based first-line regimens have low genetic-barriers requiring a the few drug resistance mutations for loss of activity.

1.2.2 Second-line cART

Second-line cART, in resource-limited settings, consists of a ritonavir-boosted PI, in combination with two NRTIs (World Health Organisation 2013). The addition of the low dose ritonavir (notated "/r") increases the trough drug level by inhibiting drug metabolism, reducing the risk of failure and drug resistance. As ritonavir-boosted PIs require multiple drug resistance mutations to lose efficacy, the second-line regimen has a high genetic barrier compared to the first-line regimens. Initially, this regimen was poorly tolerated since both lopinavir and didanosine (ARVs commonly used for second-line therapy) had gastrointestinal side effects. The WHO has subsequently recommended the use of AZT and 3TC or TDF and FTC/3TC combined with LPV/r for second-line, due to lower toxicity and better tolerability (World Health Organisation 2013).

1.2.3 The prevention of mother-to-child transmission (PMTCT) in South Africa

Apart from using ARVs for therapy, they can also be used to prevent vertical HIV transmission from mother-to-child since mothers who are virologically suppressed have the lowest risk of vertical transmission (European collective study 2005). AZT given from 14 weeks of gestation, during labour and after birth was shown to reduce transmission by about 70% (Connor et al. 1994). However, this regimen was difficult to implement in developing countries. Also, simplified regimens, based on nevirapine monotherapy showed efficacy (Guay et al. 1999; Dabis et al. 2005). Moreover, early initiation of cART in mothers not only maximally reduces HIV transmission but also provides maternal health benefits and may prevent transmission in subsequent pregnancies should the mothers remain adherent. Therefore in 2013, South Africa adopted the WHO Option B plus treatment plan as the national PMTCT plan (Ton & Frenkel 2013). Option B plus consists of cART initiation in pregnancy in all pregnant HIV-infected women who are not yet receiving cART, irrespective of disease stage or CD4 count, and which is continued uninterrupted thereafter.

1.2.4 Therapy failure

The WHO defines virologic failure as a persisting VL >1000 copies/ml (World Health Organisation 2013). A large proportion of patients failing first-line cART are expected to have HIV drug resistance (Marconi et al. 2008; Wallis et al. 2014; World Health Organisation 2012) due to the low genetic barrier of the regimen. In contrast, few patients failing a second-line ritonavir-boosted PI regimen have any PI resistance mutations (van Zyl et al. 2011; Wallis et al. 2011).

1.2.5 Accumulation of drug resistance mutations

Drug-resistant variants would replace sensitive variants under drug pressure to become the most dominant circulating population. The emergence of multi-drug-resistant viral strains requires a number of interrelated evolutionary mechanisms including a relatively high mutation rate (due to reverse transcriptase lacking proofreading), enrichment of resistant variants through selection pressure (exerted by therapy) and viral recombination events of variants harboring different mutations (zur Wiesch et al. 2011). Different drug-resistant variants could recombine when infecting the same cell, as HIV reverse transcriptase often switches templates. Subsequently, viral strains that accumulate DRMs to a range of ARVs emerge. A typical example of this scenario is when a multi-drug-resistant variant develops and replicates effectively under first-line NNRTI-based cART.

In some cases, the cost of replicating in the presence of a particular drug is reduced replication fitness (Kulkarni et al. 2012). Relative fitness is dependent on the replication environment: Viruses harbouring drug resistance mutations usually replicate less efficiently than the wild-type variant in the absence of drug pressure since they encode enzymes that often have reduced processivity compared to wild-type, but which enable them to replicate better than the wild-type in the presence of ARVs. These drug-resistant strains retain dominance until the selection pressure is removed and dominance is restored to the fitter wild-type strain. However, these resistant variants are ‘archived’ as proviruses in quiescent CD4 cells, acting as viral reservoirs, and may rapidly replace wild-type when the same or similar drugs are recycled (Siliciano & Siliciano 2004). In this way, treatment-experienced patients who have failed multiple regimens accumulate various DRMs which eventually limit their therapy options. In addition, transmission of drug-resistant viral strains limits the therapeutic options for the newly infected individual. These variants decrease the baseline susceptibility (before therapy initiation) in the general population, as observed by various studies investigating ARV resistance in treatment-naïve populations (Hamers et al. 2011; Gupta et al. 2012).

1.3 Challenges in clinical drug resistance testing

In light of the observed DRMs in ARV-naïve population, drug resistance testing (DRT) would become valuable for baseline testing in settings with high levels of transmitted drug resistance, which is associated with high therapy coverage (especially when not accompanied by viral load monitoring) (van Zyl et al. 2014; Phillips et al. 2011). While the costs and reduced accessibility of DRT in resource-limited countries have prevented the WHO from recommending its widespread implementation, DRT in itself has a list of obstacles that reduce its practicability.

1.3.1 Intra-host diversity

As HIV reverse transcriptase lacks proofreading, viral replication is prone to base substitutions at an estimated rate of roughly once per five viral genomes (Hu & Hughes 2012). This, together with a high rate of viral replication, results in a myriad of diverse viral species produced daily. The vast diversity of HIV makes viral genotyping challenging because Sanger sequencing relies on reaction priming with genome-specific primers. Using degenerate primers increases the selection tolerance to include multiple variants, although, too many degenerate bases causes non-specific priming and failed sequencing reactions.

1.3.2 Sensitivity - Subtype B tests and Subtype C prevalence

Commercially available DRT kits are readily available but require appropriately skilled lab technicians and access to diagnostic facilities with sufficient infrastructure to perform and process DNA sequencing. In addition, interpreting the sequencing results to detect HIV drug resistance often requires online drug resistance database access. In resource-limited settings such as sub-Saharan Africa, access to technology to perform drug resistance assays and interpret sequences is limited. Moreover, most commercial DRT assays target are designed for subtype B viral strains (e.g.: Applied Biosystems' ViroSeq® (California, USA) and Visible Genetics' TRUGENE® (Ontario, Canada)) and have reduced sensitivity to subtype C variants that are prevalent in sub-Saharan Africa (Aghokeng et al. 2011). However, if sufficient technical infrastructure is available, published in-house DRT methods could be adapted and implemented, enabling a ~50% reduction in DRT costs ($ 112 as opposed to the commercial $ 300) (Chaturbhuj et al., 2014). These methods could also achieve a high rate of sequencing success through improved primer selection for conserved targets and to match local circulating strains.

1.3.3 Sample turnaround time

Despite reduced costs, in-house resistance genotyping is time consuming and necessitates various "hands-on" and computational analysis steps: nucleic acid extraction, cDNA synthesis, PCR enrichment, gel electrophoresis (to confirm PCR amplification), PCR clean-up, sequencing reactions, automated Sanger DNA sequencing, sequence alignment, contig assembly and bioinformatic drug resistance interpretation. While an automated drug resistance sequence analysis pipeline such as "RECall" could shorten sequence processing (Woods et al. 2012), few interventions are capable of reducing the laboratory processing time requirements without cost implications. Because peripheral clinical care facilities lack the necessary infrastructure to process DRT samples, they are compelled to ship samples to core laboratories where in-house DRT can be performed and could increase turnaround time.

1.3.4 In-house DRT

In-house DRT (or bulk sequencing) methods are more affordable than commercial alternatives, however, unlike commercial assays, there is a lot of inter-laboratory variation in in-house assay design and performance (Parkin et al. 2012; World Health Organisation 2013). These methods require the processes mentioned above (section 1.3.3) and rely heavily on template enrichment through multiple rounds of pre- and nested PCR. In most cases, the primers used in these reactions are designed to amplify the most common strains identified in the region where testing is conducted. Primers are designed by downloading an alignment of many geographically-relevant viral sequences from an online HIV database (such as http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html), investigating nucleotide conservation surrounding the DRT relevant regions, and selecting feasible pre- and nested PCR primers.

The alternative approach is to order and use primers designed by other researchers, which are proven to amplify the required genomic regions. When either approach is successful, the nested PCR products are sequenced using internally-binding sequencing primers and the Sanger dideoxynucleotide sequencing method. Since the above methodology relies heavily on gene-specific primers, these reactions are greatly biased and primer-mismatched or minor variants are underrepresented (Rowley et al. 2010). Nevertheless, in-house DRT by PCR and Sanger sequencing can be used to detect baseline resistance or transmitted resistance in recently infected patients, and acquired drug resistance in patients failing therapy. However, these methods have limited sensitivity as they cannot reliably detect minority, drug-resistant variants below a threshold of 20% (Palmer et al. 2005).

1.3.4.1 Minority, drug-resistant populations

When therapy-naïve individuals are infected with drug-resistant HIV, the transmitted viral population is possibly monophyletic, not containing wild-type variants (Little et al. 2008; Gandhi et al. 2003). Reversion to the more fit, wild-type viral strain is estimated to take between four to ten years during which the resistant population wanes, becoming a minor variant (Little et al. 2008). While these variants are indefinitely archived in viral reservoirs, they may persist undetected by conventional DRT since Sanger genotyping methods. Nevertheless, these variants are clinically important since their presence has been shown to predict therapy failure (Metzner et al. 2009; Jourdain et al. 2010; Paredes et al. 2010; Li et al. 2012; Li et al. 2013).

1.3.4.2 Minor variant detection solutions

The solutions presented below are practical alternatives to conventional PCR and Sanger DRT, aimed at detection drug-resistant, minor variants.

1.3.4.2.1 Allele-specific PCR (AS-PCR)

During a standard PCR, the polymerase encounters a primed genomic region and proceeds to complement the single-stranded genomic region that follows (Kleppe et al. 1971). However, if the 3' end of the primer is unbound to the single-stranded template, primer extension would be inefficient. As such, AS-PCR relies on the inability of *Taq* polymerase to extend primers with un-bound 3' ends, where the mismatched base can discriminate between a DRM and the native nucleotide (Metzner et al. 2003). The sensitivity of a real-time PCR assay coupled with the AS-PCR methodology, results in an affordable DRT alternative (Bergroth et al. 2005; Metzner et al. 2005; Paredes et al. 2007). In contrast to Sanger genotyping however, AS-PCR may be more sensitive but is limited in the number of mutations that can be simultaneously sampled and thus limits mutation linkage studies. Similar to Sanger genotyping, AS-PCR relies on primer to template binding and is thus prone to false negative or positive results due to mismatches apart from the putative resistant and wild-type alleles, that the assays were designed to discriminate between (Rowley et al. 2010).

1.3.4.2.2 Oligonucleotide ligation assay (OLA)

OLAs discriminates between wild-type and mutant alleles using two, 20-mer DNA probes: one with an appended biotin and another with a hapten that corresponds to either the wild-type or mutant allele (Tobe et al. 1996). For a successful ligation reaction, both probes must perfectly complement the template (i.e.: there must be no mismatches in both probes binding areas) and the probes must bind directly adjacent to each other in the 5' to 3' direction (i.e.: with no intervening nucleotides). Any mismatches at the ligation site or mispriming would prevent ligation insomuch that a mutant-specific hapten-appended probe could not be ligated to the biotin-attenuated probe when both probes are hybridised to a wild-type template. This hybridisation stringency confers allelic specificity to OLA. Ligated probes are captured by their attenuated biotin moiety on a streptavidin coated well and the hapten is identified using an enzyme-conjugated antibody. The identity of the hapten in turn the identifies the template allele as either the wild-type or mutant (Tobe et al. 1996; Edelstein et al. 1998; Jourdain et al. 2010). Similar to AS-PCR however, OLA too is limited in the number of mutations that can be assayed in parallel and thus limits mutation linkage study. In addition, OLA may return false negative results if there are additional mutations surrounding the DRM allele being interrogated.

1.3.5 Pooled viral load testing before DRT can reduce cost

HIV DRT usually follows a detectable viral load. When the prevalence of virological failure is low, viral load testing of pooled samples could reduce diagnostic costs as most pools would be negative; with only a few positive pools requiring deconvolution to determine which of the individual samples making up the pool are positive (having a viral load above a failure threshold) (Smith et al. 2009). Pooled testing of dried plasma spots would allow sample transport from peripheral sites, and has been shown to be efficient when a threshold for therapy failure of 1000 HIV-1 RNA copies/sample is used, and could reduce reagent costs by more than half (van Zyl et al. 2011).

A novel approach that allows the identification of patients who are failing and distinguishes patients with drug resistance from those without (likely due to inadequate adherence) was proposed by Newman and colleagues (Newman et al. 2014). In short, samples are pooled and a qualitative PCR is performed. If a pool tests positive (indicating that one or more of the comprising members have virological failure) it is followed by PCR of the individual members making up a positive pool. Positive individual PCR products then undergo Sanger sequencing to detect HIV drug resistance. Negative pools or individual PCR products exclude failure and HIV drug resistance. When the expected failure rate is low, this approach combines the efficiency of pooled testing to qualitatively detect failure with in-house HIV DRT by Sanger sequencing.

1.4 Next generation sequencing (NGS)

NGS is also referred to as second generation sequencing, ultra-deep sequencing and massively parallel sequencing due to various characteristics of the sequencing methods:

- Second generation sequencing relates to these platforms having followed the first generation Sanger sequencing platforms. Second generation sequencing is a more accurate description than NGS since third generation sequencing is already commercially available;
- Massive parallel sequencing referrers to the vast number of DNA molecules that are sequenced in parallel.
- Ultra-deep sequencing is a particular application where the high coverage achieved by massive parallel sequencing is focused on a short genome target to achieve high-resolution sequencing of rare variants from biologic samples.

Second generation sequencing (referred to here as NGS) offers a viable alternative to first generation DRT methods. Because NGS identifies genomic sequences rather than differentiating single alleles, it allows for DRM linkage study that is not possible with AS-PCR and OLA, provided that the linked mutations are identified on the same DNA strand. The use of genetic barcodes or molecular identifiers (MIDs) enables pooled testing and cost savings (Rohland & Reich 2012), while the redundant sampling enables minor variant detection (Wang et al. 2007).

Various sequencing library generation options are available for NGS, however, they all require sequencing adaptors to capture the target and prime the sequencing reactions. Depending on the enrichment strategy employed, the MID sequences may be added as part of a fusion primer (as is the case of amplicon sequencing) or ligated to the sequence of interest in a fragmented library generation approach. NGS platforms require sample enrichment via bead-based emulsion PCR (Diehl et al. 2006) or bridge PCR amplification on a solid surface (Shendure et al. 2005), that is made possible by the flanking adaptor sequences. If a pooled sample testing approach is used, all the indexed samples are pooled at the same final concentration before the adaptor sequence is ligated to its bead- or solid-surface-anchored complement.

NGS platforms available in South Africa use one of three approaches to detection nucleotide incorporation, including:

- Sequencing-by-synthesis (SBS) – This approach directly detects incorporated nucleotides through either of the following methods:
    - By detecting the particular nucleotide that has been incorporated using reversible chain terminating (RCT) nucleotides linked to fluorophores (Ju et al. 2006). Solexa/Illumina's (California, USA) NGS platforms, the MiSeq and HiSeq, use this RCT approach;
    - By detecting the release of a DNA polymerisation by-product: either H+ ions, as a change in the reaction pH (Rothberg et al. 2011), or a pyrophosphate, as used in pyrosequencing. Pyrosequencing monitors the production of pyrophosphate that indirectly results in the catalysis of luciferin, producing chemiluminescence (Ahmadian et al. 2000). The emitted light is detected using a high-resolution charge coupled device (CCD) camera and the luminescence intensity amplitude is translated into the number of incorporated nucleotides (Morey et al. 2013). The Life Technologies' (California, USA) Ion Torrent Personal Genome Machine™ (PGM) is the only NGS platform that uses pH changes to detect nucleotide incorporation, while the 454 Life Sciences (Connecticut, USA) platforms, the Genome Sequencer (GS) Junior and the FLX Titanium, are the only pyrosequencing NGS platforms (now distributed by Roche Diagnostics, Basel, Switzerland).
- Sequencing-by-ligation (SBL) – This sequencing chemistry detects the appended fluorophore on the end of eight-base-long probes. Each probe interrogates two bases. During probe binding the colour of the appended fluorophore is recorded. Thus, this method employs a two-base interrogation strategy and requires that a base is dually

interrogated to determine its identity (Shendure et al. 2005). Currently, Agencourt's (Massachusetts, USA) SOLiD™ sequencer is the only NGS platform to use this novel SBL sequencing chemistry (now distributed by Applied Biosystems, Massachusetts, USA).

1.4.1 Ultra-deep and ultra-wide sequencing

Various options exist for upstream HIV NGS sample preparation although amplicon sequencing or targeted resequencing offers optimal use of NGS coverage for DRT. Since most primary DRMs related to first and second-line cART failure are found in the viral *protease* and *reverse transcriptase* genes, enriching these regions prior to sequencing allows for sufficient sampling depth to identify minority populations of drug-resistant variants. Alternatively, if minor variant detection is not required, sequencing with multiple MIDs enables researchers to pool and genotype many samples in parallel, at sufficient depth to accurately identify drug resistance. Pooled sequencing of many indexed samples with "shallow coverage" is referred to as ultra-wide sequencing (Dudley et al. 2012) and is currently limited by the number of available specific MIDs and the platform-specific minimum sequencing depth needed to confidently call a specific base.

1.4.1.1 Advantages and disadvantages

Ultra-deep sequencing can identify minor variants with >99% confidence in nucleotides identity. This "base-calling confidence" and is a function of the platform-specific base-call signal quality. In addition, the number of times that specific base was accurately identified improves the average base-call quality and in turn, the base-call confidence. Coverage depth can increase the confidence in minor variant calling provided that the sequenced library is representative of the biologic diversity in the patient. However, this cannot be assumed: despite acceptable accuracy of current NGS platforms, upstream enrichment or library preparation could introduce mutations or misrepresentation of variant frequency compared to the true population diversity in the individual from whom the sample was collected.

When many MIDs are used to sequence multiple samples in parallel, the per-sample costs are greatly reduced although, each sequencing platforms has a limit to the amount of per-run sequence data produced, which is divided between the number MIDs represented in the sequencing reaction. Thus, the more MIDs used, the lower the per-sample coverage and the lower the average base-call confidence.

Another consideration is practical batch sizes; with high coverage platforms such as Illumina Miseq the pooling of up to 96 samples greatly reduces costs. However, these large batch sizes would practically result in delays in turnaround time in lower throughput laboratories, as it may take many days to fill a batch.

1.4.2 Various platforms

There are currently four NGS technologies available in South Africa, each with their advantages and disadvantages (Hadfield & Loman 2014). Listed below are the various platforms, a brief description of their origins, a summary of their sequencing chemistries, and their specific strengths and weaknesses. As previously mentioned (see section 1.4 above), second generation sequencing platforms require specific sequences on the ends of the sequence amplicon to anchor the DNA for enrichment prior to sequencing.

1.4.2.1 454 Life Sciences

The 454 Life Sciences (Connecticut, USA) GS20 (distributed by Roche Diagnostics (Basel, Switzerland) since 2007), was the first commercially available second generation sequencing platform, introduced in 2005 (Margulies et al. 2005). The GS20 had an introductory read length of between 100 to 150 bp and generated up to 20 megabases (Mb) of sequence data per run (Margulies et al. 2005). The pyrosequencing technology used on 454 platforms (454 Life Sciences, Connecticut, USA) is based on the research conducted by Mostafa Ronaghi and Pål Nyrén at the Royal Institute of Technology (Stockholm, Sweden) (Ronaghi et al. 1996), and relies on the detection of pyrophosphates that are released after a polymerase incorporates a nucleotide (Margulies et al. 2005). The sequencing chemistry involves the steps below and is depicted in figure 1.8.1 in Appendix A:

- After the enrichment with a bead-based emulsion PCR, single-stranded, bead-bound amplicons are immobilised in a 29 µm well located on a picotitre plate, in a flow-cell. The sequencing reaction is primed in the presence of a polymerase, an adenosine tri-phosphate (ATP) sulfurylase, and a luciferase.
- Single nucleotides are washed over the flow-cell in a predetermined sequence. If the specific nucleotide that is washed over the picotitre plate is not incorporated, it is digested by an apyrase enzyme in the subsequent washing. However, when the polymerase incorporates a nucleotide, a pyrophosphate and a hydrogen ion are released.
- ATP-sulfurylase uses the free pyrophosphate to convert adenosine di-phosphate to ATP, which in turn activates the luciferase. The luciferase catalyses the substrate luciferin, releasing a visual light signal before an apyrase wash degrades unincorporated nucleotides.
- The amplitude of the emitted light is proportional to the number of nucleotides incorporated such that the luminescence produced from two incorporated nucleotides is double that observed when one nucleotide is incorporated.

Two 454 pyrosequencing platforms are currently available in South Africa: the 454 GS FLX and the smaller bench top version, the 454 GS Junior (454 Life Sciences, Connecticut, USA). While there are minor differences in these platforms adaptor sequences, the most noticeable difference is the sequence data capacity (i.e.: 450 Mb on the FLX and 35 Mb on the Junior) and the platform-specific read lengths (450 bp on the FLX and 400 bp on the Junior) (http://454.com/products/index.asp).

The greatest advantages of the 454 sequencing platforms are the long read length and the speed at which a sequencing run is completed. However, 454 platforms struggle with accurately sequencing homopolymer regions since there is not much difference in the observed luminescence when more than six bases are incorporated at once (Huse et al. 2007). In addition, the running costs for the 454 platforms are considerably more expensive than their competitors (Loman et al. 2012). In October of 2013, Roche announced that it would no longer support the 454 sequencing platforms after mid 2016 as the platform was being overshadowed by the newer sequencing technologies (https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business).

1.4.2.2 Solexa/Illumina

Solexa (California, USA) released their first NGS platform, the Genome Analyzer (GA) in 2006, before being acquired by Illumina (California, USA) in 2007 (http://www.illumina.com/). The GA had an initial read-length of 50 bp and generated 1 gigabase (Gb) of sequence data per run, that was later improved in 2007 to 150 bp read length and 50 Gb per sequencing run. Solexa (California, USA) uses an SBS approach and the sequencing chemistry first described by Shankar Balasubramanian and David Klenerman from Cambridge University (Cambridge, United Kingdom) (Balasubramanian et al. 2004). Illumina platforms employ the reversibly-terminating fluorescent labelled nucleotides sequencing strategy as described below and depicted in figure 1.8.2 in Appendix A:

- Sequence amplicons are enriched on a solid surface by bridge PCR to form amplicon colonies before a primer and a polymerase are added in solution.
- One of the four different reversibly-terminating fluorescently labelled nucleotides is incorporated into the complementary DNA strand. Once incorporated, a scanning laser determines the identity of this nucleotide by the colour of its fluorescent tag.
- A chemical wash is applied to cleave off the 3’ non-extendable part of the incorporated nucleotide, along with the fluorescent tag.

- Next, more reversibly-terminating nucleotides are washed over the reaction and the cycle is repeated. Based on the steps involved in this sequencing strategy, it is often referred to as cycling reversible-terminator or CRT sequencing.

Solexa/Illumina platforms (California, USA) currently available in South Africa include the MiSeq and the HiSeq which produce 15 Gb and 1500 Gb sequence data per run, at read lengths of 600 and 300 bp respectively (http://www.illumina.com/systems/sequencing-platform-comparison.html?mode=iframe). The greatest advantage of the Solexa/Illumina platforms (California, USA) is their high sequencing throughput (generating many gigabases of data) while their greatest disadvantage is the short read lengths in relation to their competitors, however, through bidirectional sequencing, these platforms can achieve 600 bp read lengths.

1.4.2.3 Life Technologies

Life Technologies (California, USA) Ion Torrent Personal Genome Machine™ (PGM) was first introduced to the commercial market in February of 2010 (Rusk 2011). Similar to the 454 sequencing platforms, the PGM detects bi-products of a DNA polymerisation. However, the PGM (Life Technologies, California, USA) differs from the above-mentioned NGS platforms in that it does not use optics to detect base incorporation signals, which allows a much smaller instrument footprint. Rather, the PGM (Life Technologies, California, USA) detects minute changes in the hydrogen concentration (pH) of the sequencing reaction taking place on a complementary metal–oxide–semiconductor (CMOS) chip (Rusk 2011). The CMOS chips used on the PGM (Life Technologies, California, USA) differ in the number and density of embedded reaction wells and in turn, the amount of sequence data output, insomuch that three chips are available for the PGM (Life Technologies, California, USA): the 314 chip, the 316 chip and the 318 chip that produce up to 100 mb, 1 Gb and 2 Gb of sequence data, respectively (https://tools.lifetechnologies.com/content/sfs/brochures/PGM-Specification-Sheet.pdf).

Initial PGM sequencing chemistries produced 50 bp read lengths and generated 100 Mb of sequence data. In 2011, these characteristics improved to 400 bp reads and more than 2 Gb of sequence data on a 318 sequence chip. The PGM (Life Technologies, California, USA) measures the hydrogen ions (protons) produced during complementary strand synthesis of bead-bound, emulsion-PCR-enriched amplicons. The sequencing strategy is depicted in figure 1.8.3 (in Appendix A) and uses the following steps:

- The bead-bound amplicon is primed and a single-nucleotide-species wash is flushed over the sequencing chip in a predetermined sequence.

- If the base is not incorporated into the complementary DNA strand, it is discarded in the next wash step. However, if the nucleotide is incorporated by the polymerase, a pyrophosphate and a hydrogen ion are released as bi-products of DNA polymerisation.
- The released hydrogen ion results in a localised change in the reaction pH, which is detected by the CMOS circuitry embedded beneath the sequence well. The amplitude of the detected pH change is proportional to the number of hydrogen ions released, which in turn is proportional to the number of nucleotides incorporated.
- The chip is subsequently washed, and the next predetermined nucleotide wash is applied.

The PGM is currently the only semiconductor sequencing platform commercially available in South Africa through the Stellenbosch University Central Analytical Facility (Western Cape Province, South Africa) or the Division of Biochemistry at North-West University (North West Province, South Africa). The per-base sequencing cost of the PGM is by far its biggest advantage accompanied by its quick turnaround time of roughly two hours (Loman et al. 2012). On the other hand, similar to the 454 platforms (454 Life Sciences, Connecticut, USA), the PGM (Life Technologies, California, USA) has difficulty sequencing homopolymers (Metzker 2005) as the pH change, measured as voltage amplitude, loses linearity with the number of bases incorporated for runs of six or more identical bases.

1.4.2.4 Agencourt/Applied Biosystems

The Sequencing by Oligo Ligation Detection™ (SOLiD) sequencing platform was introduced to the commercial market by Agencourt (Massachusetts, USA) in 2006 before the company was acquired by Applied Biosystems (Massachusetts, USA). At introduction, the SOLiD sequencer produced 35 bp reads and 3 Gb of sequence data per run, with greater than 99.85% accuracy owing to its SBL technology (Liu et al. 2012). The read length and total data output were soon improved to 85bp and 30 Gb per run with the introduction of the SOLiD 5500xl in 2007 (http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html). After the bead-based emulsion PCR, the novel SBL strategy depicted in figure 1.8.4 (in Appendix A), is applied as follows:

- Enriched beads with bound single-stranded amplicons are immobilised on the surface of the flow-cell. An eight-base probe with a 5’ fluorescent dye is ligated to the amplicon. The colour of the fluorescent dye is visually recorded.
- Next, the dye is cleaved off from the probe along with the last three probe bases. The cleaved dye is washed away, and the next probe is ligated.
- Since the first two bases of the probes are known and correspond to a specific fluorescent dye colour, this sequencing method interrogates two-bases-at-a-time.

- Once the amplicon has been sequenced to the end, the complementary strand is washed away and a new primer (one base shorter than the first round primer) is used to again primer the sequencing reaction. This sequencing round once again interrogates the template two-bases-at-a-time, starting from a -1 base position.
- This cycle is repeated five times, each time the primer length is decreased by one additional base and the two-base interrogation is offset by -2, -3 and -4 nucleotides from the original starting position.

One of the greatest advantages of this sequencing approach is that, since nucleotides are essentially interrogated multiple times, the error rate of this platform is exceptionally low (Liu et al. 2012). Although, due to the short read lengths characteristic of this platform, investigation of mutation linkage is limited and post-sequencing read mapping requires powerful computational hardware.

1.4.3 Homopolymer error

Of the sequencing platforms described above, two are particularly prone to read length error in homopolymer regions. These errors result from the PGM (Life Technologies, California, USA) and 454's (454 Life Sciences, Connecticut, USA) inability to discriminate more than five of the same nucleotide species incorporated in one nucleotide wash. Incorrect homopolymer read length errors present as insertion or deletions, collectively referred to as indels. On the 454 platform (454 Life Sciences, Connecticut, USA), when seven or more nucleotides are incorporated, the luciferase luminescence is incrementally more than when six nucleotides are incorporated, and the charge coupled device (CCD) camera cannot accurately distinguish the difference in luminescence amplitude (Margulies et al. 2005). Similarly on the PGM (Life Technologies, California, USA), the incremental difference between six or more hydrogen ions (released during base incorporation) causes diminutive changes in the reaction pH, which the embedded sensors struggle to identify (Bragg et al. 2013).

Sequencing chemistry on the other SOLiD (Agencourt, Massachusetts, USA) and Solexa/Illumina platforms (California, USA) circumvent homopolymer errors using the following strategies:

- The SOLiD's SBL protocol (Agencourt, Massachusetts, USA) can only incorporate one probe at a time (per sequenced fragment) since the 5' ligation-terminator (attached to the probe) must be cleaved before the ligase can incorporate the next probe (see figure 1.8.4 in Appendix A). If two probes were incorporated simultaneously, the scanning laser would detect two fluorophores with an abnormal emission wavelength.
- CRT sequencing used on the Solexa/Illumina (California, USA) platform terminates sequencing after every nucleotide incorporation, preventing more than one nucleotide

incorporation at a time. Once the terminator is chemically cleaved, the next nucleotide can be added. This one-nucleotide-at-a-time sequencing strategy produces accurate homopolymer read length calls (see figure 1.8.2 in Appendix A).

1.4.4 Comparative NGS

As the first commercially available NGS platform, Roche 454 sequencing has been the preferred technology used to identify drug-resistant HIV minority populations, despite the homopolymer read-length limitations of its' SBS chemistry. Recently, Li and colleagues demonstrated the affordability and accuracy of Illumina's CRT sequencing homopolymer reading accuracy when compared to the 454 technology (Li et al. 2014). These favorable characteristics and the 600bp bi-directional, read lengths (http://www.illumina.com/systems/sequencing-platform-comparison.html?mode=iframe) paint the Illumina sequencers as the preferred drug resistance NGS platforms in the future.

1.4.5 Application to HIV DRT and pooling potential

Homopolymer reading errors are of particular concern when resequencing HIV since 17 drug resistance mutations (DRMs) are known to occur in homopolymer regions (Dudley et al. 2012). Moreover, while the SOLiD (Agencourt, Massachusetts, USA) and MiSeq (Solexa/Illumina, California, USA) platforms are immune to homopolymer errors, the SOLiDs (Agencourt, Massachusetts, USA) read lengths do not facilitate mutation linkage studies and the MiSeq (Solexa/Illumina, California, USA) sequencing runs are costly.

To compensate for the PGM (Life Technologies, California, USA) and 454's (454 Life Sciences, Connecticut, USA) homopolymer errors, bioinformatic software uses logistically corrective mapping algorithms. By translating sequence reads into amino acids and mapping them to a consensus codon sequence, the software identifies indels as codon frame shifts. Once identified, homopolymer indels are corrected such that the codon sequence is in frame with the consensus, and the read is realigned. Two codon-aware sequence alignment programs are used for HIV mapping in this thesis:

- Datamonkey was developed by Sergei Kosakovsky Pond in 2005 and is available online at http://www.datamonkey.org (Pond & Frost 2005). This software was used in chapter three to align PGM and MiSeq HIV data.
- In chapter four we used RAMICS, developed by Simon Travers and Imogen Wright in 2014 to map 454 GS Junior sequences. This software package is accessible online at http://hiv.sanbi.ac.za/tools/#/ramics (Wright & Travers 2014).

1.4.5.1 Targeted resequencing

NGS platforms can sequence entire genomes depending on the sequence library preparation methods employed. For example, if viral RNA is isolated from patient material, the entire viral genome can be reverse transcribed, PCR amplified with random or specific primers and sequenced using a DNA fragmentation and sequencing adaptor ligation library preparation protocol. While this approach generates useful sequence data on recombination and evolution, since most first and second-line cART DRMs occur in *protease* and *reverse transcriptase*, roughly 20% of genes sequenced are relevant to drug resistance.

Targeted resequencing of these “drug-resistance-related” areas (i.e.: *protease* and *reverse transcriptase*) is achieved through PCR enrichment of these specific areas. This enrichment strategy optimally uses the available sequence coverage to improve the sampling depth and in turn, minor variant detection, and can be achieved by two sequencing strategies:

- Amplicon sequencing with fusion primers – Fusion primers contain 5’ platform-specific adaptor sequences and may include patient-specific MIDs, while the 3’ gene-specific sequence facilitates targeted enrichment of drug-resistance-related areas. Although the final amplicon should not exceed the maximum platform-specific sequencing read length, determined by the sequencing strategy (either unidirectional of bidirectional). To improve mutation linkage study, amplicons may be generated with overlapping ends, as seen in thesis chapter two. While effective, the use of fusion primers has been linked to strong primer-induced selection bias, as described in thesis chapter four.
- PCR enrichment followed by fragmentation and adaptor ligation - Enriching drug-resistance-related areas with primers that bind in conserved genomic regions reduces the primer-induced selection bias. The selection bias is reduced because most conserved-area primers should not discriminate between various viral species and should reduce the effects of PCR resampling. Considering these factors, enriching the viral *protease* and *reverse transcriptase* with conserved primers, reduces selection bias. Following this enrichment with amplicon fragmentation and adaptor ligation, results in NGS-platform-sequenceable amplicons, generated with reduced selection bias and sequenced with an optimal use of coverage. In addition, the amplicons generated before adaptor ligation are sequenceable on multiple NGS platforms, provided that the appropriate adaptor is appended. This adaptor ligation strategy was used in thesis chapter three to validate the PGM (Life Technologies, California, USA) for HIV DRT by sequencing the same sequence library on an alternative NGS platform, the Solexa/Illumina MiSeq.

1.4.5.2 Coverage and depth

Confident identification of a minor variant mutation in a sequenced library is dependent on coverage and sequence accuracy, provided that the sequenced library is representative of the biologic diversity in the patient. Therefore, a higher coverage is associated with higher base-call accuracy (Meynert et al. 2013). However, the nucleic acid enrichment process, prior to sequencing, could introduce errors either since the patient sample is inadequate to represent patient diversity (e.g. small volume and low sample HIV RNA load, resulting in a high random sampling error) or the reverse transcription and PCR amplification process could introduce errors or biased amplification of particular variants. These sampling or enrichment errors could not be overcome by increased coverage but could be improved by adequate sampling (increased sample volumes in patients with low viral loads) and enrichment processes with lower error and bias (discussed in more detail below).

When pooling many samples for NGS there are many considerations: The data limit of the sequencing platform, the sequencing platform-specific error rate (how often a base is incorrectly identified), the platform read-length (and whether mutation linkage is important for the investigator), any AT:GC bias, the target amplicon size, and the frequency (percentage) of minor variants that one aims to detect. For instance, with high coverage platforms such as the Illumina HiSeq (California, USA) many more samples can be pooled than with 454. However, the relative shorter read-length may make the Illumina platform less suitable for the study of mutation linkage across long genome targets compared to the new Pacific Biosciences (California, USA) RS II sequencing platform, which is able to sequence up to 15 kb (http://investor.pacificbiosciences.com/releasedetail.cfm?ReleaseID=876252).

1.5 NGS implementation challenges

In resource-limited countries, the greatest deterrent to NGS is the costs associated with sample preparation and sequencing. These costs include high fidelity PCR reagents, parallel reactions, fusion primer costs, commercial sample extraction and reaction purification kits, library preparation kits (including ligation of sequencing adaptors) for sequencing platforms and labour costs. In addition, the resultant sequence data analysis is complex and requires some bioinformatics or programming knowledge to translate sequence data into clinical results.

As previously described, the per sample cost may be reduced by using multiple MIDs to pool samples together on NGS platforms that provide sufficient sequencing depth to accurately sequence all variants with therapeutic relevance. This approach would, however, require seasoned laboratory personnel and sufficient sample volumes to adequately represent the variation in the patient, and to reduce random PCR sampling error, parallel processing. In addition, sequencing libraries and reactions may take a few days to complete, resulting in patients being lost to follow-up.

1.5.1 Sample preparation

The use of high fidelity PCR is advised for upstream NGS to reduce homopolymer transcription error and spurious base substitutions which may present as minor variant populations after sequencing (McInerney et al. 2014). These enzyme systems, with their proofreading ability, improve assay fidelity at the expense of assay sensitivity and processivity. Additionally, for small genome targets, such as for HIV drug resistance testing, the most efficient use of NGS coverage requires an amplicon sequencing approach which employs fusion primer PCR, but which is associated with a strong selection bias and template resampling (Berry et al. 2011). Nevertheless, the selection bias and processivity issues can be partially addressed by parallel PCRs for each patient, since limiting dilution PCR and single genome amplification are not feasible for routine diagnostics.

1.5.2 Additional precautions prior to sequencing

1.5.2.1 DNA degradation by ultraviolet light

Exposing DNA to ultraviolet (UV) light can cause degradation and point mutations that appear as minor variants after sequencing (Hori et al. 2007). To avoid this problem, when separating sequence amplicons in agarose, it is advisable to use a nucleic acid stain that allows DNA to be visualised without UV exposure. Typical examples of applicable gel stains are GR Green (Labgene, Châtel-Saint-Denis, Switzerland) and GelStar (Lonza, Basel, Switzerland), which permit DNA to be visualised over blue light with an orange filter.

1.5.2.2 No-template control validation

When using “high-resolution” sequencing with the ability to sample minor variants, low copy number reaction contamination may appear as minor variant populations if they are enriched through the fusion primer PCR (i.e.: they contain sequencing adaptors and MIDs). Screening for low-level contamination requires that a sensitive nested PCR be conducted on all no-template controls (NTCs) even when no bands are visible in the NTC after the fusion primer PCR. This nested reaction enables the sensitive detection of contamination that warrants remedial decontamination and repeating fusion primer PCRs.

1.5.2.3 Sequencing-by-synthesis (SBS) template validation

When PCR products undergo manipulation such as gel extraction it is prudent to confirm that DNA remained intact and amplifiable, by performing a screening nested PCR prior to sequencing on a next generation platform.

1.5.2.4 Sampling error and bias

Amplicon generation using fusion primers may introduce primer-induced selection bias and increase PCR template resampling. Once the template-specific 3’ end of the fusion primer binds to its genomic target, DNA polymerisation begins, and the fusion primer is incorporated into the complementary DNA strand. In the subsequent PCR rounds, this template becomes the perfect match for the same species of fusion primer, and this template will be resampled. This occurrence is referred to as PCR resampling (Liu et al. 1996). After enrichment mismatched variants will be under-represented as a result of biased priming of perfectly matched variants.

Random sampling error may misrepresent the original sample population when only a few variants are sampled from a patient with a low viral load or when PCR randomly primes a particular template. When primers randomly prime particular templates earlier than others, and due to the exponential nature of PCR, these early-selected variants will be over-represented in relation subsequently selected ones. This type of selection bias is also called “primer resampling error” (Liu et al. 1996). Various PCR enrichment methods are available to circumvent the PCR resampling error however limiting dilution PCR (LD-PCR) is currently the only means of eliminating it.

1.5.2.4.1 Limiting dilution (LD)-PCR

Currently, LD-PCR is the only approach that obviates the PCR resampling error and is used for single genome sequencing (Ramachandran et al. 2008; Salgado et al. 2010). LD-PCR requires quantifying starting cDNA and diluting the sample to one cDNA copy per pre-nested PCR. At this resolution, the PCR resampling error is eliminated since there is no competition between templates for primer hybridisation. However, LD-PCR is laborious, not feasible for diagnostic implementation and expensive to perform when considering the number of PCRs required.

1.5.2.4.2 Indexing cDNA with a unique ID tag

The primer ID strategy (described in chapter 4, section 4.1.6) involves synthesising cDNA with a gene-specific primer that contains a 5’ PCR tag and an ID of random bases (eight or ten bases depending on the assay version) (Jabara et al. 2011; Boltz et al. 2015). Each cDNA strand thus contains a unique ID sequence that is later incorporated into the sequenced amplicon. Post-NGS, the ID tag is used to determine the origin of each sequence read. Because the IDs are unique, reads with the same ID are aligned, and the consensus corrects for spurious PCR error and PCR resampling (Jabara et al. 2011; Boltz et al. 2015). While this enrichment strategy provides a reference point for accurate minor variant identification, it has been linked to biased sampling and insensitivity (Zhou et al. 2015; Keys et al. 2015; Brodin et al. 2015).

1.5.2.4.3 Emulsion PCR (em-PCR)

Emulsification of an aqueous PCR in oil essentially mimics the LD-PCR on a micro- to atto-litre scale (fully described in thesis chapter 4, section 4.1.8). Provided that template concentrations are at least three-fold less than the number of emulsion droplets produced, each reaction droplet should rarely contain more than one template. Thus, primer binding completion is eliminated along with the PCR founder effect.

To employ em-PCR for upstream NGS sample preparation, high fidelity PCR is a minimum requirement. A commercial high fidelity em-PCR kit is available from Roche Diagnostics (Basel, Switzerland) however this kit is part of their NGS amplicon library preparation protocol and is designed to enrich bead-bound amplicons in an em-PCR. Another high fidelity commercial em-PCR platform is available from RainDance Technologies (Massachusetts, USA), although this technology is not yet available in South Africa. In thesis chapter 4, we attempt an in-house high fidelity em-PCR to correct fusion primer sampling error.

1.5.3 Bioinformatics

Parallel to sequence platform development has been the improvement of software, capable of processing NGS data, since these platforms produce large sequence datasets (Li & Homer 2010). These software packages make use of read alignment algorithms to align sequence reads to a consensus sequence in a process called “mapping”. Post-sequencing data processing basically requires the exclusion of poor quality bases, read mapping to a consensus sequence, and minority population identification.

Base-call quality relies on the platform-specific confidence in a nucleotide being accurately identified. For NGS platforms like the MiSeq (Solexa/Illumina, California, USA) and the 454 FLX (454 Life Sciences, Connecticut, USA) that visually detect base incorporation, the quality score is partially based on the amplitude and wavelength of emitted fluorescence or observed luminescence. On the PGM however, these factors are substituted for pH change measured as voltage amplitudes since the PGM does not detect visual signals. Base-call signals are converted into PHRED Q-scores using platform-specific base-calling algorithms. The PHRED scoring system is a logarithmic accuracy measurement scale wherein a Q-score of 10 equals a 90% accurate base-call probability, a Q-score of 20 equals a 99% accurate probability and a Q-score of 30 equals a 99.9% accurate probability (Ewing & Green 1998; Ewing et al. 1998).

NGS bioinformatics software makes use of these Q-scores to trim poor quality-called bases usually at the 3' ends of read and in homopolymer regions. Once these bases are excluded, reads are mapped to a consensus sequence. Segminator (Archer et al. 2010) can be used to align sequence reads to an HIV consensus sequence using "K-mer mapping" (Altschul et al. 1997). K-mer mapping breaks both the reference sequence and the read sequence into shorter "words", while recording the location of the reference words. Then, the mapper aligns the reference and read words and records the reference location where the first read word maps. Finally, the mapper excludes the reference words located before the location where the first read word mapped and continues to align the rest of the read words to the reduced number of reference words (Archer et al. 2010). In essence, this mapping algorithm excludes unused parts of the reference sequence and makes mapping more efficient.

Datamonkey (Pond & Frost 2005) and RAMICS (Wright & Travers 2014) make use of codon aware mapping algorithms; Briefly, the reference sequence and the reads are translated into amino acid sequences. Next, the read amino acid sequence is mapped to the reference amino acid sequence and any indels would appear as an amino acid frame shift. Finally, identified frame shifts are interrogated on a nucleotide level, and indels are corrected before the read is realigned.

1.6 Study rationale

Considering that next generation sequencing (NGS) is a viable solution for pooled, low-cost, HIV drug resistance testing, we set out to test various up-stream sample preparation methods for amplicon sequencing on multiple NGS platforms. In an attempt to broaden the clinical application of NGS for studying HIV-1 minor variants in patients, we included understudied patient populations: adult patients who failed a second-line PI regimen, using 454 sequencing (454 Life Sciences, Connecticut, USA), and studied infants who were exposed to a PMTCT regimen consisting of NVP and AZT, using the novel Ion Torrent PGM (Life Technologies, California, USA) and Illumina MiSeq (California, USA) sequencing. After realising the limits of current library generation methods we further interrogated PCR resampling error by assaying plasmid mixtures and quantifying the relative representation in a sequenced library, followed by an attempt to reduce resampling error through partitioning reactions in minute droplets, during the first few PCR steps with an emulsified PCR mastermix. Accurate representation of minor variant populations in sequence-libraries is critical to improving the resolution of minor variant discrimination with NGS. The investigations in this dissertation, therefore, focus on the enrichment of patient samples to allow accurate minor variant drug resistance detection.

In light of the rationale above, our study hypothesis were twofold:

- Hypothesis A: NGS can be used to detect minor variant drug resistant mutations in understudied population such as 2nd line cART failed patients and in infants who failed PMTCT.
- Hypothesis B: Using water-in-oil emulsification to partition templates during fusion primer PCR can reduce PCR resampling error.

Hence, our specific objectives were as follows:

- To characterise minor variant resistance in patients failing a second-line LPV/r-based regimen.
- To characterise minor variant resistance in children exposed to the Western Cape prevention of mother-to-child HIV transmission (PMTCT) regimen.
- To compare deep sequencing results from the PGM (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) platforms to "gold standard" clonal sequencing.
- To measure primer-induced selection bias and PCR resampling attributed to the use of fusion primers.
- To test the ability of emulsification of the first few PCR cycles in reducing primer-induced selection bias.

1.7 References

Aghokeng, A.F. et al., 2011. High failure rate of the ViroSeq HIV-1 genotyping system for drug resistance testing in Cameroon, a country with broad HIV-1 genetic diversity. *Journal of clinical microbiology*, 49(4), pp.1635–1641.

Ahmadian, A. et al., 2000. Single-Nucleotide Polymorphism Analysis by Pyrosequencing. *Analytical Biochemistry*, 280(1), pp.103–110.

Altschul, S.F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* , 25(17), pp.3389–3402.

Archer, J. et al., 2010. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time--an ultra-deep approach. *PLoS computational biology*, 6(12), p.e1001022.

Balasubramanian, S., Klenerman, D. & Bentley, D., 2004. Arrayed biomolecules and their use in sequencing.

Bergroth, T., Sönnerborg, A. & Yun, Z., 2005. Discrimination of lamivudine resistant minor HIV-1 variants by selective real-time PCR. *Journal of Virological Methods*, 127(1), pp.100–107.

Berry, D. et al., 2011. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Applied and Environmental Microbiology* , 77(21), pp.7846–7849.

Boltz, V.F. et al., 2015. Analysis of Resistance Haplotypes Using Primer IDs and Next Gen Sequencing of HIV RNA Methods. *CROI*, p.1. Available at: http://www.croiconference.org/sites/default/files/posters-2015/593.pdf [Accessed July 19, 2015].

Bragg, L.M. et al., 2013. Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comput Biol*, 9(4), p.e1003031.

Brodin, J. et al., 2015. Challenges with using primer IDs to improve accuracy of next generation sequencing. *PloS one*, 10(3), p.e0119123.

Brook, I., 1987. Approval of zidovudine (AZT) for acquired immunodeficiency syndrome: A challenge to the medical and pharmaceutical communities. *JAMA*, 258(11), p.1517.

Cambiano, V. et al., 2013. Transmission of drug resistant HIV and its potential impact on mortality and treatment outcomes in resource-limited settings. *The Journal of infectious diseases*, 207 Suppl 2, pp.S57–62.

Chaturbhuj, D.N. et al., 2014. Evaluation of a Cost Effective In-House Method for HIV-1 Drug Resistance Genotyping Using Plasma Samples. *PLoS ONE*, 9(2), p.e87441.

De Clercq, E., 2009. Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *International Journal of Antimicrobial Agents*, 33(4), pp.307–320.

De Clercq, E., 2004. Antiviral drugs in current clinical use. *Journal of Clinical Virology*, 30(2), pp.115–133.

De Clercq, E., 2012. Human viral diseases: what is next for antiviral drug discovery? *Current Opinion in Virology*, 2(5), pp.572–579.

Connor, E.M. et al., 1994. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *The New England journal of medicine*, 331(18), pp.1173–1180.

D'Aquila, R.T. et al., 1996. Nevirapine, zidovudine, and didanosine compared with zidovudine and didanosine in patients with HIV-1 infection. A randomized, double-blind, placebo-controlled trial. National Institute of Allergy and Infectious Diseases AIDS Clinical Trials Group Protocol 241 Investigators. *Annals of internal medicine*, 124(12), pp.1019–1030.

Dabis, F. et al., 2005. Field efficacy of zidovudine, lamivudine and single-dose nevirapine to prevent peripartum HIV transmission. *AIDS (London, England)*, 19(3), pp.309–318.

Diehl, F. et al., 2006. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nature methods*, 3(7), pp.551–559.

Dudley, D.M. et al., 2012. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PloS one*, 7(5), p.e36494.

Edelstein, R.E. et al., 1998. Oligonucleotide Ligation Assay for Detecting Mutations in the Human Immunodeficiency Virus Type 1 pol Gene That Are Associated with Resistance to Zidovudine, Didanosine, and Lamivudine. *Journal of Clinical Microbiology*, 36(2), pp.569–572.

European AIDS Clinical Society, 2014. EACS Guidelines Version 7.1., (November), pp.1–85.

European collective study, 2005. Mother-to-child transmission of HIV infection in the era of highly active antiretroviral therapy. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 40(3), pp.458–465.

Ewing, B. et al., 1998. Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Research* , 8(3), pp.175–185.

Ewing, B. & Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research* , 8(3), pp.186–194.

Ezzell, C., 1987. AZT given the green light for clinical treatment of AIDS. *Nature*, 326(6112), p.430.

Gandhi, R.T. et al., 2003. Progressive Reversion of Human Immunodeficiency Virus Type 1 Resistance Mutations In Vivo after Transmission of a Multiply Drug-Resistant Virus. *Clinical Infectious Diseases*, 37(12), pp.1693–1698.

Guay, L.A. et al., 1999. Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *The Lancet*, 354(9181), pp.795–802.

Gulick, R.M. et al., 1997. Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *The New England journal of medicine*, 337(11), pp.734–739.

Gupta, R.K. et al., 2012. Global trends in antiretroviral resistance in treatment-naive individuals with HIV after rollout of antiretroviral treatment in resource-limited settings: a global collaborative study and meta-regression analysis. *Lancet*, 380(9849), pp.1250–1258.

Hadfield, J. & Loman, N., 2014. Omnics map. Available at: http://omicsmaps.com/ [Accessed August 7, 2015].

Hamers, R.L. et al., 2011. HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. *The Lancet Infectious Diseases*, 11(10), pp.750–759.

Hammer, S.M. et al., 1997. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. *The New England journal of medicine*, 337(11), pp.725–733.

Hirsch, M.S., 1988. Azidothymidine. *Journal of Infectious Diseases* , 157(3), pp.427–431.

Hori, M., Fukano, H. & Suzuki, Y., 2007. Uniform amplification of multiple DNAs by emulsion PCR. *Biochemical and biophysical research communications*, 352(2), pp.323–328.

Hu, W.-S. & Hughes, S.H., 2012. HIV-1 Reverse Transcription. *Cold Spring Harbor Perspectives in Medicine*, 2(10).

Huse, S.M. et al., 2007. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome biology*, 8(7), p.R143.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Jourdain, G. et al., 2010. Association between detection of HIV-1 DNA resistance mutations by a sensitive assay at initiation of antiretroviral therapy and virologic failure. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 50(10), pp.1397–1404.

Ju, J. et al., 2006. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences* , 103(52), pp.19635–19640.

Kearney, M.F. et al., 2014. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS pathogens*, 10(3), p.e1004010.

Keys, J.R. et al., 2015. Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS research and human retroviruses*, 31(6), pp.658–668.

Kleppe, K. et al., 1971. Studies on polynucleotides: XCVI. Repair replication of short synthetic DNA's as catalyzed by DNA polymerases. *Journal of Molecular Biology*, 56(2), pp.341–361.

Kulkarni, R. et al., 2012. The HIV-1 Reverse Transcriptase M184I Mutation Enhances the E138K-Associated Resistance to Rilpivirine and Decreases Viral Fitness. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 59(1), pp.47-54.

Lalezari, J.P. et al., 2003. Enfuvirtide, an HIV-1 Fusion Inhibitor, for Drug-Resistant HIV Infection in North and South America. *New England Journal of Medicine*, 348(22), pp.2175–2185.

Li, H. & Homer, N., 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics* , 11(5), pp.473–483.

Li, J.Z. et al., 2014. Comparison of illumina and 454 deep sequencing in participants failing raltegravir-based antiretroviral therapy. *PLoS one*, 9(3), p.e90485.

Li, J.Z. et al., 2013. Impact of minority nonnucleoside reverse transcriptase inhibitor resistance mutations on resistance genotype after virologic failure. *The Journal of infectious diseases*, 207(6), pp.893–897.

Li, J.Z. et al., 2012. Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS (London, England)*, 26(2), pp.185–192.

Little, S.J. et al., 2008. Persistence of Transmitted Drug Resistance among Subjects with Primary Human Immunodeficiency Virus Infection. *Journal of Virology* , 82(11), pp.5510–5518.

Liu, L. et al., 2012. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.

Liu, S.L. et al., 1996. HIV quasispecies and resampling. *Science (New York, N.Y.)*, 273(5274), pp.415–416.

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–439.

Manzardo, C. & Gatell, J.M., 2014. Stribild(R) (elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate): a new paradigm for HIV-1 treatment. *AIDS reviews*, 16(1), pp.35–42.

Marconi, V.C. et al., 2008. Prevalence of HIV-1 Drug Resistance after Failure of a First Highly Active Antiretroviral Regimen in KwaZulu Natal, South Africa. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 46(10), pp.1589–1597.

Margulies, M. et al., 2005. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. *Nature*, 437(7057), pp.376–380.

McInerney, P., Adams, P. & Hadi, M.Z., 2014. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Molecular biology international*, 2014, p.287430.

Metzker, M.L., 2005. Emerging technologies in DNA sequencing. *Genome Research*, 15(12), pp.1767–1776.

Metzner, K.J. et al., 2009. Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 48(2), pp.239–247.

Metzner, K.J.. c et al., 2003. Emergence of Minor Populations of Human Immunodeficiency Virus Type 1 Carrying the M184V and L90M Mutations in Subjects Undergoing Structured Treatment Interruptions. *Journal of Infectious Diseases*, 188(10), pp.1433–1443.

Metzner, K.J.. f et al., 2005. Detection of minor populations of drug-resistant HIV-1 in acute seroconverters. *AIDS*, 19(16), pp.1819–1825.

Meynert, A.M. et al., 2013. Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC bioinformatics*, 14, p.195.

Molina, J.M. et al., 1994. Quantification of HIV-1 virus load under zidovudine therapy in patients with symptomatic HIV infection: relation to disease progression. *AIDS (London, England)*, 8(1), pp.27–33.

Montaner, J.S. et al., 1998. A randomized, double-blind trial comparing combinations of nevirapine, didanosine, and zidovudine for HIV-infected patients: the INCAS Trial. Italy, The Netherlands, Canada and Australia Study. *JAMA*, 279(12), pp.930–937.

Morey, M. et al., 2013. A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism*, 110(1-2), pp.3–24.

Nachega, J.B. et al., 2011. HIV treatment adherence, drug resistance, virologic failure: evolving concepts. *Infectious disorders drug targets*, 11(2), pp.167–174.

Nachega, J.B. et al., 2014. Lower Pill Burden and Once-Daily Antiretroviral Treatment Regimens for HIV Infection: A Meta-Analysis of Randomized Controlled Trials. *Clinical Infectious Diseases*, 58 (9), pp.1297–1307.

Newman, H. et al., 2014. A qualitative PCR minipool strategy to screen for virologic failure and antiretroviral drug resistance in South African patients on first-line antiretroviral therapy. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology*, 60(4), pp.387–391.

Palmer, S. et al., 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *Journal of clinical microbiology*, 43(1), pp.406–413.

Paredes, R. et al., 2010. Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *The Journal of infectious diseases*, 201(5), pp.662–671.

Paredes, R. et al., 2007. Systematic evaluation of allele-specific real-time PCR for the detection of minor HIV-1 variants with pol and env resistance mutations. *Journal of Virological Methods*, 146(1–2), pp.136–146.

Parkin, N. et al., 2012. Evaluation of In-house Genotyping Assay Performance Using Dried Blood Spot Specimens in the Global World Health Organisation Laboratory Network. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 54(Suppl 4), pp.S273–9.

Phillips, A.N. et al., 2011. Effect on transmission of HIV-1 resistance of timing of implementation of viral load monitoring to determine switches from first to second-line antiretroviral regimens in resource-limited settings. *AIDS*, 25(6), pp. 843-50.

Pond, S.L.K. & Frost, S.D.W., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)*, 21(10), pp.2531–2533.

Ramachandran, S. et al., 2008. End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions. *Journal of virological methods*, 151(2), pp.217–224.

Rohland, N. & Reich, D., 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, 22(5), pp.939–946.

Ronaghi, M. et al., 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242(1), pp.84–89.

Rosenblum, M. et al., 2009. The Risk of Virologic Failure Decreases with Duration of HIV Suppression, at Greater than 50% Adherence to Antiretroviral Therapy. *PLoS ONE*, 4(9), p.e7196.

Rothberg, J.M. et al., 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356), pp.348–352.

Rowley, C.F. et al., 2010. Ultrasensitive detection of minor drug-resistant variants for HIV after nevirapine exposure using allele-specific PCR: clinical significance. *AIDS research and human retroviruses*, 26(3), pp.293–300.

Rusk, N., 2011. Torrents of sequence. *Nature Methods*, 8(1), p.44.

Salgado, M. et al., 2010. Evolution of the HIV-1 nef gene in HLA-B*57 positive elite suppressors. *Retrovirology*, 7, p.94.

Shendure, J. et al., 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741), pp.1728–1732.

Shisana, O. et al., 2014. *South African national HIV prevalence, incidence and behaviour survey, 2012*, Available at: http://www.hsrc.ac.za/uploads/pageContent/4565/SABSSM IV LEO final.pdf.

Siliciano, J.D. & Siliciano, R.F., 2004. A long-term latent reservoir for HIV-1: discovery and clinical implications. *The Journal of antimicrobial chemotherapy*, 54(1), pp.6–9.

Smith, D.M. et al., 2009. The use of pooled viral load testing to identify antiretroviral treatment failure. *AIDS (London, England)*, 23(16), pp.2151–2158.

Statistics South Africa, 2014. *Statistical release Mid-year population estimates*, Available at: http://www.statssa.gov.za/publications/P0302/P03022015.pdf.

Tobe, V.O., Taylor, S.L. & Nickerson, D.A., 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. *Nucleic Acids Research*, 24(19), pp.3728–3732.

Ton, Q. & Frenkel, L., 2013. HIV drug resistance in mothers and infants following use of antiretrovirals to prevent mother-to-child transmission. *Current HIV research*, 11(2), pp.126–136.

Wallis, C.L. et al., 2014. Drug Susceptibility and Resistance Mutations After First-Line Failure in Resource Limited Settings. *Clinical Infectious Diseases*, 59(5), pp.706–715.

Wallis, C.L. et al., 2011. Protease Inhibitor Resistance Is Uncommon in HIV-1 Subtype C Infected Patients on Failing Second-Line Lopinavir/r-Containing Antiretroviral Therapy in South Africa. *AIDS Research and Treatment*, 2011, p.769627.

Wang, C. et al., 2007. Characterization of mutation spectra with ultra-deep pyrosequencing: Application to HIV-1 drug resistance. *Genome Research*, 17(8), pp.1195–1201.

Zur Wiesch, P.A. et al., 2011. Population biological principles of drug-resistance evolution in infectious diseases. *The Lancet Infectious Diseases*, 11(3), pp.236–247.

Williams, I.G. & De Cock, K.M., 1996. The XI international conference on AIDS. Vancouver 7-12 July 1996. A review of Clinical Science Track B. *Genitourinary medicine*, 72(5), pp.365–369.

Woods, C.K. et al., 2012. Automating HIV Drug Resistance Genotyping with RECall, a Freely Accessible Sequence Analysis Tool. *Journal of Clinical Microbiology*, 50(6), pp.1936–1942.

World Health Organisation, 2013. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection. Geneva: World Health Organisation; 2013., pp.1–272. Available at: http://apps.who.int/iris/bitstream/10665/85321/1/9789241505727_eng.pdf?ua=1.

World Health Organisation, 2015. Fact Sheet No 360. *facts sheet*. Available at: http://www.who.int/mediacentre/factsheets/fs360/en/ [Accessed August 8, 2015].

World Health Organisation, 2012. *WHO HIV DRUG RESISTANCE REPORT 2012*, Geneva, Switzerland. Available at: http://apps.who.int/iris/bitstream/10665/75183/1/9789241503938_eng.pdf.

Wright, I.A. & Travers, S.A., 2014. RAMICS: trainable, high-speed and biologically relevant alignment of high-throughput sequencing reads to coding DNA. *Nucleic acids research*, 42(13), p.e106.

Zhou, S. et al., 2015. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of virology*. [Cited ahead of print]

Van Zyl, G.U. et al., 2014. Emerging antiretroviral drug resistance in sub-Saharan Africa: novel affordable technologies are needed to provide resistance testing for individual and public health benefits. *AIDS (London, England)*, 28(18), pp.2643–2648.

Van Zyl, G.U. et al., 2011. Low lopinavir plasma or hair concentrations explain second line protease inhibitor failures in a resource-limited setting. *Journal of acquired immune deficiency syndromes* (1999), 56(4), pp.333–339.

Van Zyl, G.U. et al., 2011. Pooling strategies to reduce the cost of HIV-1 RNA load monitoring in a resource-limited setting. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 52(2), pp.264–270.

1.8 Appendix A: Supplementary figures

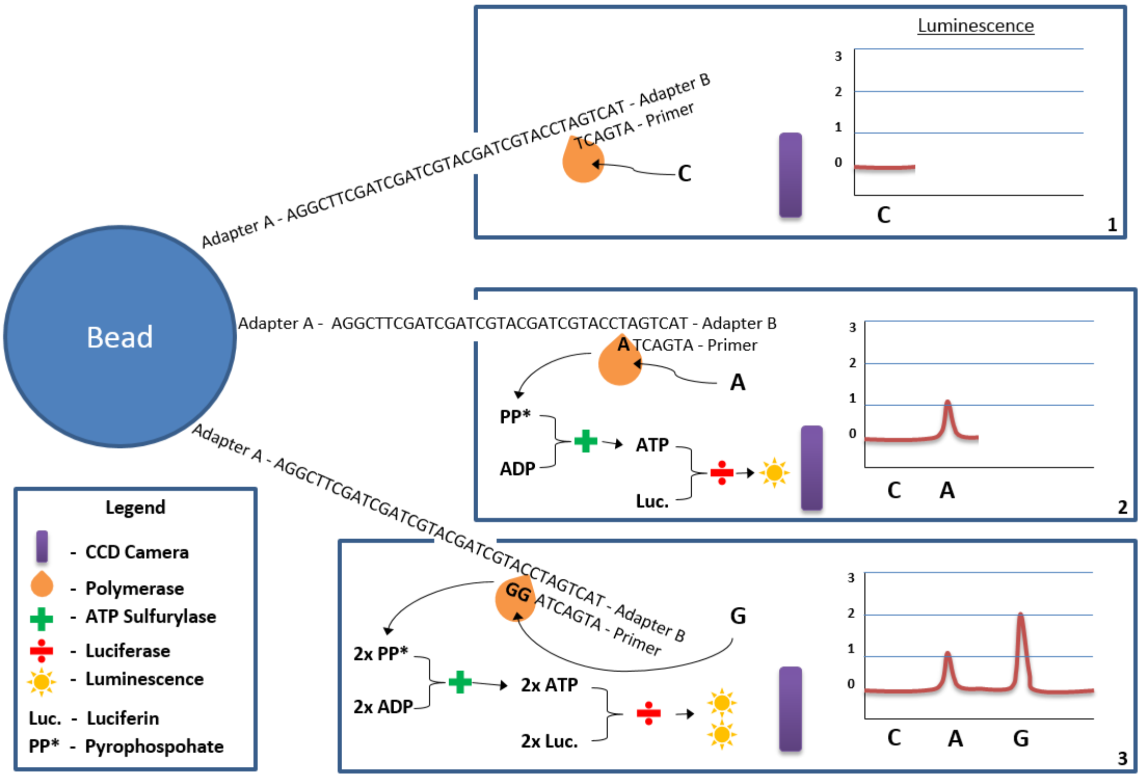


Figure 1.8.1 Life Technologies 454 sequencing strategy: Amplicons are bound to a bead via adaptor ligation. After these beads undergo clonal enrichment through emulsion PCR, they are placed onto a flow-cell where the amplicon is primed for pyrosequencing to begin. **In box 1**, cytosines (**C**) are washed over the well but cannot be incorporation into the complementary DNA strand being synthesised, since they are not commentary to the template strand. These nucleotides are then washed away with the subsequent wash step and no luminescence is observed. **In box 2**, adenosines (**A**) are washed over the well and one is incorporated into the complementary DNA strand, resulting in the release of a pyrophosphate as a DNA polymerisation by-product. ATP sulfurylase combines this free pyrophosphate with an ADP molecule to produce ATP. The ATP molecule activates luciferase which in turn catalyses luciferin, producing a chemiluminescent signal. The luminescence is captured by a CCD camera and the amplitude is converted into the number of adenosines incorporated. **In box 3**, two guanines (**G**) are incorporated into the complementary DNA strand, causing the release of two pyrophosphate molecules. Once again, ATP sulfurylase combines these two pyrophosphates with two SDP molecules to produce two ATPs. These ATPs then excite two luciferases and two luciferins are broken down to emit luminescence twice as strong as when one luciferin is catalysed. The CCD camera records this luminescence and translates the amplitude into the incorporation of two guanosines.

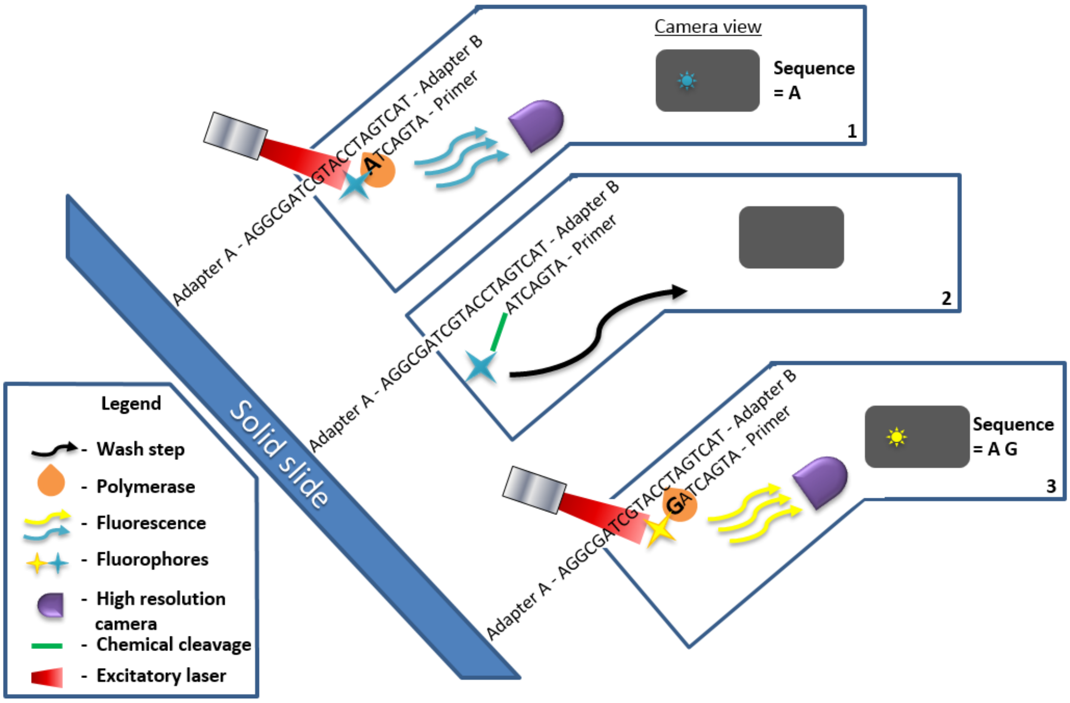


Figure 1.8.2 Solexa/Illumina sequencing strategy: Amplicons are bound to a solid surface via adaptor ligation. After these amplicons undergo clonal enrichment by bridge PCR, the amplicon is primed for CRT sequencing. **In box 1**, an adenosine (**A**) with a 3' fluorophore is incorporated into the complementary DNA strand being synthesised. Since the polymerase cannot cleave off the fluorophore, sequencing is temporarily terminated. An excitatory laser excites the fluorophore and a high-resolution camera captures the fluorescence. Since each nucleotide has a specific fluorophore, the observed fluorescence wavelength determines the identity of the incorporated nucleotide. **In box 2**, the fluorophore is chemically cleaved and washed away, and complementary strand synthesis termination is reversed. **In box 3**, a guanine (**G**) is incorporated and once again the wavelength of its' specific fluorophore is recorded by the high-resolution camera. Since the polymerase cannot cleave the fluorophore, synthesis is once again terminated until the fluorophore is chemically cleaved. By recording the sequence of detected fluorophores at a specific location on the slide, the sequence of the amplicon colony is determined. Using this strategy, Solexa/Illumina NGS platforms are able to accurately sequence homopolymer stretches, one base at a time.

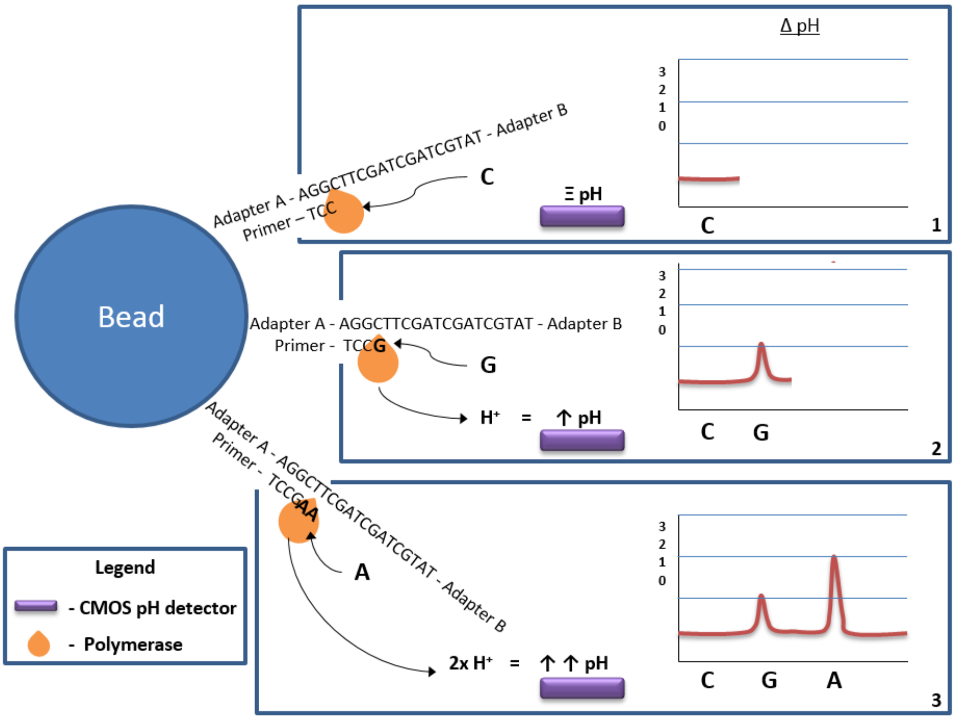


Figure 1.8.3 Life Sciences Ion Torrent sequencing strategy: Amplicons are bound to a bead via adaptor ligation. After these beads undergo clonal enrichment through emulsion PCR, they are placed onto a CMOS sequencing chip where the amplicon is primed for semiconductor sequencing to begin. **In box 1**, cytosines (**C**) are washed over the sequencing chip but cannot be incorporation into the complementary DNA strand being synthesised, since they are not commentary to the template strand. These nucleotides are then washed away with the subsequent wash step and no changed is observed in the reaction pH. **In box 2**, guanosines (**G**) are washed over the sequencing chip and one is incorporated into the complementary DNA strand, resulting in the release of a hydrogen molecule (a proton) as a DNA polymerisation by-product. The release of this proton is detected by the semiconductor pH sensor embedded in the sequencing chip. This pH change is translated into the incorporation of a single nucleotide. **In box 3**, two adenines (**A**) are incorporated into the complementary DNA strand causing the release of two protons. This results in a two-fold change in the reaction pH which is recorded by the semiconductor pH sensor and translated into the incorporation of two nucleotides.

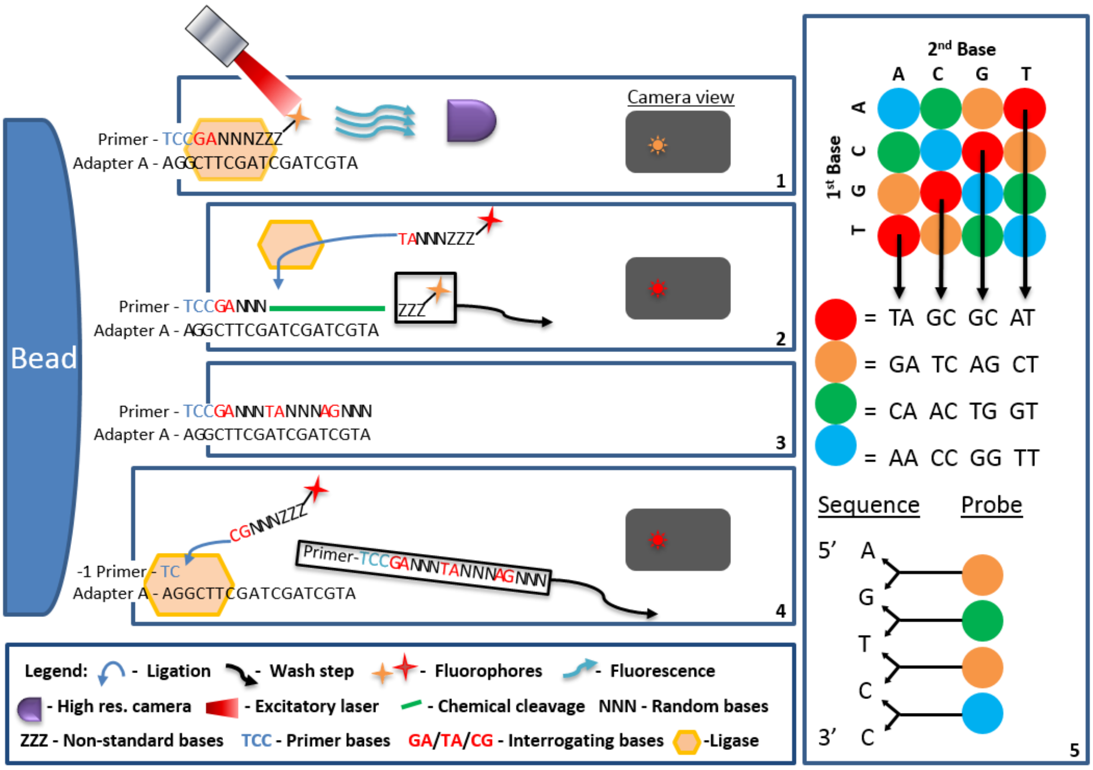

Bead
Primer - TCCGANNNZZZ
Adapter A - AGGCTTCGATCGATCGTA
Camera view
1
TANNNZZZ
Primer - TCCGANNN
Adapter A - AGGCTTCGATCGATCGTA
ZZZ
2
Primer - TCCGANNNTANNNAGNNN
Adapter A - AGGCTTCGATCGATCGTA
3
CGNNNZZZ
-1 Primer - TC
Adapter A - AGGCTTCGATCGATCGTA
Primer-TCCGANNNTANNNAGNNN
4
2nd Base
A C G T
1st Base
A C G T
= TA GC GC AT
= GA TC AG CT
= CA AC TG GT
= AA CC GG TT
Sequence
Probe
5'
A
G
T
C
3'
C
5
Legend: - Ligation - Wash step - Fluorophores - Fluorescence
- High res. camera - Excitatory laser - Chemical cleavage NNN - Random bases
ZZZ - Non-standard bases TCC - Primer bases GA/TA/CG - Interrogating bases -Ligase

Figure 1.8.4 Agencourt/ABI sequencing strategy: Amplicons are bound to a bead via adaptor ligation. After these beads undergo clonal enrichment through emulsion PCR on a solid surface, they are primed for sequencing-by-ligation (SBL). **In box 1**, an eight-base probe with a 3' fluorophore is ligated to the template DNA strand. The identity of the first two bases of the probe are known and correspond to one of four fluorophores. This probe interrogates bases one and two of the template sequence and once ligated, the fluorophore is identified using an excitatory laser and a high-resolution camera. **In box 2**, the last three bases of the probe are chemically leaved, along with the 3' fluorophore. The next probe is ligate to the template, interrogating bases six and seven, before the last three bases and the fluorophore are again cleaved. **In box 3**, the template has been fully complemented by the ligation of multiple probes. **In box 4**, the complementary DNA strand and the first primer are denatured and washed away before a new, "-1 primer" (with one less base on the 3' end) is bound to the template DNA strand. Next, another probe is ligated to the template strand, this time binding at the minus one position and interrogating template bases minus one and one. This sequence of ligation and cleavage events are repeated until each base has been interrogated twice (each time with a separate probe). By recording the sequence of detected fluorophores at a specific location on the solid surface, the sequence of the amplicon is determined. This strategy enables SBL NGS platforms to accurately sequence homopolymer stretches, using two base interrogation. **Box 5** shows the various fluorophore colours and the first and second bases which they represent. Using this strategy, each fluorophore represents four nucleotide pairs and 16 pairs in total. The identity of a specific base can only be determined once both fluorophores interrogating that base, are known.

2 Deep sequencing reveals minor *protease* resistance mutations in patients failing a protease inhibitor regimen.

2.1 Background

Antiretroviral therapy (ART) reduces human immunodeficiency virus (HIV) replication through the inhibition of various steps in the replication cycle (Clavel & Hance 2004). HIV-infected adults who qualify for treatment are initiated on first-line combination antiretroviral therapy (cART) (Palella et al. 1998; Gulick et al. 1997). The World Health Organisation (WHO) recommends a non-nucleoside reverse transcriptase inhibitor (NNRTI) and two nucleoside reverse transcriptase inhibitors (NRTIs), one of which is lamivudine (3TC), for first-line therapy in resource limited settings (World Health Organisation 2010). These combinations have low genetic barriers since therapy failure requires few drug resistance mutations (DRMs) to confer a high level of drug resistance. Nevertheless patients may remain on these low barrier regimens for many years when maintaining good adherence as the risk of therapy failure decreases with good adherence (Rosenblum et al. 2009).

At the time of this study, virologic failure (VF) was defined as a persistent viral load (VL) >1000 c/ml, after a period of adherence intensification. Patients with VF qualified for second-line cART consisting of two NRTIs and a protease inhibitor (PI), lopinavir (LPV) boosted with a low dose of ritonavir (LPV/r) (National Department of Health 2010). The addition of ritonavir increases the half-life and trough levels resulting in a higher genetic barrier. Study participants were included from a larger study monitoring the relationship between low plasma and hair drug titres and second-line therapy failure (van Zyl et al. 2011). For the purpose of this study, patients with a VL>500 HIV-1 RNA c/ml obtained by HIV-1 real-time viral load assay (Abbott Laboratories, Illinois, USA), received HIV drug resistance testing (DRT) with an in-house PCR and sequencing method. Participants had experienced VF on an NNRTI-based first-line cART with at least the 3TC drug resistance mutation (DRM) M184V/I, in addition to VF on the second-line PI-based cART.

Six of the seven patients exhibited insufficient plasma-LPV concentrations (<0.87 µg/ml) and inadequate LPV and ritonavir drug-hair titres (<3.15 and 0.54 ng/mg-hair respectively), as assayed by liquid chromatography-mass spectrometry. These low concentrations were indicative of poor therapy adherence: LPV-plasma and hair concentrations less than 1µg/ml and 3.63ng/mg, respectively were shown to predict therapy failure (van Zyl et al. 2011).

2.2 Methods

2.2.1 Inclusion criteria

Study participants were part of a larger study investigating the relationship between second-line cART failure and plasma and hair-LPV/r concentrations (van Zyl et al. 2011). Inclusion in our study required previous first-line cART failure with at least the lamivudine DRM M184V/I and current second-line failure (defined for the purpose of this study as a VL >500 c/ml), as well as sufficient (1 ml) stored plasma from a historical EDTA blood sample. As part of routine clinical practice, study participants had their viral loads measured. For the study, in-house HIV drug resistance testing (DRT) by PCR and Sanger sequencing were performed at both first and second-line cART failure, while next generation sequencing (NGS) was only performed at second-line failure.

2.2.2 In-house DRT

In-house DRT (or bulk sequencing) consisted of a two-step reverse transcription-PCR (RT-PCR), a nested PCR and a sequencing reaction. These experiments are described below.

2.2.2.1 RT-PCR

Viral RNA was extracted using the QIAamp® UltraSence® (Qiagen, Nimburg, Netherlands) virus kit according to the manufacturer specified protocol and stored at -80 °C until needed. RT-PCR was a variation of a previously described method (Plantier et al. 2005) and was carried out using the pre-nested primers listed in table 2.2.2.1.1 (below), the Access RT-PCR kit (Promega, Wisconsin, USA) and the reaction conditions that follow.

Table 2.2.2.1.1 RT-PCR pre-nested primers

Description	Name	Sequence (5’-3’)	HXB2 binding
Forward primer	5’PROT1	TAATTTTTTAGGGAAGATCTGGCCTTCC	2082 to 2109
Reverse primer	Mj4	CTGTTAGTGCTTTGGTTCCTCT	3399 to 3420 (RC)

(RC) = Reverse complement.

The RT-PCR was compiled using the reagents listed below at their indicated concentrations. All reactions were prepared on ice in a mastermix and aliquoted before the extracted RNA was added.

Reagent	Final concentration (in 50 µl)
1. 5 x AMV/*Tfl* Reaction Buffer	1 x
2. dNTP Mix (10mM of each)	0.2 mM
3. 10 µl extracted RNA	Unknown
4. 25 mM $MgSO_4$	1 mM
5. 5'prot1 FWD Primer	1 µM
6. Mj4 REV Primer	1 µM
7. AMV Reverse Transcriptase (5 U/µl)	0.1 U/µl
8. *Tfl* DNA Polymerase (5 U/µl)	0.1 U/µl
9. 23 µl Nuclease free water	1 x

Thermocycling was done using an ABI 9700 thermocycler (Applied Biosystems, Massachusetts, USA) with the heat lid activated at 105 °C and the following cycling conditions:

45 °C for 45 min (reverse transcription)

94 °C for 2 min

94 °C for 30 sec
40x 60 °C for 1 min
68 °C for 2 min

68 °C for 7 min

4 °C soak

On completion, reactions were purified using the QIAquick® PCR clean-up kit (Qiagen, Nimburg, Netherlands) following the manufacturer's spin column protocol, before the PN-PCR products were eluted in 50µl of elution buffer. This pre-nested product was used for both the bulk sequencing nested PCR (described below) and the 454 fusion primer nested PCR (described in section 2.2.3).

2.2.2.2 Nested PCR

Nested PCR was performed using the GoTaq® Flexi kit (Promega, Wisconsin, USA) according to the manufacturer's specifications, using the primers listed in table 2.2.2.2.1 below.

Table 2.2.2.2.1 Nested PCR primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward Primer	5'prot2	TCAGAGCAGACCAGAGCCAACAGCCCCA	2136 to 2163
Reverse Primer	NE135	CCTACTAACTTCTGTATGTCATTGACAGTCCAGCT	3300 to 3334 (RC)

(RC) = Reverse complement.

The reaction mastermix was prepared on ice and aliquoted in a DNA-free area before pre-nested template was added in a post-amplification, nested PCR hood. The mastermix contained the following:

	Reagent	Final concentration (in 50 µl)
1.	5 x Reaction Buffer	1 x
2.	dNTP Mix (10mM of each)	0.2 mM
3.	25 mM $MgCl_2$	2 mM
4.	5'prot2 FWD Primer	500 nM
5.	Mj4 REV Primer	500 nM
6.	2 µl PN-PCR template	Unknown
7.	GoTaq® Flexi DNA Polymerase (5 U/µl) (Pormega)	1.25 U/rxn
8.	30.75 µl Nuclease free water	1 x

Nested PCR thermocycling conditions are seen below:

94 °C for 2 min

$$40x \left\{ \begin{array}{l} 94\ °C \text{ for } 30 \text{ sec} \\ 55\ °C \text{ for } 30 \text{ sec} \\ 72\ °C \text{ for } 1 \text{ min} \end{array} \right.$$

72 °C for 7 min

4 °C soak

Once completed, nested amplicons were purified using the QIAquick® PCR clean-up kit (Qiagen, Nimburg, Netherlands) and the products were separated on a 1% agarose gel prepared in TBE buffer (data not shown).

2.2.2.3 Sanger sequencing

Sequencing reactions used an in-house optimised method using the BigDye® Terminator Direct Cycle (V3.1) sequencing kit (Applied Biosystems, Massachusetts, USA), and each of the primers listed in table 2.2.2.3.1.

Table 2.2.2.3.1 Sequencing reaction primers

Description	Name	Sequence (5'-3')	HXB2 binding	Tm
Forward Primer	POL1D	TCCCTCAAATCACTCTTTGGC	2251 to 2271	56.3 °C
Forward Primer	AK11	GTACCAGTAAAATTAAARCCAG	2571 to 2592	48.4 °C
Forward Primer	POL3D	CAGTACTGGATGTGGG	2869 to 2884	48.7 °C
Forward Primer	ABB20-3F	ATCAGTACAATGTGCTTCCA	2980 to 2999	51.6 °C
Reverse Primer	AK12	TGGTGTYTCATTRTTTRYACTAG	2947 to 2969 (RC)	50.6 °C
Reverse Primer	POL3REV	CTGAAAAATATGCATCCCCC	2882 to 2901 (RC)	51.1 °C
Reverse Primer	JA217	CTTTTATTTTTTCTTCTGTCAATGG	2622 to 2646 (RC)	49.5 °C
Reverse Primer	R2051	TATRTTGACAGGTGTAGGT	2486 to 2504 (RC)	49.5 °C

(RC) = Reverse complement.

Sequencing reaction mastermix was comprised of the following reagents, aliquoted in a 96 well plate:

	Reagent	Final concentration (in 10 µl)
1.	5 x Terminator sequencing reaction buffer	1.5 x
2.	BigDye® Terminator reaction mix	1 x
3.	Fwd or Rev Sequencing Primer	5 pMoles
4.	Nested PCR Product	15 to 20 ng

Sequencing reactions were performed using the following in-house-optimised reaction conditions in an ABI 9700 thermocycler (Applied Biosystems, Massachusetts, USA):

$$25\ \text{x} \left\{ \begin{array}{l} 95\ ^\circ\text{C for 10 sec} \\ 45 \text{ or } 50\ ^\circ\text{C for 5 sec*} \\ 60\ ^\circ\text{C for 4 min} \end{array} \right.$$

* Primers were annealed at the sequence-primer-specific annealing temperature.

On completion, the sequencing reaction was purified using an in-house-optimised BigDye® X-terminator (Applied Biosystems, Massachusetts, USA) clean up protocol (below). The SAM Solution™ (Applied Biosystems, Massachusetts, USA) used in the purification reaction was heated to room temperature before use.

	Reagent	Volume added to Mastermix
1.	SAM™ solution (Applied Biosystems)	49 µl
2.	X-terminator solution (Applied Biosystems)	11 µl

Fifty-five microlitres of the above reaction mixture was added to each sequencing PCR in the 96 well plate before the wells were sealed and the reaction was vortexed for 30 to 45 min. The purified sequencing reaction was then centrifuged at 1000 xg for 5 min to pellet the X-terminator particles and the sequencing reaction was read on an ABI Genetic Analyser 3130xL (Applied Biosystems, Massachusetts, USA).

2.2.2.4 Post-sequencing analysis

Sequences were analysed using Sequencher V4.7 software (Gene Codes Corporation, Ann Arbor, USA) to exclude poor quality bases and to construct sequence contigs. Contiguous sequences were submitted to the Stanford University online HIV drug resistance database (http://hivdb.stanford.edu) for HIV drug resistance interpretation.

2.2.3 NGS sample preparation

An overview of the NGS enrichment strategy can be seen in Appendix B, along with the list of DRMs sampled by each of the overlapping sequence amplicons.

2.2.3.1 Nested fusion primer PCR

Since enriched *pol* fragments were available from the RT-PCR (section 2.2.2.1 above), the nested PCR was prepared using previously published 454 primers, amplifying the entire *protease* gene and the first 245 codons of *reverse transcriptase* in three overlapping fragments (Varghese et al. 2010). These primers were modified with the addition of six bases to the four-base molecular identifier (MID) and the 3' end of two primer species were extended to improve 3' conservation. The detailed primer schematics are attached in Appendix B, however, the condensed format of the 454 fusion primers are seen in table 2.2.3.1.1 below:

Table 2.2.3.1.1 Condensed fusion primers

Description	Name	Mandatory additions – *MID*# – Gene-specific Sequence	HXB2 binding
Forward Primer	PR-F	Adaptor-A-TCAG-*MID*-CCTCARATCACTCTTTGGC	2253 to 2271
Reverse Primer	PR-R	Adaptor-B-TCAG-*MID*-YTTGGGCCATCCATTCCTGG	2589 to 2608 (RC)
Forward Primer	RTA-F	Adaptor-A-TCAG-*MID*-TGCACAYTAAATTTTCCAATTAGa*	2535 to 2557
Reverse Primer	RTA-R	Adaptor-B-TCAG-*MID*-ACTAGGTATGGTGAATGCAG	2932 to 2951 (RC)
Forward Primer	RTB-F	Adaptor-A-TCAG-*MID*-CTRGATGTGGGRGATGCAta*	2874 to 2891
Reverse Primer	RTB-R	Adaptor-B-TCAG-*MID*- TGTWTWGGCTGTACTGTCC	3265 to 3284 (RC)

(RC) = Reverse complement. # Ten-base MIDs varied for each patient. * Bases added for 3' conservation are in lower case.

Expand High Fidelity PCR (Roche Diagnostics, Basel, Switzerland) system was used in conjunction with each of the fusion primer pairs listed in table 2.2.3.1.1 above (i.e.: PR, RTA and RTB), to produce the 454 FLX (454 Life Sciences, Connecticut, USA) sequence amplicons. The reagents and reaction conditions are described below:

	Reagent	Final concentration (in 50 µl)
1.	10 x Reaction buffer	1 x
2.	dNTP Mix (10 mM of each)	0.2 mM
3.	$MgCl_2$ (25 mM)	3 mM
4.	PR/RTA/RTB Forward Primer	800 nM
5.	PR/RTA/RTB Reverse Primer	800 nM
6.	Expand HiFi Polymerase (Roche)	2.625 U/rxn
7.	2.5 µl Pre-nested product	Unknown
8.	37.25 µl Nuclease free water	1 x

Since each primer pair had a unique optimal annealing temperature, each reaction was amplified separately in an ABI 9700 thermocycler (Applied Biosystems, Massachusetts, USA) using the thermocycling conditions seen below.

94 °C for 2 sec

40 x
- 94 °C for 30 sec
- *45-50 °C for 30 sec
- 72 °C for 1 min

72 °C for 7 min

4 °C soak

*The PR fragment was annealed at 48 °C, the RTA fragment at 45 °C and the RTB fragment at 50 °C

2.2.3.2 Amplicon size selection and purification

On completion, nested amplicons were separated in a 0.8% agarose gel prepared in sodium borate buffer. To prevent DNA degradation by ultraviolet light (UV), a non-UV-dependant DNA stain, GelStar (Lonza, Maryland, USA), was used to visualise amplicons over blue light with an orange exclusion filter on a Dark Reader (Clare Chemicals, Colorado, USA). Once separated, amplicons were size selected by excising specific bands from the electrophoresis gel using a different scalpel blade for each fragment. These gel slices were weighed and gel purified using the Wizards Gel and PCR purification kit (Promega, Wisconsin, USA) according to the manufacture's specification. Amplicons were eluted in 30µl of warm nuclease free water and quantified in triplicate with a NanoDrop ND1000 (Thermo Scientific, Massachusetts, USA), according to the manufacturer directions. These size-selected amplicons were sent to Inqaba Biotech in Pretoria of South Africa, for NGS on the 454 FLX Titanium platform (454 Life Sciences, Connecticut, USA).

2.2.4 FLX Titanium sequencing

For NGS library preparation, the 454 Life Sciences FLX Titanium Lib-A emPCR protocol was followed according to the manufacturers' specifications for bidirectional amplicons sequencing (Roche Diagnostics, Basel, Switzerland).

2.2.5 Data analysis

After sequencing, data was available in standard flowgram format (.sff) which contains both the amplicons nucleotide sequence and the quality score assigned to each base-call. Using CLC Genomics Workbench V5.1 (CLC Bio, Aarhus, Denmark), these sequence strings (referred to here as "reads") were de-multiplexed according to patient-specific MIDs. Reads were quality filtered with a 95% base-call quality confidence: In short, a 5-base sliding window interrogated bases starting at the 3' end of the sequence. If the average quality score within the window was less than 95%, the last base was truncated and the window was incremented by one base towards the 5' end. If the average score was more that 95%, the window was once again incremented but no bases were removed.

A South African HIV-1 subtype C consensus sequence was generated from 317 whole genomes, sequenced between 1997 and 2010 and downloaded from HIV online database (attainable at http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). These sequences were aligned using the Geneious align tool in Geneious V6.0 (Biomatters, Auckland, New Zealand) to produce the consensus sequence that was used as a reference for read alignment.

The "cleaned" reads were mapped to the generated subtype C reference sequence using Segminator V1.3.3, an online read mapping tool (available at http://www.phylogenetictrees.com/segminator.php#2) designed specifically for HIV NGS data analysis (Archer et al. 2010). Segminator results were available in comma separated format (.csv) and were imported into Excel 2007 (Microsoft, Washington, USA). In Excel (Microsoft, Washington, USA), the range of possible amino acids were pared against *protease* and *reverse transcriptase* codons and minority, drug-resistant variants were seen as subpopulations of the total coverage at a DRM loci. For each patient, the list of observed minority DRMs greater than 0.5% prevalence was submitted to the Stanford University drug resistance database (version 6.0.11) for susceptibility analysis (available online at: http://sierra2.stanford.edu/sierra/servlet/JSierra).

2.3 Results

2.3.1 Cohort

The study cohort demographics are seen in table 2.3.1.1 below as well as the first and second line therapy formulations. At therapy failure events, the patient viral loads (as determined by the Abbot Laboratories (Illinois, USA) HIV-1 real-time viral load assay) and the CD4 cell counts were available in viral c/ml and CD4 cells/µl respectively, as seen below.

Table 2.3.1.1 Demographic and therapy information of patients

	Demographic information			First line therapy					Second line therapy (AZT, DDI, LPV/r)					
Study no	**Age (current)**	**Gender**	**Ethnic group**	**ART**	**Therapy duration (months)**	**Failure period (months)**	**VL**	**CD4**	**Therapy duration (months)**	**Failure period (months)**	**VL**	**CD4**	**LPV-plasma**	**LPV-hair**
1	24	F	mixed race	d4T, 3TC, NVP	33	2	4000	276	10	7	2100000	358	LDL	0.54
2	35	F	mixed race	d4T, 3TC, NVP(EFV)*	37	20	2500	137	28	18	1300000	132	LDL	0.87
3	46	F	mixed race	d4T, 3TC, NVP	7	2	18000	278	16	10	520	592	7.35	6.28
4	31	F	mixed race	AZT, 3TC, NVP	19	7	3300	367	21	15	94000	195	LDL	LDL
5	51	M	mixed race	d4T, 3TC, EFV	12	1	120000	115	9	7	320000	160	LDL	0.43
6	48	M	mixed race	d4T, 3TC, NVP	7	1	5400	131	21	14	91000	168	0.81	0.73
7	30	M	mixed race	d4T, 3TC, NVP	6	1	15000	102	12	7	300000	199	0.33	Not done

* Single drug switch from NVP to EFV in August 2007. Second line therapies: AZT: Zidovudine, DDI: Didanosine; LPV/r: Lopinavir with low dose ritonavir. CD4: most recent CD4 count in cells/µl while on the current regimen. VL: Viral load. Lopinavir plasma concentration: µg/ml. Lopinavir hair concentration: ng/mg hair. Failure period: Time VL>500 copies/ml. LDL: Lower than the detection limit.

2.3.2 In-house DRT summary

As expected, bulk sequencing detected multiple NNRTI DRMs when patients failed first-line cART. However, Sanger genotyping detected no PI-related DRMs after second-line failure, since patients showed poor adherence to the PI-based therapy, resulting in inadequate resistance selection pressure. Nevertheless, bulk sequencing detected remnants of the NNRTI DRMs, K103N (patient 5) and V179D (patient 7) at second-line failure. While the K103N detection was substantiated by its NGS prevalence (~55%), V179D was quantified at 6.12% of the sequenced sample. In addition, mutations at amino acid 103 (in *reverse transcriptase*) in patient 1 totalled 53% of the sampled population (44.66% K103R, 5.13% K103N and 3.21% K103E), however, they were undetected by bulk sequencing. These inconsistencies allude to possible primer-induced selection bias and PCR re-sampling error. The comparative NGS results are displayed in table 2.3.3.1 on the next page.

2.3.3 NGS summary

For NGS, 155 074 sequence reads obtained from the FLX Titanium (454 Life Technologies, Connecticut, USA) and the highest locus coverage was seen at the amplicon overlap between the *protease* (PR) and the first *reverse transcriptase* region (RTA), as indicated in figure 2.3.3.1 below. There was an average per-fragment coverage of approximately 9 400 reads or roughly 4 700 in both forward and reverse directions. The increased sensitivity characteristic of redundant sampling was able to detect variants present at less than 0.1% however, an arbitrary prevalence threshold of <0.5% was use to report drug-resistant minor variants.

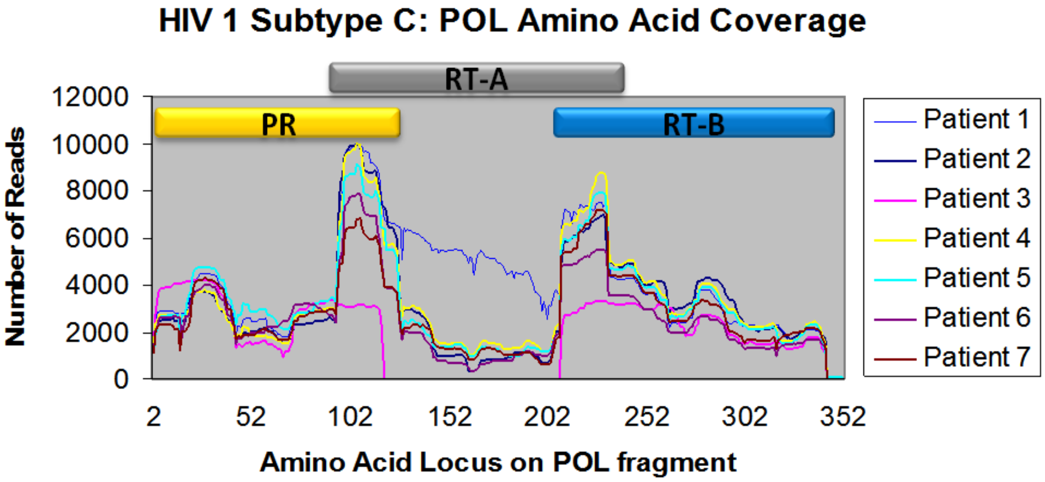


Figure 2.3.3.1 Coverage across HIV *pol*: The figure above shows the NGS coverage depth for all patients across the first 352 amino acids of HIV *pol.* The highest coverage is seen where amplicons overlap. There was no coverage across RT-A for patient 3 since the amplicon failed to amplify.

Table 2.3.3.1 Comparative in-house DRT and NGS DRT results

Patient	NNRTI failure episode		PI failure episode						
	Bulk sequencing (mutations)		Time on LPV/r	Viral load	Bulk sequencing (mutations)		NGS mutations (frequency)		
	Reverse transcriptase	*Protease*	(Days)	(copies/ml)	RT (NNRTI-only)	*Protease*	NRTI-mutations (percentage)	NNRTI mutations (percentage)	PI mutations (percentage)
1	A62V, M184I, *V108I, Y181C, H221Y*	None	280	2100000	None	None	K65R(1.11); D67N(0.93); D67E(0.86); K219R(0.71)	V90I(0.77); A98E(0.91); K101E(5.88); K103R(44.66); K103N(5.13); K103E(3.21); V179I(0.51); Y181C(0.58); F227L(0.82); F227S(0.57); K238R(0.50)	**I54T(0.50)**
2	M184V, *V106M*	None	841	1300000	None	None	K65R(3.81); D67N(8.39); F77L(0.73); M184V(8.02); L210S(0.69); T215I(0.77)	V90I(0.78); K101R(0.69); K103R(2.23); Y181C(1.93); F227S(0.51)	L23P(0.51)
3	M184V, *V90I, K103N, Y181C*	None	476	520	None	None	V118A(0.54); K219E(0.54)	V179I(5.78)	**I54M(0.67)**
4	M184V, *V108I, Y181C, H221Y*	None	616	94000	None	None	K65R(2.73); D67N(1.51); F116S(0.53); M184V(2.55)	K101E(0.61); K101R(0.61); P225T(1.50); F227L(0.55); K238T(3.31); K238R(0.74)	M46V(0.67); **F53L(0.56)**; F53S(0.51)
5	M184V	None	266	320000	K103N	None	K65R(1.59); K65E(0.53); D67N(2.60); M184V(3.14)	K101E(1.34); K101R(0.63); K103N(55.15); V179D(0.98); P225T(0.97); F227L(0.63); F227S(0.56); K238T(2.91)	**F53L(0.73)**; **N88S(0.68)**; **N88D(0.58)**
6	M184V, *K103N*	None	617	91000	None	None	K65R(1.09); K65E(0.55); K219E(0.50)	K103E(0.81); G190E(0.69)	**V82A(0.62**)
7	M41L, K65R, V75I, M184V, *K103R, V179D*	None	366	300000	V179D	None	K65R(2.19); T215A(0.69); K219R(0.70); K219E(0.51)	V90I(1.04); V179D(6.12)	None

NNRTI failure episode - NRTI mutations are in standard typescript and NNRTI mutations in *italics*.
NGS mutations - PI resistance mutations are in **bold** and other amino acid variants at PI resistance loci in standard type-script

NGS confirmed the bulk sequencing results and detected minor PI-resistant population undetected by the bulk sequencing (by PCR and Sanger sequencing), in five of the seven patients. Moreover, NGS was able to detect the remnants of lamivudine resistance (M184V/I) in patient 4, nevirapine/efavirenz resistance (K103N/E/R) in patients 1, 2, 5 and 6, and nevirapine/efavirenz/etravirine resistance in patients 1 and 2 (Y181C). The complete list of mutations detected by both bulk sequencing and NGS are summarized in table 2.3.3.1 above.

2.4 Discussion

2.4.1 Summary findings

The present study used amplicon sequencing on the 454's next generation sequencing (NGS) platform to detect drug-resistant minority HIV populations in patients who failed the first-line non-nucleoside reverse transcriptase inhibitor (NNRTI)-based combination antiretroviral therapy (cART) with M184V/I and second-line protease inhibitor (PI)-based cART. Second-line failure NGS drug resistance testing (DRT) results were compared to bulk sequencing results (available at both failure events), where viral load (VL) and CD4 cell count data were also available. Six of the seven patients included in this study had very low plasma- and hair-lopinavir concentrations (<0.81 µg/ml and 0/87 ng/ml respectively) and VLs >90 000 c/ml blood. Patient 3 had mid-to-high lopinavir adherence that was confirmed by high pressure liquid chromatography performed on this patient's plasma and hair samples extracts. Despite patient 3's a low VL of 520 c/ml and the absence of major PI resistance, remnants of first-line therapy NRTI and second-line NNRTI therapy resistance were identified by NGS.

Bulk sequencing detected no major PI DRMs at second-line failure while NGS detected only minority population with major PI DRMs (<0.7% of the sequenced sample). After second-line failure, bulk sequencing identified K103N in patient 5, from a population that NGS quantified at ~55%. However, bulk sequencing failed to identify K103R in patient 1, a variant that was quantified at 44.66% of the sequenced sample. Surprisingly, bulk sequencing also detected V179D in patient 7, a variant that contributed only 6.12% of that patient sample and was below the conventional Sanger sequencing detection threshold of ~20% (Palmer et al. 2005).

2.4.2 Resistance evolution

Resistance evolution begins with primary mutations which appear soon after the onset of cART (Bangsberg et al. 2006). Considering protease inhibitors, primary DRMs confer drug resistance at a fitness cost. In patients with intermittent adherence, wild-type virus may outgrow these less fit variants with the result that minor drug-resistant variants would not be detectable by HIV drug resistance testing through PCR and Sanger 'bulk' sequencing. If viral replication continues on the current regimen, compensatory mutations could arise through successive mutations or through recombination events, which would result in viral strains with drug resistance and improved fitness (Maguire et al. 2002; Rosenbloom et al. 2012).

The emergence of multi-drug-resistant viral strains requires ongoing viral replication in the presence of therapy and involves a number of mechanisms: a high mutation rate (due to reverse transcriptase lacking proofreading), enrichment of resistant variants through selection pressure (exerted by therapy), and viral recombination events of variants harboring different mutations. For example: Under first-line cART selection pressure (Tenofovir, Lamivudine and Efavirenz), viral strains with an evolutionary survival advantage (e.g.: a single DRM) are selectively enriched. These variants have characteristic single base-pair mutations such as K65R, M184V, V106M or K103N, which can arise sequentially or through recombination with other single mutation variants. At the time of first-line failure, two or more of these mutations are often detected by bulk sequencing as major variant drug resistance mutations.

2.4.3 Low frequency minor variants

The low frequency variants identified in this study are probably single mutation variants that lack the compensatory or secondary mutations needed to improve replication competency. They do not predominate as they experience inadequate drug pressure to favor a specific variant. Under intermittent selection pressure, they lack the survival benefit with their compromised fitness, resulting in an insufficient resistant population size to allow recombination and multi-drug-resistant variants to emerge. In addition, more resent research by Swenson et al, suggests that the genotypic detection of drug resistance mutations at low frequencies are predictive of therapy failure (Swenson et al. 2014).

2.4.4 Therapy failure and adherence

For low genetic barrier first-line regimens, resistance develops quickly under intermittent therapy adherence which results in insufficient blood-drug concentrations to effectively suppress viral replication (Luber 2005). Upon failure, patients are switched to the second-line therapy with a higher genetic barrier. At the time when we conducted this study, second-line cART consisted of two NRTIs (one of which was didanosine or DDI), and the boosted PI, LPV/r. Both of these ARVs have gastrointestinal side effects and consequentially, the second-line therapy is poorly tolerated to the extent that individuals who fail on this regimen did so as a result of poor adherence rather than viral resistance. For the patient group sampled in this study, adherence was determined by measuring plasma and hair drug concentrations and was seen to be less than adequate (van Zyl et al. 2011). These findings were in concordance with a previous study that correlated low PI-hair concentrations with therapy failure (Gandhi et al. 2009). Poor adherence and the detection of minor variant drug resistance are both predictive of therapy failure, with the hazard ratio for failure in patients with minor variant drug resistance, versus without, the highest in adherent patients (Li et al. 2012).

2.4.5 Sanger DRT

The routine diagnostic DRT uses low fidelity PCR which limits the accuracy of DRM detection in homopolymer regions. Bulk sequencing is however capable of sampling and identifying majority drug-resistant variant populations greater than 20% of the sequenced sample. In 30 to 40% of second-line failures, DRMs are not detected by Sanger DRT since failure, in these cases, is due to poor therapy adherence (El-Khatib et al. 2010; van Zyl et al. 2011). Consequently, the inadequate selection pressure maintains these variants below the assay sensitivity threshold.

2.4.6 454 NGS and DRT

As the first commercially available massive parallel sequencing platform, 454 Life Sciences platforms (now distributed by Roche Diagnostics, Basel, Switzerland) are well established in amplicon sequencing for HIV DRT (Simen et al. 2014). In 2005, the 454 Genome Sequencer 20 had an introductory read length of up to 100bp which improved to ~450bp through bidirectional sequencing on the FLX Titanium system (454 Life Sciences, Connecticut, USA) at the time of our study. While running costs for the 454 pyrosequencing platforms are high in comparison to other commercially available NGS platforms (Loman et al. 2012), the use of sample molecular identifiers (MIDs) incorporated into the fusion primers, facilitated sample pooling and reductions in per sample costs. However, the advantage of the FLX platform was the relatively long read length (compared to other NGS platforms at this time), allowing for mutation linkage study. As seen in the present study, the median read length was only 278 bp after quality filtering, although DRM linkage detection was improved through long read-length and an overlapping amplicon approach.

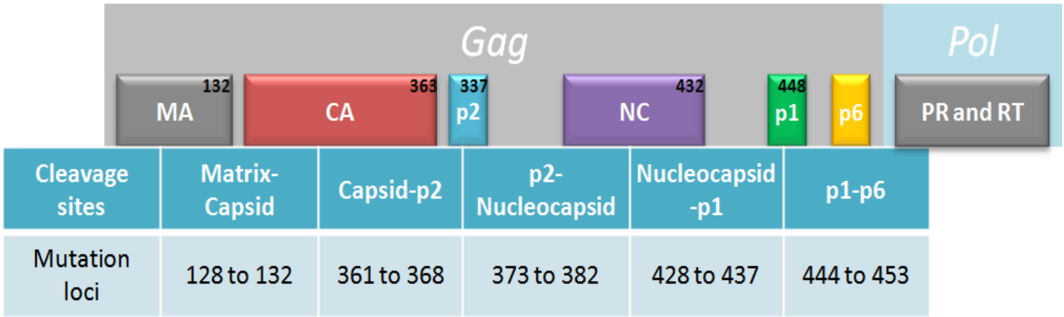
Hindering the utility of the 454 NGS platforms however, is the inability to accurately sequence homopolymer regions, a fault owing to the sequencing chemistry and detection methods. On 454 sequencers, bases are sequentially washed over the picotitre plate and the luminescence amplitude resulting from base incorporations is proportional to the number of incorporated bases (Morey et al. 2013). This luminescence recorded by a charge coupled device or CCD camera and the intensity is translated into the number of bases incorporated (the specific base that was washed over the flow-cell when the image was captured). For example: The luminescence signal amplitude for two incorporated bases is double that of one incorporated base. However, when more than five bases are incorporated at once, the change in luminescence is incremental and difficult for the CCD camera to distinguish, insomuch that homopolymer lengths are sometimes over- or under-called, presenting as a base insertion or deletion (collectively referred to as “indels”) (Marinier et al. 2015).

Misalignment due to spurious indels may appear in 454 sequences and could result in the false detection of DRMs which are not present, since 17 known DRMs fall within or just after homopolymer regions in HIV-1 subtype B (Dudley et al. 2012). Considering K65R in subtype C viruses, this DRM occurs at the end of a homopolymer stretch, where misalignment due to spurious “indels” could result in the spurious detection of K65R (Varghese et al. 2010). The high rate of emergence of K65R in HIV-1 subtype C has been ascribed to a high frequency of enzyme dissociation at this site resulting in a higher rate of mutation; this has been shown to be dependent on the particular RNA template sequence in HIV subtype C (Wainberg et al. 2011; Coutsinos et al. 2009). When detected with NGS, K65R could either be the result of a true high occurrence in HIV-1 subtype C or due to homopolymer sequencing error. Regardless of this DRMs origin, the use of high fidelity polymerase systems is advised for upstream NGS sample preparation since it improves accuracy in homopolymer amplification (McInerney et al. 2014). In our study, we used a high fidelity PCR system for our nested PCR however, our two-step reverse transcription/pre-nested PCR (RT-PCR) was performed with a low fidelity enzyme system.

2.4.7 Compensatory mutations in *gag*

Due to the high genetic barrier of second-line protease inhibitor regimens, drug resistance resulting in regimen failure often require multiple drug resistance mutations and compensatory mutations to yield a virus that is sufficiently resistant and fit to replicate. The evolution of drug resistance to high genetic barrier regimens is an intricate process (Flynn et al. 2015). In addition to the resistance mutations in *protease*, second-line failure may be associated with compensatory mutations in the *gag* gene. Other researchers have documented the co-evolution of *protease* and *gag* (Fun et al. 2012), suggesting that during viral replication, a less fit drug-resistant protease requires compensatory mutation in its substrates found in *gag*, which the viral protease cleaves from the gag-pol poly-protein (van Maarseveen & Boucher 2006).

Since most protease inhibitors have similar functions and target the substrate binding region of the viral protease, primary resistance mutations occurring in this region enable protease functionality at a great fitness cost (Nijhuis et al. 2007). The secondary or compensatory mutations arise in the *gag-pol* poly-protein cleave sites restoring viral fitness (Flynn et al. 2015). These compensatory mutations are seen around the matrix (MA) -capsid (CA), CA-p2, p2-nucleocapsid (NC), NC-p1 and p1-p6 cleavage sites depicted in figure 2.4.7.1 below.



Cleavage sites	Matrix-Capsid	Capsid-p2	p2-Nucleocapsid	Nucleocapsid -p1	p1-p6
Mutation loci	128 to 132	361 to 368	373 to 382	428 to 437	444 to 453

Figure 2.4.7.1 Compensatory mutations in *gag*: The figure above depicts the gag-pol poly-protein with the amino acid number where the viral protease cleaves the protein subunit, seen in black. In the table below the diagram are the loci between which PI therapy failure compensatory mutations are known to occur. (Figure adapted from Fun 2012 and Flynn 2015)

Under intermittent selection pressure (as seen in this cohort), single mutation drug-resistant variants may wax and wane as conditions for their selection are created during periods of improved adherence, followed by replacement by wild-type during periods of poor adherence (Rosenbloom et al. 2012). This would prevent drug resistance evolution and is supported by models which suggest that the highest risk for protease inhibitor resistance is at levels of moderate to high levels of adherence (Gardner et al. 2009). At a low frequency, single drug-resistant variants may be more likely to recombine with wild-type virus than with other strains carrying DRMs and compensatory mutations and would remain below the expected 20% sensitivity threshold of bulk sequencing (Palmer et al. 2005). Since the development of minor variant drug resistance in patients failing a second-line, LPV/r-based regimen is poorly characterized we set out to investigate this with NGS. Our findings indicate a lack of HIV drug resistance evolution in a poorly adherent cohort receiving this high genetic barrier, second-line regimen.

2.4.8 Study limitations

This study had the following limitations.

- Specimens were only available to perform deep sequencing at the time of second-regimen failure. Additional samples for NGS before therapy initiation and at first-line therapy failure would have enabled a more informative investigation of drug resistance development.
- The use of high fidelity PCR is advised for up-stream sample preparation for NGS (McInerney et al. 2014) as spurious PCR-induced errors would be detected with the high resolution of NGS (Brandariz-Fontes et al. 2015). In this study, low fidelity PCR was used for the pre-nested PCR which limited the accurate identification of minor variants.

- cDNA was not quantified prior to sequencing: While viral load (VL) data was available for study patients at the time of second-line failure and NGS, determining the number of cDNA species sampled is essential in estimating the limit of minor variant sampling. Low viral loads increase the minor variant threshold. Patient 3 had a low viral load of 520 copies/ml and was therefore more subject to PCR resampling error.
- Specific primers may result in biased amplification: While the RT-PCR priming strategy used here does improve reverse transcription sensitivity, the use of specific primers may result in a sampling bias, which could influence the downstream representation of variants in the sequenced population.
- The read-length was relatively short: The mean sequencing read-length seen in this study was 278 bases. There was great reduction in coverage distal to the primers, characterised by a steep drop in base-call confidence for longer reads. With longer read-lengths, coverage would have been better at the overlapping regions and the distal ends of the first and last amplicon.

2.5 Conclusion

This study compared in-house DRT by PCR and Sanger sequencing to NGS for detection of drug resistance in patients failing a second-line LPV/r-based regimen. PCR and Sanger sequencing (bulk sequencing) detected no major PI DRMs at second-line failure while NGS detected only minority population (<0.7% of the sequenced sample). Bulk sequencing identified K103N in patient 5, from a population that NGS quantified at ~55%. However, bulk sequencing failed to identify K103R in patient 1, a variant that was quantified at 44.66% of the sequenced sample. Surprisingly, bulk sequencing also detected V179D in patient 7, a variant that contributed only 6.12% of that patient sample and below the conventional Sanger sequencing detection threshold of ~20% (Palmer et al. 2005).

Further investigations of protease inhibitor resistance in patients who fail second-line regimens are required. Primary mutations in *protease* might confer resistance but require secondary compensatory mutations to regain fitness (Fun et al. 2012). The study of co-evolution of *gag* and *envelope* may share more light on protease inhibitor resistance development (Rabi et al. 2013; Flynn et al. 2015).

Novel methods may also improve the overall accuracy of NGS. The introduction of the primer ID approach to targeted NGS resequencing has enabled the indexing of individual cDNA species sampled prior to PCR enrichment. Two variations of the primer ID method have been published (Jabara et al. 2011; Boltz et al. 2015); this method collapses all reads with the same random identifier into one and thereby allows for bioinformatic correction of resampling or PCR replication error.

In this investigation NGS improved the DRM detection in five of the seven patients included in our study, although no major variants with protease inhibitor resistance were identified. This suggests that, in the absence of sustained selection pressure, the evolution of variants with multiple PI resistance mutations is halted at an early stage, whereas the potency of the regimen suppresses viral replication in patients with good adherence. This lack of evolution may help to explain the surprisingly low prevalence of protease inhibitor resistance, in previously PI-naïve patients, failing LPV/r-based regimens.

2.6 References

Archer, J. et al., 2010. The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time-an ultra-deep approach. *PLoS computational biology*, 6(12), p.e1001022.

Bangsberg, D.R. et al., 2006. Adherence-resistance relationships for protease and non-nucleoside reverse transcriptase inhibitors explained by virological fitness. *AIDS (London, England)*, 20(2), pp.223–231.

Boltz, V.F. et al., 2015. Analysis of Resistance Haplotypes Using Primer IDs and Next Gen Sequencing of HIV RNA Methods. *CROI*, p.1. Available at: http://www.croiconference.org/sites/default/files/posters-2015/593.pdf [Accessed July 19, 2015].

Brandariz-Fontes, C. et al., 2015. Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results. *Scientific reports*, 5, p.8056.

Clavel, F. & Hance, A.J., 2004. HIV drug resistance. *New England Journal of Medicine*, 350(10), pp.1023–1035.

Coutsinos, D. et al., 2009. Template usage is responsible for the preferential acquisition of the K65R reverse transcriptase mutation in subtype C variants of human immunodeficiency virus type 1. *Journal of virology*, 83(4), pp.2029–2033.

Dudley, D.M. et al., 2012. Low-cost ultra-wide genotyping using Roche/454 pyrosequencing for surveillance of HIV drug resistance. *PloS one*, 7(5), p.e36494.

El-Khatib, Z. et al., 2010. Viremia and drug resistance among HIV-1 patients on antiretroviral treatment: a cross-sectional study in Soweto, South Africa. *AIDS (London, England)*, 24(11), pp.1679–1687.

Flynn, W.F. et al., 2015. Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of correlated mutations in Gag and protease. *PLoS computational biology*, 11(4), p.e1004249.

Fun, A. et al., 2012. Human Immunodeficiency Virus gag and protease: partners in resistance. *Retrovirology*, 9, p.63.

Gandhi, M. et al., 2009. Protease inhibitor levels in hair strongly predict virologic response to treatment. *AIDS (London, England)*, 23(4), pp.471–478.

Gardner, E.M. et al., 2009. Antiretroviral medication adherence and the development of class-specific antiretroviral resistance. *AIDS (London, England)*, 23(9), pp.1035–1046.

Gulick, R.M. et al., 1997. Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy. *The New England journal of medicine*, 337(11), pp.734–739.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Li, J.Z. et al., 2012. Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS* (London, England), 26(2), pp.185–192

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–439.

Luber, A.D., 2005. Genetic Barriers to Resistance and Impact on Clinical Response. *Journal of the International AIDS Society*, 7, p.69.

Van Maarseveen, N. & Boucher, C., 2006. Resistance to protease inhibitors. In A. M. Geretti, ed. London.

Maguire, M.F. et al., 2002. Changes in Human Immunodeficiency Virus Type 1 Gag at Positions L449 and P453 Are Linked to I50V Protease Mutants In Vivo and Cause Reduction of Sensitivity to Amprenavir and Improved Viral Fitness In Vitro. *Journal of Virology*, 76(15), pp.7398–7406.

Marinier, E., Brown, D.G. & McConkey, B.J., 2015. Pollux: platform independent error correction of single and mixed genomes. *BMC bioinformatics*, 16, p.10.

McInerney, P., Adams, P. & Hadi, M.Z., 2014. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Molecular biology international*, 2014, p.287430.

Morey, M. et al., 2013. A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism*, 110(1-2), pp.3–24.

National Department of Health, 2010. *The South African Antiretroviral Treatment Guidelines*, Available at: http://apps.who.int/medicinedocs/documents/s19153en/s19153en.pdf

Nijhuis, M., van Maarseveen, N.M. & Boucher, C.A.B., 2007. HIV protease resistance and viral fitness. *Current opinion in HIV and AIDS*, 2(2), pp.108–115.

Palella, F.J.J. et al., 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *The New England journal of medicine*, 338(13), pp.853–860.

Palmer, S. et al., 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *Journal of clinical microbiology*, 43(1), pp.406–413.

Plantier, J.-C. et al., 2005. HIV-1 resistance genotyping on dried serum spots. *AIDS (London, England)*, 19(4), pp.391–397.

Rabi, S.A. et al., 2013. Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *The Journal of Clinical Investigation*, 123(9), pp.3848–3860.

Rosenbloom, D.I.S. et al., 2012. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. *Nature Medicine*, 18(9), pp.1378–1385.

Rosenblum, M. et al., 2009. The Risk of Virologic Failure Decreases with Duration of HIV Suppression, at Greater than 50% Adherence to Antiretroviral Therapy. *PLoS ONE*, 4(9), p.e7196.

Simen, B.B. et al., 2014. An international multicenter study on HIV-1 drug resistance testing by 454 ultra-deep pyrosequencing. *Journal of virological methods*, 204, pp.31–37.

Swenson, L.C. et al., 2014. HIV Drug Resistance Detected During Low-Level Viremia Is Associated with Subsequent Virologic Failure. *AIDS*, 28(8), pp.1125–1134.

Varghese, V. et al., 2010. Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. *PloS one*, 5(6), p.e10992.

Wainberg, M.A., Zaharatos, G.J. & Brenner, B.G., 2011. Development of antiretroviral drug resistance. *The New England journal of medicine*, 365(7), pp.637–646.

World Health Organisation, 2010. *Antiretroviral therapy for HIV infection in adults and adolescents - Revision 2010*, Available at: http://www.who.int/hiv/pub/arv/adult2010/en/index.html.

Van Zyl, G.U. et al., 2011. Low lopinavir plasma or hair concentrations explain second line protease inhibitor failures in a resource-limited setting. *Journal of acquired immune deficiency syndromes (1999)*, 56(4), pp.333–339.

2.7 Appendix B: Supplementary data

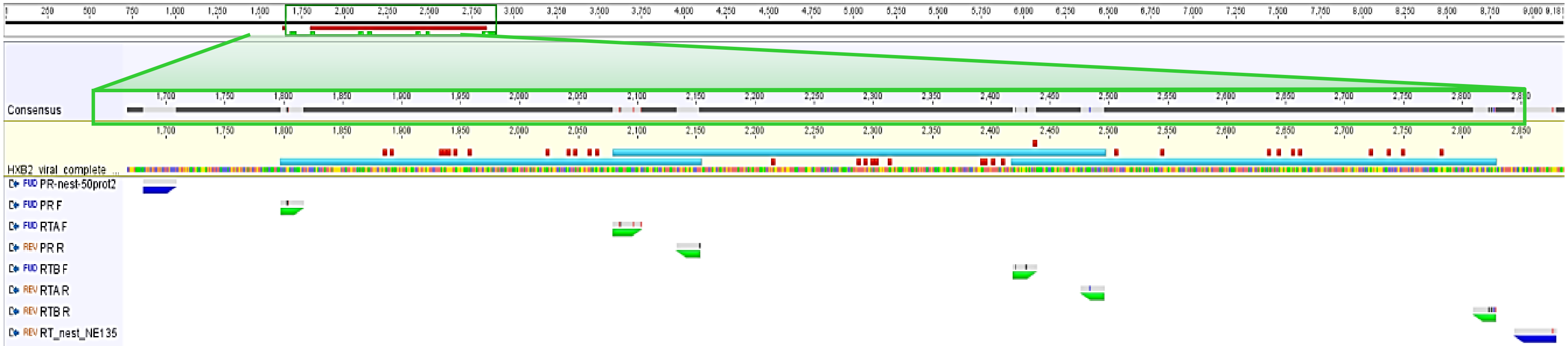


Figure 2.7.1 Schematic overview of NGS amplicon generation: This figure was generated in Geneious V6.0 (Biomatters) by mapping the pre-nested primers and the gene-specific regions of the fusion primers to the HXB2 reference sequence (seen at the top of the figure). It displays the overlapping amplicon generation strategy on the region of the viral genome enriched by the pre-nested PCR (green-boarded window). The pre-nested primers (dark blue) flank the nested fusion primers (green) and the three overlapping amplicons (light blue). The amplicon on the left amplifies the entire *protease* region (PR), and amplicons RT-A and RT-B amplify the first ~350 amino acids of *reverse transcriptase*, with RT-A overlapping with the 3'end of the PR amplicon and the 5'end of the RT-B amplicon.

Table 2.7.1 DRMs distribution: The table below sequentially lists the DRMs sampled and the amplicon on which they appear.

Amplicons	Drug resistance mutations															
PR	K20R	L23P	M36I/L	M46V	F53L/S	I54T/M	D60E	I62V	L63P/S	A71T	T74S	V82A	I84V	N88S/D	L89M	I93L
RT-A	K65R/E*	D67N*	F77L*	V90I#	A98E#	K101E/R#	K103R/N/E#	F116S*	V118A*							
RT-B	V179I/D#	Y181C#	M184V*	G190E#	L210S*	T215I/A*	K219R/E*	P225T#	F227L/S#	K238R/T#						

* NRTI resistance mutations # NNRTI resistance mutations

Table 2.7.2 Forward fusion primers: Tabulated below are the schematics of the forward fusion primers. At the 5' end is the **bolded** adaptor-A sequence, followed by the underlined key sequence, the *italicised* molecular identifier (MID) and finally, the gene-specific region at the 3' end. The "Adaptor-A" sequence mediates the binding of sequence amplicons to beads via a bead-bound adaptor compliment, after which the amplicon is enriched through bead-based emulsion PCR, as part of NGS library preparation. The "KEY" sequence is the first four bases read by the pyrosequencer to determine the baseline luminescence intensity for each picotitre sequence well in the 454 sequencing flow-cell.

NAME	5'- Adaptor-A sequence - KEY sequence - *MID* - Template-specific-sequence -3'	HXB2 binding	Codon
PR-F-MID-1	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ACGAGTGCGT*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-2	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ACGCTCGACA*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-3	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*AGACGCACTC*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-4	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*AGCACTGTAG*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-5	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ATCAGACACG*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-6	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ATATCGCGAG*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
PR-F-MID-7	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*CGTGTCTCTA*-CCTCARATCACTCTTTGGC	2253→2271	PR1-PR7
RT-A-F-MID-1	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ACGAGTGCGT*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-2	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ACGCTCGACA*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-3	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*AGACGCACTC*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-4	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*AGCACTGTAG*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-5	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ATCAGACACG*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-6	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*ATATCGCGAG*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-A-F-MID-7	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-*CGTGTCTCTA*-TGCACAYTAAATTTTCCAATTAGa	2535→2557	PR95-RT3
RT-B-F_MID-1	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-ACGAGTGCGT-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-2	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-ACGCTCGACA-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-3	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-AGACGCACTC-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-4	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-AGCACTGTAG-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-5	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-ATCAGACACG-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-6	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-ATATCGCGAG-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114
RT-B-F_MID-7	5'-**CGTATCGCCTCCCTCGCGCCA**TCAG-CGTGTCTCTA-CTRGATGTGGGRGATGCAta	2874→2891	RT109-114

Bases added to improve 3' conservation with South African subtype C are in lower case. PR = *protease.* RT= *reverse transcriptase*

Table 2.7.3 Reverse fusion primers: Below are the schematics of the reverse fusion primers. At the 5' end is the **bolded** adaptor-B sequence, followed by the underlined key sequence, the *italicised* molecular identifier (MID) and finally, the reveres-complemented gene-specific region at the 3' end. The "Adaptor-B" sequence mediates the binding of sequence amplicons to beads for bead-based emulsion PCR. The "KEY" sequence is the first four bases read by the pyrosequencer to determine the baseline luminescence intensity for each picotitre sequence well.

NAME	5'- Adaptor-B sequence - KEY sequence - *MID* - Template-specific-sequence -3'	HXB2 binding	Codons
PR-R-MID-1	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGAGTGCGT*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-2	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGCTCGACA*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-3	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGACGCACTC*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-4	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGCACTGTAG*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-5	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATCAGACACG*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-6	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATATCGCGAG*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
PR-R-MID-7	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*CGTGTCTCTA*-YTTGGGCCATCCATTCCTGG	2608→2589 (RC)	RT14-RT20
RT-A-R-MID-1	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGAGTGCGT*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-2	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGCTCGACA*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-3	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGACGCACTC*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-4	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGCACTGTAG*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-5	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATCAGACACG*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-6	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATATCGCGAG*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-A-R-MID-7	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*CGTGTCTCTA*-ACTAGGTATGGTGAATGCAG	2951→2932 (RC)	RT128-RT134
RT-B-R-MID-1	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGAGTGCGT*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-2	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ACGCTCGACA*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-3	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGACGCACTC*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-4	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*AGCACTGTAG*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-5	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATCAGACACG*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-6	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*ATATCGCGAG*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245
RT-B-R-MID-7	5'-**CTATGCGCCTTGCCAGCCCGC**TCAG-*CGTGTCTCTA*-TGTWTWGGCTGTACTGTCC	3284→3265 (RC)	RT239-RT245

PR = *protease.* RT= *reverse transcriptase* (RC) = Reverse complement

2.8 Published article

JVI
Journals.ASM.org

Deep Sequencing Reveals Minor Protease Resistance Mutations in Patients Failing a Protease Inhibitor Regimen

Randall Fisher,[a] Gert U. van Zyl,[a,b] Simon A. A. Travers,[c] Sergei L. Kosakovsky Pond,[d] Susan Engelbrecht,[a,b] Ben Murrell,[e,f] Konrad Scheffler,[e] and Davey Smith[d]

Division of Medical Virology, Stellenbosch University, Stellenbosch, South Africa[a]; National Health Laboratory Service, Coastal, Tygerberg, South Africa[b]; South African National Bioinformatics Institute, University of the Western Cape, Bellville, South Africa[c]; Department of Medicine, Division of Infectious Diseases, University of California, San Diego, San Diego, California[d]; Computer Science Division, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Aftrica[e]; and Biomedical Informatics Research Division, eHealth Research and Innovation Platform, Medical Research Council, Tygerberg, South Africa[f]

Standard genotypic antiretroviral resistance testing, performed by bulk sequencing, does not readily detect variants that comprise <20% of the circulating HIV-1 RNA population. Nevertheless, it is valuable in selecting an antiretroviral regimen after antiretroviral failure. In patients with poor adherence, resistant variants may not reach this threshold. Therefore, deep sequencing would be potentially valuable for detecting minority resistant variants. We compared bulk sequencing and deep sequencing to detect HIV-1 drug resistance at the time of a second-line protease inhibitor (PI)-based antiretroviral regimen failure. Eligibility criteria were virologic failure (HIV-1 RNA load of >500 copies/ml) of a first-line nonnucleoside reverse transcriptase inhibitor-based regimen, with at least the M184V mutation (lamivudine resistance), and second-line failure of a lopinavir/ritonavir (LPV/r)-based regimen. An amplicon-sequencing approach on the Roche 454 system was used. Six patients with viral loads of >90,000 copies/ml and one patient with a viral load of 520 copies/ml were included. Mutations not detectable by bulk sequencing during first- and second-line failure were detected by deep sequencing during second-line failure. Low-frequency variants (>0.5% of the sequence population) harboring major protease inhibitor resistance mutations were found in 5 of 7 patients despite poor adherence to the LPV/r-based regimen. In patients with intermittent adherence to a boosted PI regimen, deep sequencing may detect minority PI-resistant variants, which likely represent early events in resistance selection. In patients with poor or intermittent adherence, there may be low evolutionary impetus for such variants to reach fixation, explaining the low prevalence of PI resistance.

The full article is available online at http://jvi.asm.org/content/86/11/6231.full.pdf+html

3 Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure

3.1 Background

3.1.1 Prevention of mother-to-child transmission in South Africa

From 2004 to 2010, the human immunodeficiency virus (HIV) prevention of mother-to-child transmission (PMTCT) therapy in the Western Cape Province of South Africa consisted of a dual regimen of Zidovudine (AZT) and Nevirapine (NVP) (Draper & Abdullah 2008). When this study cohort was recruited (October 2006 to October 2009), pregnant women received AZT from their second trimester and single dose NVP (sdNVP) intrapartum. Neonates received sdNVP and AZT for one week and were formula fed to prevent postpartum HIV transmission. The HIV transmission rate was <10% during this period (Draper & Abdullah 2008). In 2010, the national guidelines replaced infant sdNVP with daily NVP for the first six weeks after birth causing the PMTCT-failure rate further declining to <3% (Barron et al. 2013).

Children, infected despite prophylactic antiretrovirals (ARVs) are at high risk of acquiring transmitted ARV drug resistance mutations (DRMs) (Ton & Frenkel 2013). Studies using allele-specific PCR (AS-PCR) demonstrated that even minority population variants containing non-nucleoside reverse transcriptase inhibitor (NNRTI) DRMs have been shown to affect NNRTI-containing regimen outcomes (Jourdain et al. 2010; Paredes et al. 2010; Li et al. 2012; Li et al. 2013). In South Africa, however, PMTCT-failed children under the age of three years receive a first-line ritonavir-boosted protease inhibitor regimen: Lopinavir (LPV) boosted with low dose ritonavir (LPV/r) (Palumbo et al. 2010; Violari et al. 2012). The prevalence of minor variant DRMs to NVP nevertheless remains important in resource-limited settings with little or no LPV/r access for infant cART or where NNRTIs are required in second-line regimens.

3.1.2 World Health Organisation Option B-Plus

In 2013, the Western Cape adopted the World Health Organisation (WHO) Option B-Plus (B+) for PMTCT (Western Cape Provincial Govt. 2012). B+ prescribes combination antiretroviral therapy (cART) for all HIV-positive pregnant women, continued after pregnancy, despite their CD4 cell count or WHO diseased state, (World Health Organisation 2012). The implementation of option B+ meant a reduction in pre-therapy initiation testing (omitting CD4 count testing), translating into a cost saving of about eight USD per patient initiated on therapy (Phillips et al. 2014). These savings exclude the additional financial benefits of HIV-negative neonates and immunocompetent pregnant women, not susceptible to opportunistic infections.

Option B+ has been shown to decrease mother-to-child transmission in Malawi (Sinunu et al. 2014), however, antiretroviral interventions have been linked to an increase in circulating drug resistance mutation accumulation (Sungkanuparph et al. 2012; Hattori et al. 2010). A recent manuscript highlighted the necessity of drug resistance testing (DRT) (van Zyl et al. 2014), whether by allele-specific quantitative PCR (AS-qPCR) (Rowley et al. 2010; Palmer et al. 2012), oligonucleotide ligation assay (OLA) (Jourdain et al. 2010) or *pol*-targeted genotyping (Claassen et al. 2011), whenever possible to ensure therapy success and to limit transmitted drug resistance (TDR) (Estill et al. 2013; Braithwaite et al. 2011).

3.1.3 HIV drug resistance testing

When HIV is transmitted despite maternal cART, there is a high risk that infants could be infected with viruses harbouring HIV DRMs. Transmitted DRMs are especially of concern in settings without access to LPV/r infant formulas (which require refrigeration) and where NVP-based regimens are therefore used for infant therapy. These concerns emphasize the importance of access to HIV DRT for surveillance of PMTCT failures and for clinical management (van Zyl et al. 2014). However, the current DRT assays are sensitive to primer mismatches (Rowley et al. 2010) and have other limitations:

- AS-qPCR and OLA are limited to specific individual point mutations and thus cannot detect or determine linkage.
- Sanger genotyping is quite expensive (300 USD according to Chaturdhuj (Chaturbhuj et al. 2014)), and is insensitive to variants contributing less than 20% of the sequenced sample (Palmer et al. 2005).

3.1.4 Next generation sequencing

By means of sample barcoding and pooling, next generation sequencing (NGS) potentially offers an affordable solution allowing for mutation linkage surveillance and in addition, is capable of identifying drug-resistant minor viral populations. Commercial NGS platforms available in South Africa include Life Technologies' (California, USA) Ion Torrent Personal Genome Machine™ (PGM) and Applied Biosystems' SOLiD™ sequencer (Massachusetts, USA), Roche Diagnostics' 454 GS Junior and FLX Titanium (454 Life Sciences, Connecticut, USA), and Illumina's MiSeq and HiSeq (California, USA) (http://omicsmaps.com/). While each sequencing platform has its own strengths and weaknesses concerning sequence read length, accuracy, depth and scalability, the PGM (Life Technologies, California, USA) boasts the lowest per base cost (Loman et al. 2012).

Common to both the 454 (454 Life Sciences, Connecticut, USA) and PGM (Life Technologies, California, USA) platforms is the way in which bases are sequentially washed over the picotitre plate (in the case of 454 Life Sciences) or sequencing chip (in the case of Life Technologies PGM). On 454 platforms, a high-resolution charge coupled device or CCD camera captures the light emitted as bases are incorporated into the complementary DNA strand, and the chemiluminescence intensity amplitude is translated into the number of incorporated bases (Morey et al. 2013). For example, when two bases are incorporated, luminescence is twice as bright as when one base is incorporated. However, when five or more bases are incorporated, the luminescence increase is incremental and difficult to discriminate. The PGM (Life Technologies, California, USA) monitors base incorporation by measuring pH change in wells, embedded in a sequencing chip, and struggles to differentiate the pH change between five or more base incorporations (Morey et al. 2013). Consequently, both 454 (454 Life Sciences, Connecticut, USA) or PGM (Life Technologies, California, USA) sequencing is prone to homopolymer sequencing error, characterised by spurious insertions or deletions at homopolymer regions (Marinier et al. 2015).

3.1.4.1 NGS Homopolymer error and HIV

Despite this disadvantage, Life Sciences' 454 platforms (Connecticut, USA) have been established for HIV DRT as the first platform used for targeted resequencing with fusion primers. Moreover, it initially provided the longest read-length which allowed the study of mutation linkage. Compared to another commercial NGS platform from Illumina (California, USA), the MiSeq, the PGM (Life Technologies, California, USA) has 400bp read lengths while the MiSeq, 300bp (using bi-directional sequencing in 2013). The MiSeq (Illumina, California, USA) however, is less prone to homopolymer sequencing error unlike the Life Technologies' PGM (California, USA) whose homopolymer errors may be corrected using post-sequencing bioinformatics software, such as Datamonkey (Kosakovsky Pond et al. 2009) and RAMICS (Wright & Travers 2014). These open-source online data analysis packages are able to map sequence reads to a reference sequence in a codon-aware manner: Once the sequence reads have been quality filtered, they are converted to an amino acid sequence and mapped to an amino acid reference sequence. Homopolymer regions with erroneous indels display a frame shift in the amino acid sequence. When these frame shifts are identified, a nucleotide interrogation-error correction protocol is initiated. Miscalled indels are corrected and the read, realigned.

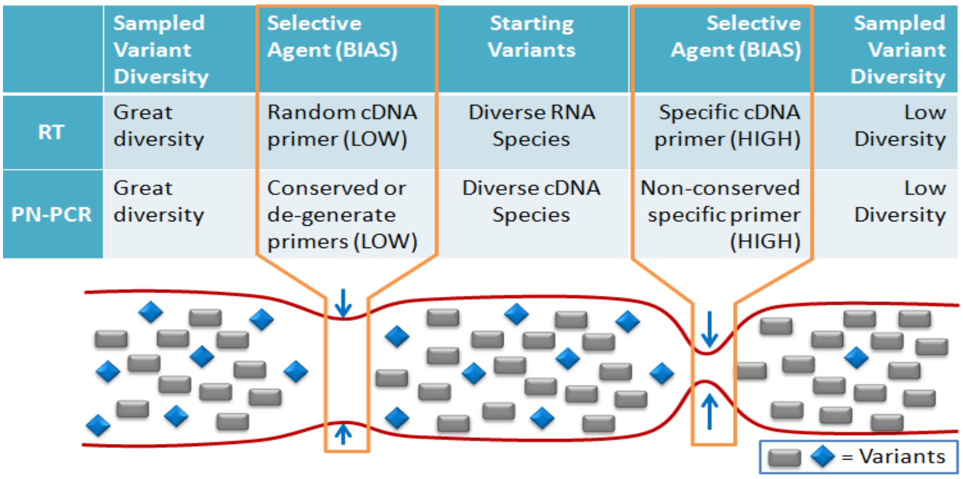
The combination of post-sequencing homopolymer error corrections and the low per-base sequencing cost of PGM (Life Technologies, California, USA) sequencing make it ideal for minor HIV variant sequencing of TDR before therapy initiation (baseline) and at therapy failure. To date, few studies have sequenced HIV using the Ion Torrent platform (Life Technologies, California, USA) (Kijak et al. 2014; Chang et al. 2013; Archer et al. 2012) but neither has validated their findings through clonal sequencing.

3.1.5 Sampling error

This study used a high fidelity PCR system to minimise PCR substitution and DNA gel stains that do not require ultraviolet light to protect DNA from damage. Moreover, the most crucial factor we tried to counteract was sampling error introduced during PCR enrichment. Sampling error is either primer-induced bias or random error, both of which obscure the variant prevalence, as depicted in figure 3.1.5.1 below. These errors have the greatest impact on variant representation when they occur early on in the NGS target enrichment process (e.g.: during reverse transcription (RT) or pre-nested PCR), and reduce the number of variants represented in downstream sequencing events. We addressed these errors separately below:

To avoid primer mediated selection bias during RT, cDNA synthesis was optimised using random pentadecamers instead of the conventional specific primer-RT-PCR combination. Although pentadecamers have a random nucleotide composition, they have similar thermodynamic properties as other 15 base-long specific oligonucleotides (i.e.: GC-rich 15-mer primers have a higher melting temperature (Tm) than AT-rich 15-mer primers). These properties were exploited to prime cDNA synthesis in a GC-rich “island” downstream from the pre-nested amplicon by using a high RT priming Tm. To prevent primer mediated bias during pre-nested PCR (PN-PCR), primers complementary to the more conserved areas of HIV-1 subtype C *reverse transcriptase* were used.

Random error is especially problematic with low frequency events and is predicted by the Poisson distribution: variance equals the mean of the distribution (therefore the coefficient of variation is highest for low frequency events). Practically, in the first few rounds of PCR, a particular randomly selected or primed template would become the most abundant at the endpoint. In any event, if a primer shows bias or randomly primes a particular template, once the primer is incorporated into the synthesised strand, the template has been mutated to perfectly complement the same primer species. This phenomenon is better known as PCR resampling (Liu et al. 1996). Particular template species are over-enriched with reference to other templates, so that the enriched population no longer represents the distribution of variants in the original template population. PCR resampling can be overcome through limiting dilution PCR (LD-PCR) (Ramachandran et al. 2008) or it can be corrected by random ID tagging of the original cDNA population with post-sequencing bioinformatics (Jabara et al. 2011).



	Sampled Variant Diversity	Selective Agent (BIAS)	Starting Variants	Selective Agent (BIAS)	Sampled Variant Diversity
RT	Great diversity	Random cDNA primer (LOW)	Diverse RNA Species	Specific cDNA primer (HIGH)	Low Diversity
PN-PCR	Great diversity	Conserved or de-generate primers (LOW)	Diverse cDNA Species	Non-conserved specific primer (HIGH)	Low Diversity

Figure 3.1.5.1 Low selection bias results in greater diversity: The figure above demonstrates the effect of sampling bias on sampled variant diversity. In reverse transcription, random RT primers reduce the selection bias and produce more diverse cDNA species, in contrast to a specific RT primer. Similarly, degenerate PN-PCR primers, or primers bringing in conserved genomic regions, improve the diversity available for further interrogation. Using highly-specific primers reduce diversity but are essential for certain applications such as Sanger genotyping.

While LD-PCR overcomes primer binding competition, it requires cDNA template dilution to one copy per PN-PCR and proves to be laborious and expensive when considering the total number of reactions needed. The cDNA ID approach compensates for PCR resampling by tagging each reverse transcribed species with a unique cDNA tag. This ID tag is incorporated into the sequence amplicon and, once sequencing is completed, bioinformatics is used to condense sequenced variants by their IDs. The sampled mutations are stacked and the consensus sequence is used for further analysis. While not completely compensatory, in this study a paralleled reaction approach was used by dividing two pentadecamer RT reactions into 14 PN-PCRs and subsequently pooling PN-PCR duplicates into seven nested PCRs.

3.1.5.1 Fusion primer sampling error

The selection bias introduced when using fusion primers was overcome with an alternative amplicon preparation method. Fusion primers are generally 50 to 60 bases long and consist of sequencing platform-specific adaptors, barcodes (or MIDs) and a gene-specific region. During PCR, the gene-specific region of the fusion primers binds to a template and the primer is incorporated into the synthesised DNA. Once complemented by the polymerase, this incorporated primer has generated the best match template for the same primer species in subsequent PCR cycles. This event effectively increases the required annealing temperature of the reaction, biases selection towards initially selected variants and reduces diversity. To compensate, we employ pre- and nested PCR to generate large sequence amplicons (approximately 1 kb) which are later randomly fragmented before sequence platform adaptors with sample indexing MIDs are ligated.

3.1.6 Study rationale

A recent study screening PMTCT-exposed children (less than two years of age) prior to cART initiation using AS-PCR and 454 NGS, found a good correlation between K103N and Y181C (Hunt et al. 2014). While the advantages of AS-PCR are many it is limited with respect to the number of mutations that can be detected simultaneously. This also limits its utility for DRM linkage studies, in contrast to NGS and other sequence-based genotyping methods. Here, we optimised sample preparation reactions while implementing the concepts described above. We not only corroborated our NGS results on another NGS platform (the MiSeq (Illumina, California, USA)) and by clonal sequencing, but we also conducted the first PGM (Life Technologies, California, USA) genotyping study reporting on HIV drug resistance from infants who had been exposed to the Western Cape (South Africa) dual regimen of AZT and NVP.

3.2 Methods

Deep sequencing of HIV has been well established on Life Sciences’ 454 platforms (Connecticut, USA) (Simen et al. 2014) but data on the Life Technologies’ (California, USA) Ion Torrent Personal Genome Machine™ (PGM) and Illumina MiSeq platforms (California, USA) had been limited (Archer et al. 2012). These new platforms provide higher coverage and the potential to pool many samples in one run and have recently started to displace 454 (Life Sciences, Connecticut, USA) sequencing for other applications. Similar to 454 (Life Sciences, Connecticut, USA) sequencing, the PGM (Life Technologies, California, USA) has been reported to be prone to homopolymer sequencing error. We therefore performed this study to:

- Detect minor variant resistance mutations in children exposed to the Western Cape prevention of mother-to-child HIV transmission (PMTCT) regimen and;
- Compare deep sequencing results on the PGM (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) platforms to clonal sequencing of PCR products, as reference.

We followed the following step-wise approach:

- Selected and collated suitable plasma samples from infants matching our inclusion criteria: PMTCT exposed and combination antiretroviral therapy-naïve. Historical plasma samples were stored at -20°C and had to be extracted before they could be processed in our study.
- We investigated different methods for reverse transcription and set out to adapt a previously published HIV reverse transcription protocol (Salazar et al. 2007) for cDNA synthesis with random primers, to limit sampling bias during this first step; as specific primers are known to introduce sampling bias when sampling diverse populations.
- We optimised pre- and nested PCRs and performed parallel PCR reactions in an effort to minimise sampling error.
- To validate our next generation deep sequencing results on PGM (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) platforms, we compared our results to clonal sequencing of the same PCR products.
- We corrected for homopolymer sequencing error using a post-sequencing, codon-aware bioinformatic analysis workflow.

The methods shown below demonstrate reaction optimisation and the rationale behind the optimisation steps taken to achieve the objectives listed above.

3.2.1 Reactions optimisations

3.2.1.1 Reverse transcription optimisation

The CHAVI–MBSC–2 protocol was specifically developed for the sequencing of the envelope gene of single HIV genomes (Salazar et al. 2007). This protocol uses a gene-specific primer to initiate cDNA synthesis (i.e.: OFM19) and thus selects for specific variants. As the use of specific primers during reverse transcription could result in biased cDNA synthesis (and skew the proportion of minor variants detected) we investigated the use of random primers. Reverse transcription kit manufacturers suggest the use of random hexamers to produce cDNA from the majority of RNA species present in the extracted sample. As a "promiscuous" primer species, hexamers compensates for the diversity and efficiently primes cDNA synthesis. In our initial reverse transcription optimisations, a high concentration of random hexamers appeared to yield more cDNA than specific cDNA primers, when the cDNA was quantified with an in-house qPCR (table 3.2.1.1.1 below).

Table 3.2.1.1.1 Quantification of cDNA: Below are the tabulated results of two in-house qPCRs, quantifying cDNA synthesised using either 2 µM of the CHAVI-prescribed HIV-1 RT-PCR primer (OFM19), or 697 pM random hexamers. In both experiments, template RNA was extracted from patient plasma that had been virologically quantified with a commercial viral load assay. While both primer species catalysed little of the input RNA to cDNA, random hexamers was more efficient than the proposed counterpart.

Input RNA Copy No.	RT-PCR Primer	Ct	Copy No.	% cDNA Synthesized
~480	R. Hex	37.4	3.54E1	7.4
	OFM 19	37.8	1.84E1	3.8
~650	R. Hex	36.5	1.14E2	17.5
	OFM 19	38.2	3.84E1	5.9

However, we observed that although a qPCR (targeting a short fragment) suggested a high cDNA yield, we could not amplify the cDNA using a larger fragment PCR. This dichotomy suggested that the cDNA largely consisted of short fragments, due to the initiation of primer extension at multiple sites along the RNA template during reverse transcription (figure 3.2.1.1.1 below). It is known that a high hexamer concentration is associated with the generation of short fragment cDNA due to frequent hybridisation to the RNA template with extension truncated at the next priming site due to the lack of 3' to 5' exonuclease activity of the reverse transcriptase.

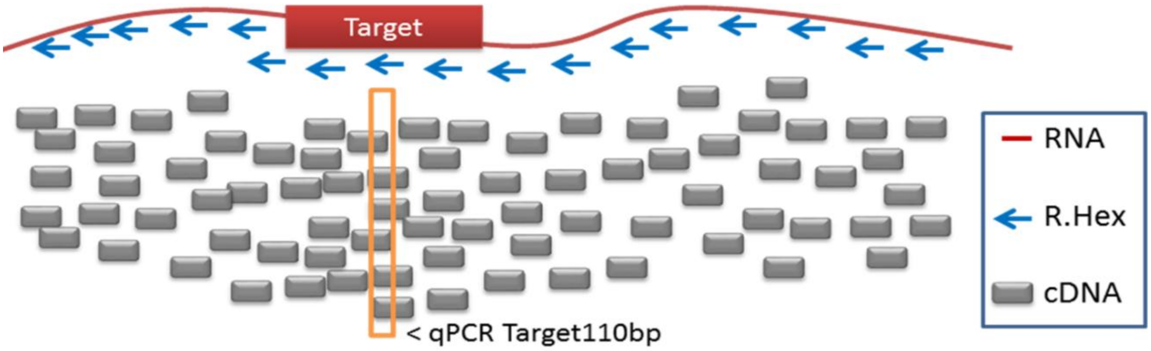


Figure 3.2.1.1.1 Random Hexamer – qPCR dichotomy: The figure above depicts our assumption of the random hexamer – qPCR dichotomy. If many viral RNA species (red) are hybridised by hexamers (blue arrows) in numerous locations, many short cDNA fragments are catalysed (grey rectangles) by the 3' to 5' exonuclease-deficient reverse transcriptase. Quantifying the cDNA species with a short fragment qPCR (orange), would yield a higher copy number than the actual number of cDNA templates that contain the full PCR target.

To produce more full length cDNA, reverse transcription kit manufacturers suggest titrating hexamers down to lower concentrations, ensuring less frequent reverse transcription priming. As part of the standard hexamer hybridisation protocol, they also advise linearizing RNA at 72 °C then allowing the hexamers to anneal at 25 °C, before adding the reverse transcriptase and reaction mixture. To this end we again synthesised cDNA using the modified CHAVI-MBSC-2 protocol and a variety of hexamer concentrations. We also decided to evaluate reaction efficacy using a 1.2 kb pre-nested PCR (PN-PCR), sensitive to 50 HIV DNA copies (figure 3.2.1.1.2 below), and gauge reverse transcription efficiency through endpoint band intensity, as seen in figure 3.2.1.1.3 below. The HIV-1 DNA control used for assay optimisation was the pZAC plasmid: an infectious clone containing a full HIV-1 subtype C genome isolated from a South African patient (Jacobs et al. 2012).

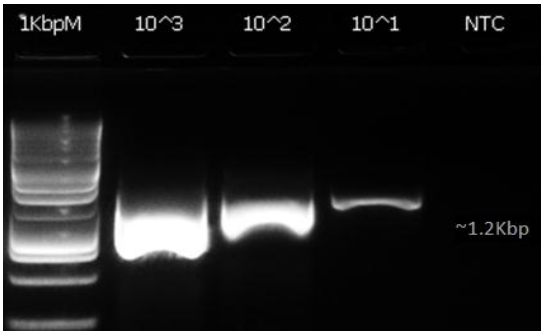


Figure 3.2.1.1.2 PN-PCR sensitivity: The figure depicts the results of reaction sensitivity testing for the 1.2 kbp PN-PCR amplicon, shown to be sensitive to 50 copies of the HIV-1 Subtype C plasmid, pZAC (lane 4 from the left). The experiment was validated by an NTC displaying no amplification. Also visible is the specific amplicon band, waning with the input copy number.

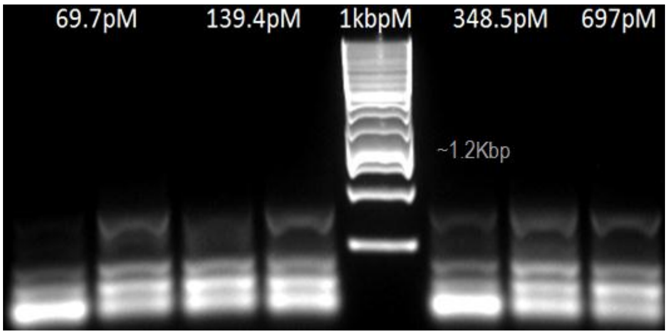


Figure 3.2.1.1.3 Unamplifiable cDNA: The figure above displays the gel electrophoresis result of the 1.2 kbp PN-PCR, amplifying the cDNA synthesised using various random hexamer concentrations, as listed above each lane. Bright, nonspecific bands are visible in all lanes below the 500 bp band (in relation to the 1 kbp DNA marker). However, no specific amplicons were produced, regardless of the hexamer concentration employed for cDNA synthesis.

Failing to generate the pre-nested amplicon, we decided to examine the possible molecular interactions between our random primers and HIV RNA and soon realised that in case of random primers, the location and frequency of priming events would be determined not only by the concentration of primers with a good template sequence match, but also the stability of the primer template complex during the reverse transcription reaction.

The melting temperature (Tm) is the temperature at which 50% of primers would be hybridized to the putative template as a template-primer complex. As temperatures increase above the Tm, these complexes are unstable and primer extension becomes increasingly unlikely. Random hexamers are single-stranded strings of six random nucleotides and, as an oligonucleotide, they exhibit a wide range of Tms. Using a neighbour-joining Tm calculator (available online at: https://www.exiqon.com/ls/Pages/ExiqonTMPredictionTool.aspx) revealed that hexamers consisting of alternating thymines (T) and adenosines (A) display a Tm of -27 °C, and hexamers consisting of guanine (G) and cytosine (C), a Tm of 35 °C.

In view of the manufacturer prescribed annealing temperature of 25 °C, GC-rich random hexamers would predominate cDNA priming. While a lower annealing temperature would allow more AT-rich hexamers to prime cDNA synthesis, lower temperatures also result in more RNA secondary structure and in turn, truncated cDNA. When examining South African HIV-1 subtype C consensus, we found regions with short stretches where at least five bases were Gs or Cs, as displayed in figure 3.2.1.1.4 below.

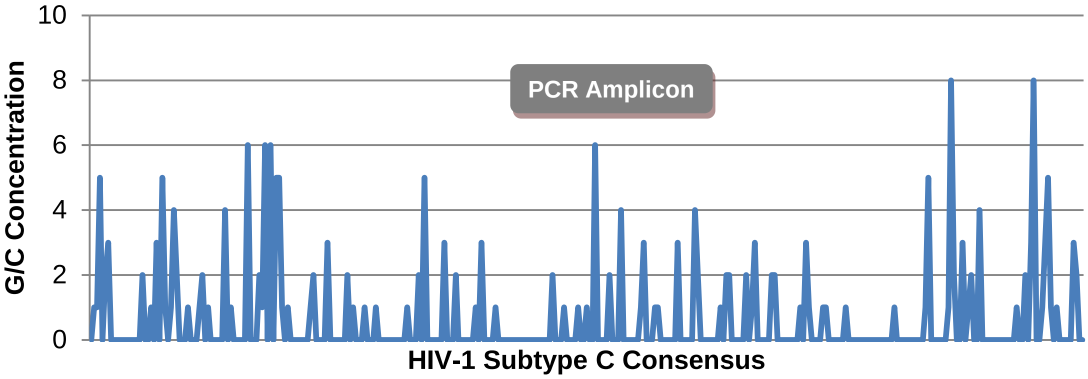


Figure 3.2.1.1.4 GC-rich islands in HIV-1 Subtype C, ZA: For analysis a six base sliding window was used. The window steps forward one base at a time, and a positive event (counted as 1) was defined as any time when the preceding six bases contained at least five G's or C's. To visualise GC rich areas, the number of positive events were plotted per 20 base window for nucleotides 800 to 7800 (Subtype C). The pre-nested PCR target that we used for sample enrichment, seen in grey above, spans a region of the genome that contains many G/C rich islands; which suggest multiple hexamer priming cites within this template.

Considering the binding energy of the matching random hexamers, our PCR target contained a region where a large proportion of hexamers would bind with sufficient stability to allow primer extension at the prescribed temperature of 25 °C used to prime the reaction. This region was also closely downstream from our qPCR target (2981 to 3133 on HXB2). The binding of hexamers inside of our target could have explained the dichotomy observed. We therefore investigated random pentadecamers and a higher annealing temperature (with an expected higher specificity) to prime reverse transcription as this would bias binding to another region with larger GC rich areas (roughly 6800 to 7100 on HXB2) upstream from (and not in) our PCR target (2136 to 3419 on HXB2). Figure 3.2.1.1.5 below displays our speculated variation in the produced cDNA when using random hexamer and random pentadecamers.

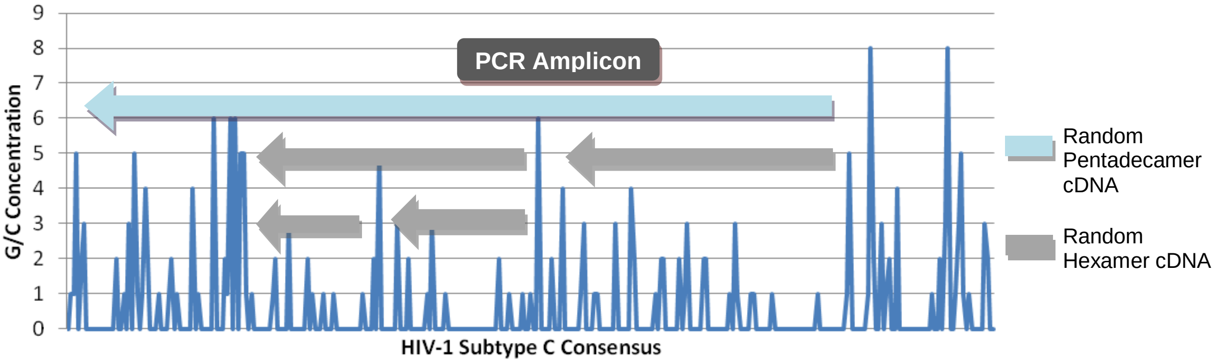


Figure 3.2.1.1.5 Hexamer and pentadecamer cDNA species: The figure above displays the possible cDNA species produced when using random hexamers (annealed at 25 °C) or random pentadecamers (annealed at 48 °C). None of the random hexamer derived cDNA species contain the full pre-nested PCR amplicon in contrast to cDNA species catalysed using random pentadecamers.

Random pentadecamers are 15-base oligonucleotides with Tms ranging from 25 to 76 °C as calculated using the nearest neighbour Tm method. They display intermediary properties when compared to random hexamers and gene-specific primers as they have Tms closer to that of specific primers without the selection bias, similar to the hexamers. Using a higher Tm, we were able to randomly prime a GC-rich region of HIV RNA and produce longer cDNA species that contained our full PN-PCR amplicon. This reaction was optimised with respect to pentadecamers annealing temperature and concentration (figure 3.2.1.1.6 below) before it was used to process our cohort samples.

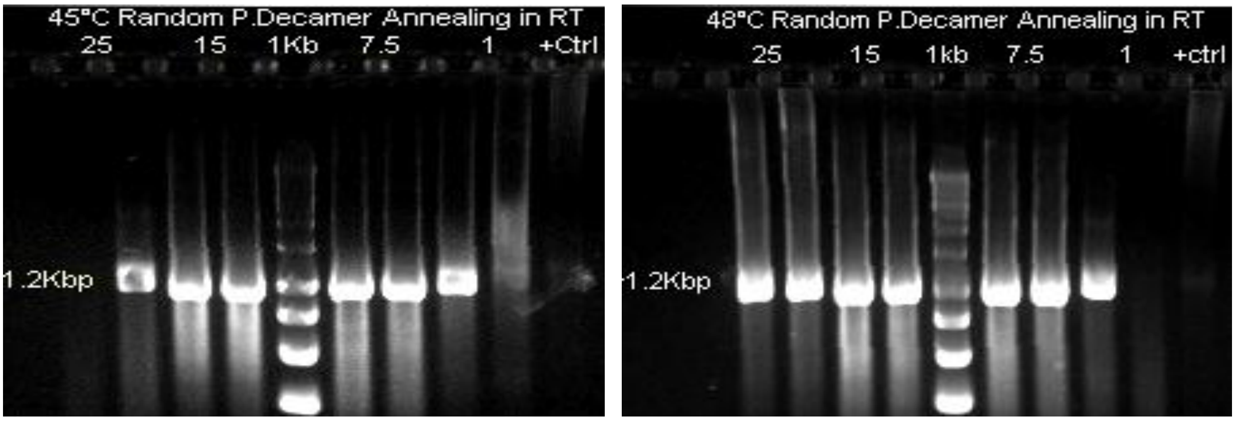

45°C Random P.Decamer Annealing in RT
25
15
1Kb
7.5
1
+Ctrl
1.2Kbp
48°C Random P.Decamer Annealing in RT
25
15
1kb
7.5
1
+ctrl
1.2Kbp

Figure 3.2.1.1.6 Reverse transcription optimisation: The figure above on the left shows the successful amplification of the 1.2 kbp PN-PCR amplicon from cDNA generated using random pentadecamers at the µM concentration shown on top of each lane and a 45 °C annealing temperature. Pentadecamers used in the reaction shown above on the right were annealed at 48 °C and showed brighter bands than those seen in the pre-nested reaction annealed at 45 °C. From these results it was evident that a pentadecamers concentration of 7.5 µM with a 48 °C annealing temperature produced the most cDNA including our sequence amplicon. The PN-PCR positive control (+ctrl) reaction amplified 50 copies of the HIV-1 subtype C plasmid, pZAC. This reaction showed poor amplification since 50 copies was the assayed limit of detection.

3.2.1.2 Pre-nested PCR optimisation

The PN-PCR was optimised with respect to enzyme system used, $MgCl_2$ concentration and annealing temperature. Initially, we matched Invitrogen’s Pfx50 (California, USA) high fidelity polymerase system against Roche Diagnostic’s Expand High Fidelity^{PLUS} (Basel, Switzerland) using the manufacturer’s standard protocols and 5 000 copies of a linearized HIV-1 subtype C plasmid, pZAC. Figure 3.2.1.2.1 below displays the successful Roche enzyme system and the failed Invitrogen reaction.

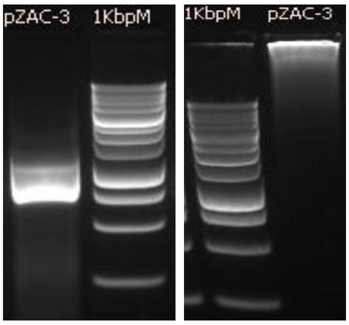


Figure 3.2.1.2.1 Pre-nested high fidelity PCR system: The gel figure on the far left displays the successful production of the 1.2 kbp PN-PCR amplicon using the Roche Diagnostics enzyme system (Basel, Switzerland) and 5 000 copies of the HIV-1 pZAC plasmid. The gel figure on the right shows the failed pre-nested amplification using Invitrogen’s enzyme system (California, USA). While this system generates some high concentration PCR product (smear at the top of the well), the size is incorrect. For both systems the standard manufacturer prescribed conditions were used.

Next, the Expand High Fidelity^{PLUS} (Roche Diagnostics, Basel, Switzerland) PN-PCR was optimised with respect to $MgCl_2$ final concentration, ranging from 0.5 mM to 6 mM, as depicted in figure 3.2.1.2.2 below. Empirically, 3.5 mM (lane eight from the left) produced the brightest band compared to other concentrations, and was used for further experimentation. The variation in band size between the positive control and the reactions could be attributed to a higher volume of the positive control loaded in the well. Since no nonspecific bands are seen, we assumed that the amplified product was the correct one.

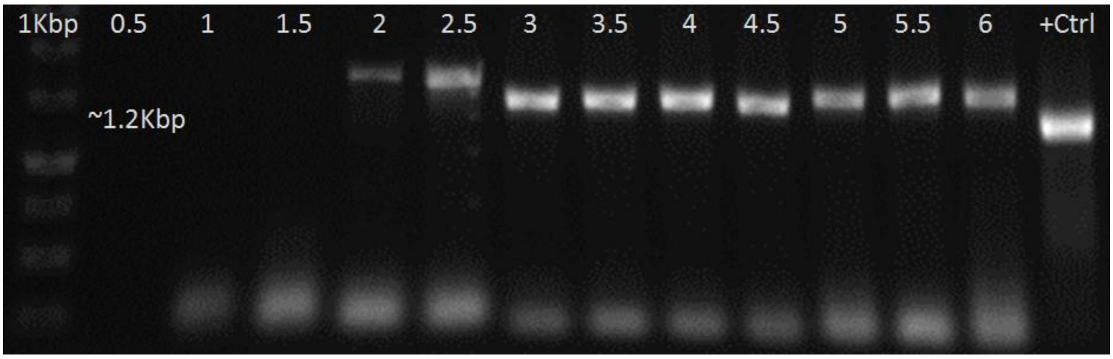


Figure 3.2.1.2.2 PN-PCR $MgCl_2$ optimisation: In the figure above depicts the $MgCl_2$ optimisation of the PN-PCR. While band intensity is not clearly visible in this gel picture, the brightest band was seen when using 3.5 mM final $MgCl_2$ concentration.

Finally, the pre-nested reaction annealing temperature was optimised between 62 and 66 °C as the manufacturers (Roche Diagnostics, Basel, Switzerland) suggest higher cycling temperatures result in better enzyme processivity. Gel figure 3.2.1.2.3 below shows that an annealing temperature of 64 °C exhibited more efficient amplification than higher or lower annealing temperatures.

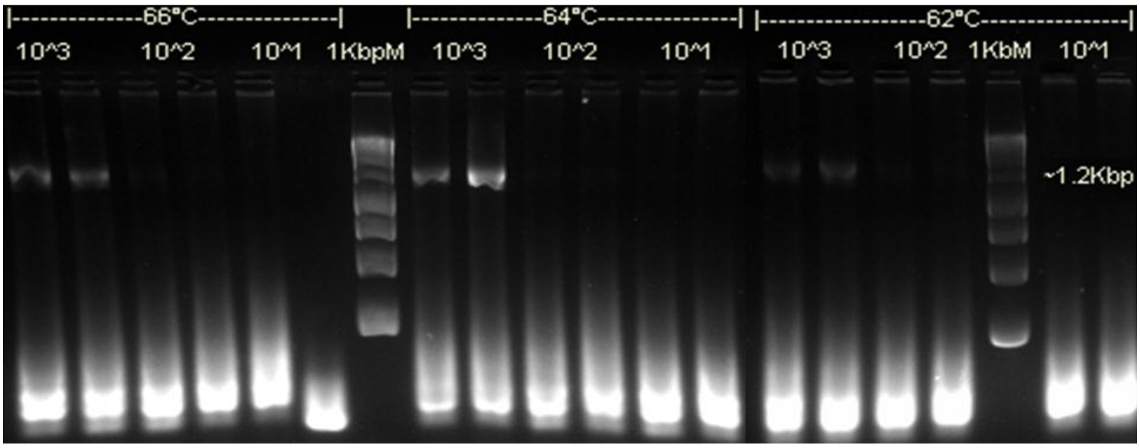

|-------------66°C--------------|
|-------------64°C--------------|
|-----------------62°C----------------|
10^3
10^2
10^1
1KbpM
10^3
10^2
10^1
10^3
10^2
1KbM
10^1
~1.2Kbp

Figure 3.2.1.2.3 PN-PCR annealing temperature optimisation: The above figure displays the effect of various annealing temperatures on the processivity of the PN-PCR, when amplifying a 1.2 kbp amplicon from 5 000 and 500 copies of the HIV-1 Subtype C plasmid, pZAC. An annealing temperature of 64°C seems to generate a brighter band than the other two temperatures tested however, none of the Tms improved assay sensitivity.

Once optimised, the sensitivity of the PN-PCR was determined using a linearized subtype C plasmid, since linear DNA closely represents cDNA. The pZAC plasmid was linearized by digestion with *Not* I (New England Biolabs, Massachusetts, USA), a restriction enzyme known to cut the plasmid once, outside the pre-nested target. Once gel purified, the linearized plasmid was quantified and reconstituted in 1 log/µl stocks. Gel figure 3.2.1.2.4 below shows the results of the sensitivity testing using 5 µl of plasmid stocks, freshly diluted in nuclease free water. In figure 3.2.1.2.5, the linearized plasmid was diluted in nuclease free water with 20 µg/µl glycogen DNA carrier (Roche Diagnostics, Basel, Switzerland). The addition of the carrier appears to increase the sensitivity of the assay however, this is not entirely true as glycogen merely blocks the binding of DNA to plastics and increases the number of template molecules available to the primers and polymerase. As glycogen is routinely used in our laboratory as a carrier, it was readily available.

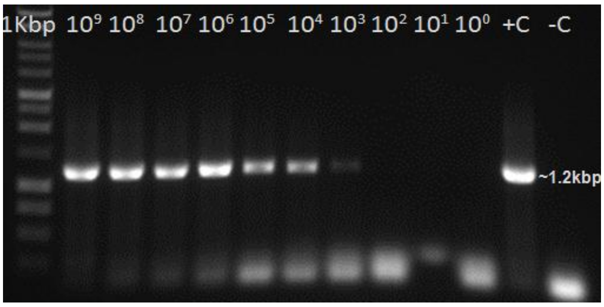

1Kbp 10^9 10^8 10^7 10^6 10^5 10^4 10^3 10^2 10^1 10^0 +C -C
~1.2kbp

Figure 3.2.1.2.4 Optimised pre-nested sensitivity test: Gel figure 3.2.1.2.4 above displays the results of the sensitivity testing of our optimised PN-PCR. Sensitivity was estimated at 5 000 linearized plasmid copies per reaction.

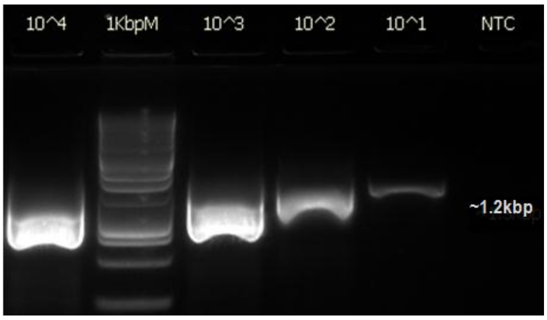


Figure 3.2.1.2.5 Improved pre-nested sensitivity: The addition of glycogen (a passive DNA carrier) to the DNA standards diluents improved the apparent sensitivity of the reaction however, the true sensitivity of the “un-assisted” PN-PCR can be seen in figure 3.2.1.2.5 above. The addition of 20 µg/µl glycogen increases the apparent reaction sensitivity one hundredfold to 50 linearized plasmid copies per reaction.

3.2.1.3 Nested PCR optimisation

Nested reaction optimisation began by amplifying 20 000 and 2 000 plasmid copies at various annealing temperatures between 65 °C and 55 °C, with a final primer concentration of 500 nM. Gel figure 3.2.1.3.1 below depict the results of this optimisation reaction showing more efficient amplicon synthesis at 65 °C annealing temperature.

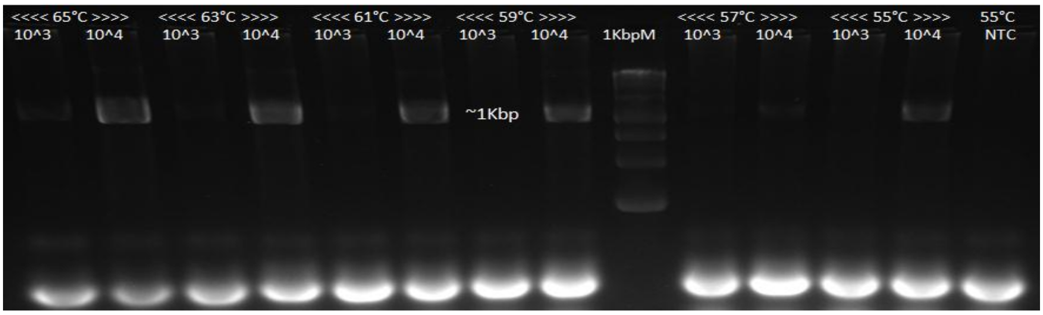


Figure 3.2.1.3.1 Nested PCR annealing temperature optimisation: This figure shows 20 000 and 2 000 plasmid copies amplified in our pre-nested PCR at various annealing temperatures, incremented by two degrees from 55 °C to 65 °C. While the DNA marker is faintly visible, successful amplification is seen at approximately 1 kbp. Neither of the tested annealing temperatures improved assay sensitivity however, an annealing temperature of 65 °C seemed to increase assay efficiency.

Next, $MgCl_2$ was titrated between 3.5 and 2.9 mM final concentration to find the optimum. Once again, 20 000 and 2 000 plasmid copies were used to test any improvement in sensitivity. Gel figure 3.2.1.3.2 below displays the PCR results showing optimal amplification of the nested amplicon at a 3.5mM $MgCl_2$ final concentration. In addition, very faint bands are visible in lanes three and four (from the left) corresponding to the 200 copies input reaction and demonstrating improved assay sensitivity. These bands are also seen in the 2.9 mM $MgCl_2$ reaction, however, we selected the higher $MgCl_2$ concentration for sample processing since it would allow for a less stringent PCR, considering HIV-1 diversity resulting in a high probability of primer template mismatches.

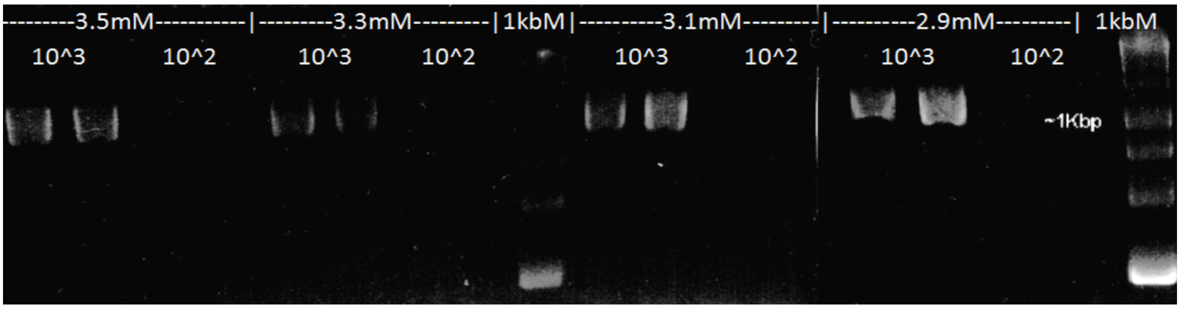

---------3.5mM-----------|---------3.3mM---------|1kbM|----------3.1mM---------|----------2.9mM---------| 1kbM
10^3
10^2
10^3
10^2
10^3
10^2
10^3
10^2
~1Kbp

Figure 3.2.1.3.2 Nested PCR $MgCl_2$ optimisation: The gel figure above depicts the successful amplification of the nested amplicon in all reaction sampling 2 000 plasmid copies with a range of MgCl2 concentrations from 2.9 mM to 3.5 mM. Faint bands are seen in the 200 plasmid copy input reactions for all $MgCl_2$ concentrations except 3.1 mM. Since a higher MgCl2 concentration tolerates more primer mismatches, the 3.5 mM concentration was used for further analysis as we intended sampling diverse viral populations.

Finally, primers were titrated at a 500, 300 and 200 nM final concentration, sampling 2 000 plasmid and using the previously optimised $MgCl_2$ concentration and annealing temperature. Gel figure 3.2.1.3.3 shows more efficient amplification using 200 nM rather than 500 nM final primer concentration and the previously optimised reaction conditions. For some unknown reason the 300nM reaction was unsuccessful however, since it contained an intermediate primer concentration, we assume that if would have display intermediate efficiency.

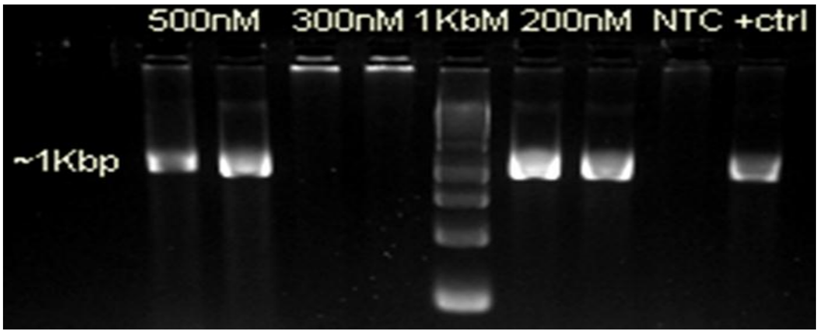


Figure 3.2.1.3.3 Nested PCR primer concentration optimisation: Above brighter bands are seen in the 200 nM primer input reactions than in the 500 nM reaction. For unknown reasons, the reaction with 300 nM primers failed to amplify however, the results were validated by the proper functioning of the reaction controls.

With the sample preparation reactions optimised, we turned our attention towards reducing PCR-enrichment-associated resampling error. To this end, we planned a feasible workflow that would allow us to sample many viral variants through paralleled pre-nested and nested PCRs. Figure 3.2.1.3.4 below illustrates our workflow to prepare the amplicon libraries for NGS. The figure also shows the conventional diagnostic DRT that was previously performed on the patient included in this study.

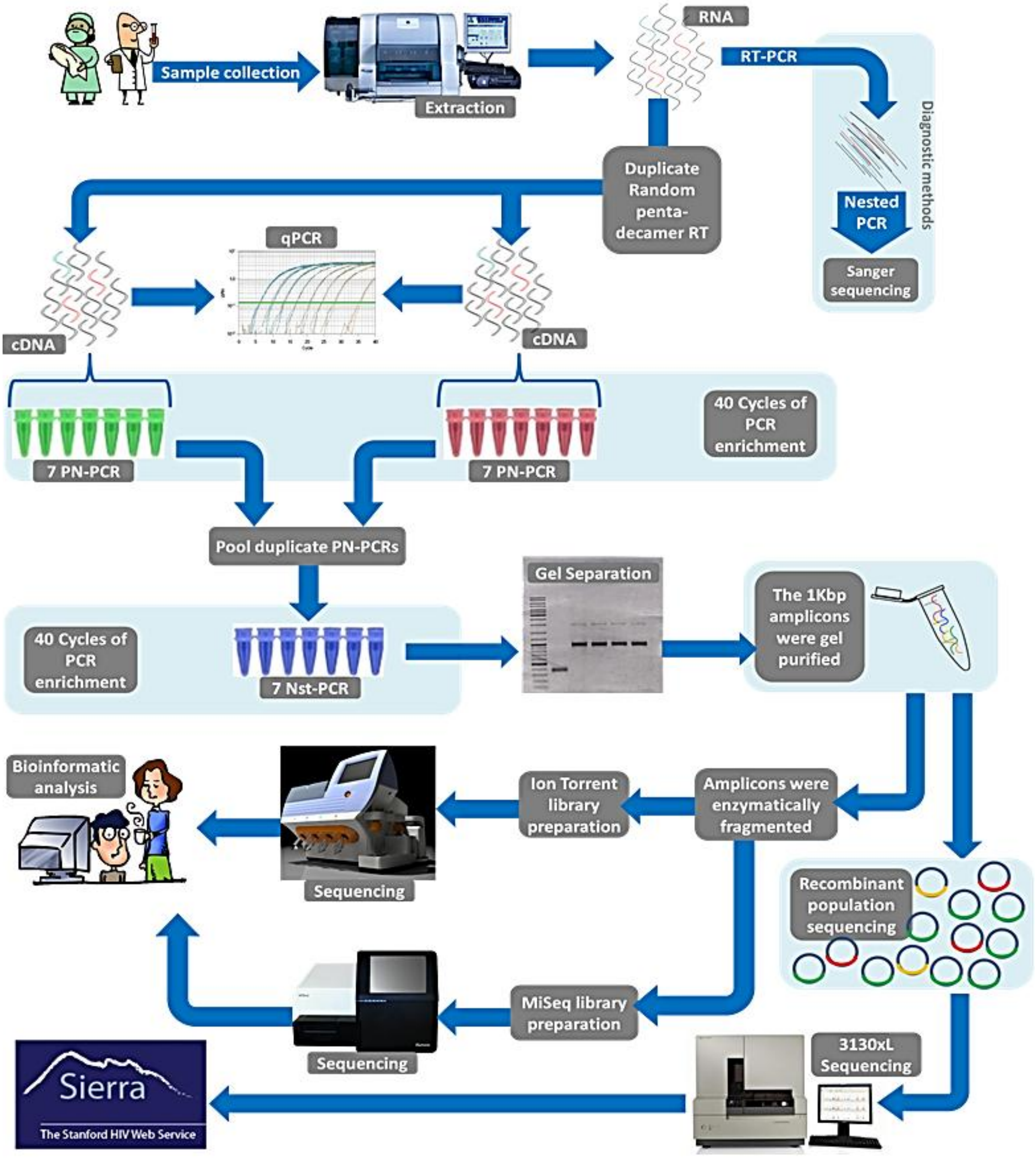

Sample collection
Extraction
RNA
RT-PCR
Diagnostic methods
Nested PCR
Sanger sequencing
Duplicate Random penta-decamer RT
qPCR
cDNA
cDNA
7 PN-PCR
7 PN-PCR
40 Cycles of PCR enrichment
Pool duplicate PN-PCRs
40 Cycles of PCR enrichment
7 Nst-PCR
Gel Separation
The 1Kbp amplicons were gel purified
Bioinformatic analysis
Sequencing
Ion Torrent library preparation
Amplicons were enzymatically fragmented
Recombinant population sequencing
Sequencing
MiSeq library preparation
3130xL Sequencing
Sierra
The Stanford HIV Web Service

Figure 3.2.1.3.4 Sample processing workflow: The optimised sample preparation workflow is illustrated above. Plasma from patients fitting our inclusion criteria were extracted on the EasyMag$^{®}$ (bioMérieux, Marcy l'Êtoile, France) system before the cDNA was generated through random pentadecamer-primed reverse transcriptions, performed in duplicate. After RNAse-H digestion, the two cDNA reactions were quantified by qPCR (to determine the number of species sampled) and amplified through 40 cycles in 14 paralleled PN-PCRs (to limit PCR sampling error and assay all available material from a patient). Duplicate PN-PCRs were pooled to form the template for seven paralleled nested PCRs, allowing an additional 40 amplification cycles. Nested PCR products were gel separated and purified before an aliquot was sent for Ion Torrent (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) sequencing. Gel purified products of patients characterised as harbouring diverse viral population with minor variants, were ligated into a bacterial vectors. These recombinants were later transformed, enriched through culturing, extracted and sequenced using an in-house, M13 sequencing protocol and the Applied Biosystems (ABi) 3130*xl* Genetic analyser (Massachusetts, USA). Deep sequencing results were analysed by a bioinformaticist and clonal population sequences were analysed using Stanford University’s online drug resistance profiling tools, Sierra.

3.2.2 Sample processing

Samples from a sub-study of the “Factors affecting the evolution of HIV-1 antiretroviral resistance mutations in patients treated at the Infectious Diseases Clinic at Tygerberg Academic Hospital and Referring Hospitals” were included in this study (Ethics Number: N06/05/081). As the included patients were all younger than 16 months, their parents gave written informed consent for their inclusion. In accordance with the Stellenbosch University’s Research Ethics Policy, participants were anonymized and assigned study numbers before data analysis.

3.2.2.1 Patient samples and nucleic acid extraction

Infants fitting our inclusion criteria were selected from a larger study investigating antiretroviral resistance in combination antiretroviral therapy-naïve patients. As such these infants were HIV-positive and PMTCT exposed but otherwise ARV untreated with approximately 1ml of plasma available from previous EDTA blood samples. Untreated children often have high viral loads and therefore limited plasma volumes were adequate to sample minor variants. The isolated plasma was then extracted on the EasyMag® DNA and RNA extraction machine (bioMérieux, Marcy l'Êtoile, France) and the resultant 25 µl nucleic acid extract was stored at -80 °C. To preserve the integrity of the viral RNA, the extract was only defrosted for reverse transcription when needed. Diagnostic drug resistance testing was conducted as described elsewhere (Claassen et al. 2011).

3.2.2.2 Reverse transcription (RT)

Reverse transcriptase is prone to dissociate from the RNA template when it encounters RNA secondary structure, which results in premature termination of cDNA synthesis. It is therefore important to ensure that RNA is denatured before and remains denatured during reverse transcription. We consequently synthesised cDNA using the CHAVI-MBSC-2 protocol with Invitrogen’s SuperScript® III RT (California, USA), and include a denaturing step. Moreover the thermo-stability of this reverse transcriptase allows for cDNA extension at higher temperatures when less RNA secondary structure is expected, and which would increase yield and produce longer cDNA fragments.

An accurate representation of the isolated viral population is essential for minor variant quantification and requires initial sampling methods with little or no selection bias. To this end, we tested random hexamers to prime the RT reaction, which seemed to generate lots of quantifiable cDNA, as measured with a short fragment qPCR. The resultant cDNA was however, discontinuous and un-amplifiable using our pre-nested PCR (PN-PCR), sensitive to >50 DNA copies. We therefore investigated and implemented the use of pentadecamers and a higher temperature

(48 °C) to prime reverse transcription, as this would bias binding to a larger GC rich area upstream from, and not in, our PCR target.

Our final modified protocol is as follows:

Each reaction was prepared in duplicate, in a 200 µl thin-walled PCR tube (eight tube strip) and all thermocycling was carried out in ABI 9700 thermocyclers (Applied Biosystems, Massachusetts, USA). The following reagents were prepared in a mastermix and aliquotted at the relevant concentrations:

	Reagent	Final concentration (in 27 µl)
1.	Random pentadecamers	11.11 µM
2.	dNTP Mix	0.741 µM
3.	12 µl extracted RNA	Unknown
4.	10 µl Nuclease free water (N.F.H_2O)	1 x

The reaction was then heated to 72 °C for 2 min to linearize the RNA and then at 48 °C for 5 min to allow the pentadecamers to anneal. Next, the reaction was rapidly cooled (-20 °C) in an aluminium chilling block (UltraCruz^{TM}, Texas, USA) before the following reagents were added at the required concentration, making up a 40 µl final reaction volume:

	Reagent	Final concentration (in 40 µl)
1.	RiboLock (RNAse inhibitor) (Thermo Scientific)	80 U/rxn
2.	DTT	5 µM
3.	SuperScript® III RT (Invitrogen)	200 U/rxn
4.	5 x Reaction buffer	1 x

This reaction was incubated at 56 °C for 1.5 hrs before another 200 U of SuperScript® III (Invitrogen, California, USA) RT enzyme was added. The reaction was once again incubated at 56 °C for 1.5 hrs before the enzymes were deactivated by heating the reaction to 85 °C for 5 min. Once on ice, four units of *E. coli* RNAse H (Thermo Scientific, Massachusett, USA) were added and the reaction was incubated at 37 °C for 20 min. The resultant cDNA was stored at -20 °C until needed for the PN-PCR.

3.2.2.3 cDNA quantification

As it is critical to estimate the starting cDNA population size prior to PCR enrichment (as the starting population size would determine the limit for minor variant detection) we performed quantitative PCR (qPCR) to quantify the cDNA starting concentration. A 5 µl aliquot of patient-derived cDNA was quantified using a SYBR green qPCR, GoTaq® HotStart DNA polymerase system (Promega, Wisconsin, USA) and the primers seen in table 3.2.2.3.1 below. qPCRs were performed on an ABI 7900HT real time PCR platform (Applied Biosystems, Massachusetts, USA) and analysed using the supplied Applied Biosystems Software (SDS version 2.3) (Massachusetts, USA). A melt curve analysis step was included at the end of the qPCR assay to discriminate between double-stranded DNA from PCR amplification and nonspecific signals (primer-dimer) fluorescence.

For the standard curve, plasmid DNA standards (pZAC – a South African HIV-1 subtype C plasmid) were amplified in triplicate at 1 log dilutions from 5 x 10^6 to 5 x 10^0 input copies. Non-template controls and patient cDNA samples were assayed in duplicate.

Table 3.2.2.3.1 qPCR primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward primer	F1_F/RealT	RGCTCTMTTAGAYACAGGAGCAGAT	2315 to 2339
Reverse primer	F1_R_RT	ACTTTGATAAAACCTCCAATTCCYCC	2394 to 2419 (RC)

(RC)= Reverse complement

The following qPCR was compiled using the listed reagents at their indicated concentration:

Reagent	Final concentration (in 25 µl)
1. 5 x Reaction buffer	1 x
2. F1_F/RealT FWD Primer	400 nM
3. F1_R_RT REV Primer	400 nM
4. dNTP Mix	200 µM
5. $MgCl_2$	3 mM
6. 25 mM ROX reference dye	500 µM
7. 50 x SYBR green Gel stain	0.5 x
8. GoTaq® HotStart DNA Polymerase	0.625 U/rxn
9. 5 µl synthesised cDNA / pZAC	Unknown / Varied
10. 9.62 µl N.F.H_2O with 1 mg/ml Glycogen (Roche)	1 x

The following thermocycling conditions were used:

95 °C for 120 sec*

40 x
94 °C for 30 sec*
54 °C for 30 sec*
72 °C for 30 sec*

Melt curve analysis

95 °C for 15 sec
65 °C for 5 sec
Thermal ramp at 0.2 °C/sec*
95 °C for 5 sec*

(*acquisition on green channel)

On completion, these results were analysed using the SDS (V2.3) software (Applied Biosystems, Massachusetts, USA). An acceptable standard curve was defined as having a Y-intercept of greater than 38 cycles, an efficiency coefficient greater than 0.97 and a slope gradient between -3.2 and -3.6. The final batch of cDNA samples (A339, A345, A364, A428, A437 and A93) could not be quantified due to 7900HT (Applied Biosystems, Massachusetts, USA) laser malfunction. For each sample, the sampled copy number is calculated by multiplying the estimated cDNA copy number by seven, since, from each 40 µl cDNA reaction, one 5 µl aliquot was used for qPCR and the remaining seven aliquots, for PN-PCRs. The total number of variants sampled per patient is the sum of the sampled copy number for duplicated cDNAs.

3.2.2.4 Pre-nested PCR

PCR induced errors, while infrequent, may misrepresent the sampled starting population in downstream next gen genotyping. Even low frequency PCR errors, not seen with bulk Sanger sequencing, could be observed with deep sequencing due to the high template coverage. It is therefore critical that a high fidelity enzyme system is used during PCR to minimize PCR error, such as misincorporation resulting in base substitutions, insertions or deletions. To this end, we optimised our pre- and nested PCRs using Roche Diagnostics' Expand High Fidelity^{PLUS} (Basel, Switzerland), an enzyme system with six-times better fidelity than regular *Taq* DNA polymerase.

Once 5 µl of the cDNA had been removed and quantified, the remaining 35 µl were split between seven paralleled PN-PCR's (5 µl/rxn) to limit random PCR sampling error. This phenomenon is known as PCR resampling and results from one initially amplified template being resampled in subsequent rounds of PCR. When this occurs, the proportion of variants in the enriched population is different to that of the starting population. Since the RTs were performed in duplicate, a total of 14 pre-nested reactions originated from each patient sample. Our optimised Expand High Fidelity^{PLUS} (Roche Diagnostics, Basel, Switzerland) reaction used the following PN-PCR primers in table 3.2.2.4.1 below:

Table 3.2.2.4.1 PN-PCR primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward primer	50Prot2	TCAGAGCAGACCAGAGCCAACAGCCCCA	2136-2163
Reverse primer	NE135	CCTACTAACTTCTGTATGTCATTGACAGTCCAGCT	3334-3300 (RC)

(RC)= Reverse complement

The pre-nested reactions were compiled using the listed reagents at their indicated concentration:

	Reagent	Final concentration (in 50 µl)
1.	5 x Reaction buffer	1 x
2.	50Prot2 FWD Primer	500 nM
3.	NE135 REV Primer	500 nM
4.	dNTP Mix	200 µM
5.	$MgCl_2$	3.5 mM
6.	Expand HiFi ^{Plus} Polymerase (Roche)	2.625 U/rxn
7.	5 µl cDNA	Unknown
8.	24.25 µl N.F.H_2O with 1 mg/ml Glycogen (Roche)	1 x

The following thermocycling conditions were used for this PN-PCR:

94 °C for 2 min

40 x
- 94 °C for 30 sec
- 64 °C for 30 sec
- 72 °C for 90 sec

72 °C for 7 min

4 °C soak

On completion, the duplicate pre-nested reactions were combined to form seven, 100 µl PN-PCR products for each patient sample. These pre-nested products were then stored at -20 °C until needed for nested PCR.

3.2.2.5 Nested PCR

These nested PCR's were performed in replicates of seven, resulting in a total of seven nested reactions per patient sample. Expand High Fidelity^{PLUS} (Roche Diagnostics, Basel, Switzerland) was once again used for all nested PCR's along with the following nested PCR primers, shown in table 3.2.2.5.1:

Table 3.2.2.5.1 Nested PCR primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward primer	F1_F	CTCTCTTAGACACAGGAGCAGAT	2317 to 2340
Reverse primer	F3_R_alt_3	CCATTTGTCAGGATGGAGTTCATA	3267 to 3243 (RC)

(RC)= Reverse complement

The nested reaction was compiled using these listed reagents at their indicated concentration:

	Reagent	Final concentration (in 50 µl)
1.	5 x Reaction buffer	1 x
2.	F1_F FWD Primer	200 nM
3.	F3_R_alt_3 REV Primer	200 nM
4.	dNTP Mix	200 µM
5.	$MgCl_2$	3.5 mM
6.	Expand HiFi ^{PLUS} Polymerase (Roche)	2.625 U/rxn
7.	4 µl Pre-nested product	Unknown
8.	25.25 µl N.F.H_2O with 1 mg/ml Glycogen (Roche)	1 x

The following thermocycling conditions were used for this nested PCR:

94 °C for 2 min

40 x
- 94 °C for 30 sec
- 65 °C for 30 sec
- 72 °C for 80 sec

72 °C for 7 min

4 °C soak

On completion, these nested products were then stored at -20 °C until needed for separation by electrophoresis.

3.2.2.6 Gel electrophoresis

An additional precaution taken to prevent point mutations and DNA degradation due to ultraviolet light exposure during gel electrophoresis was the use of GelStar™ nucleic acid stain (Lonza, Basel, Switzerland). This gel stain allows for the visualisation of DNA over blue light and through an orange exclusion filter. 50 µl of each nested PCR products were separated on a 1% agarose gel made in sodium borate buffer, stained with 1x GelStar™ (Lonza, Basel, Switzerland). Electrophoresis gels were photographed over a blue light with orange exclusion.

3.2.2.7 Re-PCR

Once photographed and documented, reactions which yielded less than five out of seven positive nested PCR's were repeated to verify results, and to exclude non-detection due to operator error. However, if the number of positive samples remained less than five out of seven on repeat, these cases were excluded from the investigation, as this would indicate that the cDNA sampled was too close to the limit of detection to allow an accurate estimation of low frequency variants. Random sampling error is at its highest close to the limit of detection as the coefficient of variation for the number of events detected, predicted by the Poisson distribution, is the highest when the number of cDNA species sampled is at its lowest. Electrophoreses was repeated as described above. Of the 46 initial samples processed, only 15 samples were included since most of the others either failed to produce an amplicon after the nested reaction or failed to produce an amplicon in more than five out of seven positive nested PCRs.

3.2.2.8 Purification and quantification

Nested PCR reactions yielding five or more positive results (~1 kbp amplicon) were excised from the electrophoresis gel, pooled together and weighed on a fine balance. Amplified DNA was purified from the gel slices using a Promega's (Wisconsin, USA) Wizard® SV Gel and PCR clean-up system according to the manufacturer's specification. Briefly: 10 µl of membrane binding buffer was added for every 10 mg of gel slice; this mixture was then vortexed for two minutes and incubated at 65 °C for ten minutes; with the gel dissolved, the mixture was centrifuged through a silica-membrane column at 14700 xg; this membrane was then washed twice with 500 and 700 µl membrane wash solution before nuclease-free water (preheated to 50 °C) was used to elute the DNA from the purification column.

Two microliters of these purified products were quantified on a NanoDrop® ND1000 (Thermo Scientific, Massachusetts, USA) in triplicate, before 30 µl of each sample was sent for PGM sequencing at the Stellenbosch University Central Analytical Facility. The remaining 14 µl we stored at -20 °C until needed for clonal sequencing.

3.2.2.9 PGM and MiSeq library preparation

The IonXpress™ Plus library preparation manual, rev. M was followed for Ion Torrent amplicon library construction (Life Technologies, California, USA). The purified amplicons were digested using the Ion Shear Enzyme II kit and protocol (Life Technologies, California, USA), followed by barcode ligation and size selection using the E-Gel system (Invitrogen, California, USA), for an average insert size of 200 bp. Template enrichment and sequencing was performed according to manufacturer recommendation with the Ion 200 Sequencing V1 kit 3.3.2 (Life Technologies, California, USA).

For the MiSeq (Illumina, California, USA) library construction, the Nextera XT DNA sample preparation manual, rev. C was followed (Illumina, California, USA). The digestion and adaptor ligation was performed simultaneously followed by the PCR mediated index addition. AmPure XP reagent (Beckman-Coulter, California, USA) was used to size select 300-500 bp fragments and sequencing was performed using the MiSeq Reagent Kit V2 (Illumina, California, USA).

3.2.2.10 Ion Torrent and MiSeq bioinformatic analysis

A South African HIV-1 Subtype C consensus sequence was generated from 317 whole genomes, sequenced between 1997 and 2010 and downloaded from HIV online database (available at http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html). These sequences were aligned using the Geneious align tool in Geneious V6.0 (Biomatters, Auckland, New Zealand). Reads were filtered using instrument quality scores and aligned to the subtype C reference sequence using a codon-aware version of the Smith-Waterman algorithm that corrects for homopolymer errors. Figure 3.2.2.10.1 below illustrated the remedial action of the codon-aware algorithm that considers both nucleotide and amino-acid homology and directly penalizes for length miscall.

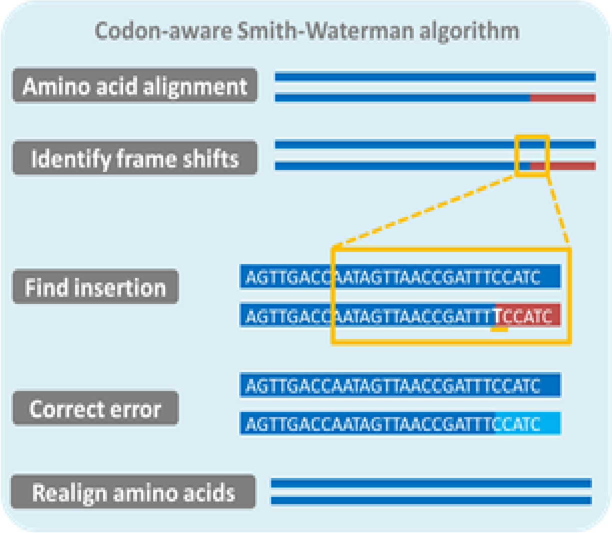

Codon-aware Smith-Waterman algorithm
Amino acid alignment
Identify frame shifts
Find insertion
AGTTGACCAATAGTTAACCGATTTCCATC
AGTTGACCAATAGTTAACCGATTTTCCATC
Correct error
AGTTGACCAATAGTTAACCGATTTCCATC
AGTTGACCAATAGTTAACCGATTTCCATC
Realign amino acids

Figure 3.2.2.10.1 Datamonkey codon-aware mapping algorithm: This figure illustrates the corrective action of the codon-aware version of the Smith-Waterman read mapping algorithm. The algorithm maps quality filtered sequence reads to an HIV-1 Subtype C reference sequence in codon space. When a frame shift is encountered in the mapping, the algorithm interrogates the nucleotide sequence and removes erroneously called homopolymer bases. The corrected sequence it then realigned to the consensus.

Bioinformatics analysis was performed by our collaborators from the University of California San Diego, Bioinformatics Core (Sergei Kosakovsky Pond and Ben Murrell). A mixture of multinomial probabilistic model was used to distinguish sequencing error from true low-frequency variants with posterior probabilities of ≥99.99%, excluding bases identified as errors. For each sample, they computed the mean of all pair-wise Tamura-Nei 93 distances between reads with at least 100 overlapping base pairs to quantify nucleotide diversity. As bidirectional sequencing was provided on the MiSeq (Illumina, California, USA), only minor variants recorded by both reverse and forward sequencing, at a threshold of ≥0.5%, were included in further analysis.

3.2.2.11 Phylogenetic analysis

We constructed a maximum likelihood tree (GTR + CAT model in FastTree), based on the first 630 nucleotides of *reverse transcriptase* (the region that was uniformly well-covered for all sequencing runs) using NGS majority consensus and Sanger bulk sequences.

3.2.2.12 Clonal sequencing

3.2.2.12.1 Amplicons

As quality control for NGS, amplicons sequenced on the PGM, and which revealed the presence of quantifiable drug-resistant variants (patients A124, A313, A157 and A158) were cloned for bi-directional, M13-based Sanger sequencing. Since the M13 primer binding region is vector-derived and identical in each vector, this clonal sequencing approach was unbiased.

3.2.2.12.2 Ligation

One-hundred-and-sixty-six nano grams of A124, 178 ng of A313, 249.3 ng of A157 and 193.7 ng of A158 purified amplicons were ligated into 100 ng of the pGEM®-T Easy Vector (Promega, Wisconsin, USA) using the reaction mixtures and thermal incubations described below. The ligation reaction was validated using the relevant reaction controls that included a ligation positive control insert and a background/no-insert control. Each reaction was prepared individually (i.e.: a mastermix was not prepared) in a 200 µl thin-walled PCR tube and all thermal incubations were carried out in an ABI 9700 thermocycler (Applied Biosystems, Massachusetts, USA), with the heat-lid activated at 105 °C.

	Reagent	Final concentration (in 20 µl)
1.	2 x Rapid ligation reaction buffer	1 x
2.	T4 DNA Ligase (3 U/µl) (Promega)	3 U
3.	Control DNA insert (4 ng) or DNA amplicon	Various
4.	10 µl Nuclease free water (N.F.H_2O)	1 x
5.	pGEM® T-Easy vector (50 ng/µl) (Promega)	100 ng

Ligation reactions were then incubated at 25 °C for 2 hrs, followed by 4 °C for 16 hrs, 75 °C for 5 min and an indefinite hold at 4 °C. On completion, recombinants were immediately transformed using an in-house optimised transformation protocols seen below.

3.2.2.12.3 Transformation of A124 and A313 recombinants

The recombinant A124 and A313 constructs were transformed using the following optimised protocol, along with a transformation efficiency control plasmid:

- Prepare and chill a 5x concentrate KCM transformation additive solution consisting of 5.1 mM KCl, 1.5 mM $CaCl_2$ and 2.5 mM $MgCl_2$.
- Pre-chill 1.5 ml tubes and SOC media while thawing 250 µl JM109 chemically competent bacteria (Promega, Wisconsin, USA) on ice.
- Once the competent bacteria have thawed, add 62 µl of the KCM additive and mix by gently swirling and flicking the bacteria stock.
- Add 5 µl of the ligation constructs to the pre-chilled tubes.
- Add 50 µl JM109 cells (Promega, Wisconsin, USA) with the 1x KCM additive to each pre-chilled tube. Mix once by slow up and down pipetting.
- Incubate tubes on ice for 1 min then heat shock at exactly 42 °C for 50 sec.
- Incubate reactions on ice for 2 min before adding 950 µl cold SOC.

The reaction was then incubated at 37 °C for 1.5 to 2 hrs with gentle shaking before being spread on to a 400 cm^2, semisolid, ampicillin-selective LB media culture plate for IPTG/X-Gal mediated blue/white colony selection. These plates (Q-trays) were incubated at 37 °C for 16 hrs, before being stored at 4 °C for colour intensification.

3.2.2.12.4 Transformation of A157 and A158 recombinants

The recombinant A157 and A158 constructs were transformed using the following optimised protocol, along with a transformation efficiency control plasmid:

- Pre-chill 1.5 ml tubes and SOC media while thawing 250 µl JM109 chemically competent bacteria (Promega, Wisconsin, USA) on ice.
- Add 5 µl of the ligation constructs to the pre-chilled tubes.
- Add 50 µl JM109 cells (Promega, Wisconsin, USA) to each transformation tube. Mix once by slow up and down pipetting.
- Incubate tubes on ice for 1 min and heat shock at exactly 42 °C for 50 sec.
- Incubate reactions on ice for 2 min before adding 950 µl cold SOC.

The reaction was then incubated at 37 °C for 1.5 to 2 hrs with gentle shaking before being spread on four, semisolid Ampicillin-selective LB media culture plates, for IPTG/X-Gal mediated blue/white colony selection.

3.2.2.12.5 Culturing and recombinant constructs recovery

Using the Q-Pix II automated colony picker (Molecular Devices, California, USA), 276 and 253 white colonies were selected for samples A124 and A313 respectively. These colonies were inoculated into 200 µl of ampicillin-selective liquid LB media in a 96 welled plate before being sealed and incubated overnight at 37 °C with orbital shaking at 200 rpm. On completion, 100 µl of each culture was diluted in an equal volume of a sterile, 50% glycerol solution before being frozen at -80 °C overnight. These glycerol stocked cultures were then couriered to the European branch of Macrogen (Seoul, South Korea) for plasmid extraction and M13-mediated forward and reverse sequencing.

From samples A157 and A158, 250 white colonies were inoculated by hand into 4 ml of ampicillin-selective liquid LB media and incubated overnight, at 37 °C with orbital shaking at 200 rpm. These hand-picked cultures were miniprepped using the manufacturer's specified protocol (Promega, Wisconsin, USA) and 2 µl of the purified constructs, quantified on a NanoDrop® ND1000 (Thermo Scientific, Massachusetts, USA).

3.2.2.12.6 A157 and A158 recombinant sequencing

Between 150 and 450 ng of the purified pGEM® constructs extracted from A157 and A158 cultures, were sequenced using the following in-house optimised BigDye® Terminator (V3.1) Cycle Sequencing kit reaction (Applied Biosystems, Massachusetts, USA), primed with standard M13 forward or reverse oligos.

	Reagent	Final concentration (in 10 µl)
1.	5 x Terminator sequencing reaction buffer	1.5 x
2.	BigDye® Terminator reaction mix (ABi)	1 x
3.	M13 Fwd or Rev Primer	5 pMoles
4.	pGEM®-T DNA construct	150 to 450 ng

Sequencing PCRs were performed using the following in-house-optimised reaction conditions in an ABI 9700 thermocycler:

95 °C for 10 sec	
	95 °C for 5 sec
25 x	72 °C for 4 min
4 °C soak	

On completion, the sequencing PCR reaction was purified using an in-house-optimised BigDye® X-terminator clean-up (Applied Biosystems, Massachusetts, USA), seen below. The SAM solution™ (Applied Biosystems, Massachusetts, USA) used in the purification reaction was heated to room temperature before use.

	Reagent	Volume added to Mastermix
1.	SAM™ solution (Applied Biosystems)	49 µl
2.	X-terminator solution (Applied Biosystems)	11 µl

Fifty-five microlitres of the above reaction mixture was added to each sequencing PCR in a 96 well plate, before the wells were sealed and the reaction, vortexed for 30 to 45 min. The purified sequencing PCR product was then centrifuged at 1000 xg for five minutes to pellet the X-terminator particles and the sequencing reaction was read on an ABI Genetic Analyser 3130*xl* (Applied Biosystems, Massachusetts, USA).

3.2.2.12.7 A157 and A158 false positive correction

Unsuccessful sequencing reactions were repeated and inconclusive sequencing results were substituted. For species characterised as containing no HIV insert as revealed by the National Centre for Bioinformatics Information (NCBI) blast results, alternative white colonies were inoculated into fresh LB broth. These cultures were then miniprepped, quantified and sequenced as previously described. These methods resulted in 252 positively screened species for both samples A157 and A158.

3.2.2.12.8 A124 and A313 bioinformatic analysis

The .fastq sequence files downloaded from Macrogen were uploaded into Stanford University's Calibrated Population Resistance (CPR) tool (V6.0) (available online at http://cpr.stanford.edu/cpr.cgi), as all the NNRTI mutations of interest were included in the surveillance drug resistance mutations list (SDRM) (Bennett et al. 2009). The CPR tool aligns the nucleotide sequences to an HIV-1 Subtype B, HXB2 amino acid sequence using a local alignment program, before translating them into an in-frame amino acid sequence. CPR then excludes sequences containing undocumented insertion, deletions and stop codons, as well as poor quality sequences. The bioinformatic process is explained in full detail in the CPR release notes (available at http://cpr.stanford.edu/pages/releaseNotes.html). Analysed results were exported to Excel 2010 (Microsoft, Washington, USA), where drug-resistant mutation frequencies are viewed as percentages.

3.2.2.12.9 A157 and A158 bioinformatic analysis

Chromatograms from the in-house sequencing reactions were imported into Geneious V6.0 (Biomatters, Auckland, New Zealand) and blasted against the NCBI nucleotide sequence database to determine the sequence specificity. Sequences were named according to their sample origin followed by their construct number and finally, the sequence direction (e.g.: Sequence name "8.214F" notates a sequence from sample A158; is the 214th construct sequenced; and is the forward sequence for that construct). The 252 forward and reverse sequences for each sample were converted to .fasta format using a 60% base-call confidence threshold.

Each forward and corresponding reverse sequence was aligned to produce one contiguous “read” for each recombinant. The drug resistance profile of these 504 contiguous reads (252 contigs for both A157 and A158) were generated using Stanford University’s Sierra HIV drug resistance database (available at http://sierra2.stanford.edu/sierra/servlet/JSierra?action=sequenceInput). The results were exported in .csv format and imported into Excel (Microsoft, Washington, USA) where mutation frequencies were calculated.

3.3 Results

3.3.1 Patient samples and nucleic acid extraction

We conducted a retrospective cross sectional study in 15 HIV-infected and PMTCT-failed infants, born between October 2006 and October 2009. At the time of sample collection all participants were treatment-naïve (baseline sampling). The PMTCT regimen consisted of maternal zidovudine (AZT) from 28 weeks gestation, intrapartum nevirapine (NVP), and neonatal single dose NVP with seven days of AZT. The median age of participants at sample collection was 3.4 months (inter-quartile range (IQR) 2.4–4.5), and 11 (73%) of the 15 infants were female. Other clinically relevant patient information is summarized in table 3.3.1.1.

Table 3.3.1.1 Patient demographics: The table below displays clinically relevant information of our study participants when sampled at baseline. The median viral load at sampling was 5.8 $\log_{10}$ HIV RNA copies/ml (IQR 5.1–6.2); median CD4 count was 1693 cells/µl (IQR 650–2025) and the median CD4% of 26% (IQR 21–35%). All patients' viruses were genotyped as HIV-1 subtype C species. Values indicated with 'NA' were not available.

		Age	Baseline sampling		HIV-1
Study No.	Gender	(months)	CD4 count	Viral load	Subtype
A124	Female	16	1208	2,900,00	Subtype C
A137	Female	4	4150	29,000	Subtype C
A144	Female	2	2036	82,000	Subtype C
A157	Female	2	540	660,000	Subtype C
A158	Female	2	2913	650,000	Subtype C
A202	Female	3	430	NA	Subtype C
A207	Female	2	NA	NA	Subtype C
A297	Male	4	2030	180,000	Subtype C
A300	Female	4	1744	120,000	Subtype C
A302	Female	2	801	500,000	Subtype C
A312	Male	5	2009	3,200	Subtype C
A313	Female	14	1886	>1,000,000	Subtype C
A318	Male	7	518	2,300,000	Subtype C
A326	Male	4	1641	>3,000,000	Subtype C
A364	Female	3	600	4,300,000	Subtype C

3.3.2 cDNA quantification

Once reverse transcribed, the viral RNA was digested and the cDNA was quantified using our in-house qPCR as described in the methods (section 3.2.2.3). The qPCRs were conducted in four batches of which only the first three batches were successful. The fourth qPCR batch run was unsuccessful due to mechanical failure on the ABI 7900HT (Applied Biosystems, Massachusetts, USA), resulting in sample A364 cDNA not being quantified. The results of the successful first qPCR batch are seen in figures 3.3.2.1 and 3.3.2.2 below, as well as in tables 3.3.2.1 and 3.3.2.2.

This first batch quantified cDNA duplicates from patients A124, A137, A144 and A157 using a linearized HIV-1 subtype C plasmid (pZAC) standard in triplicate. An acceptable standard curve gradient was defined as being between -3.2 and -3.6 (defined as the change in Ct divided by the change in labelled concentration), with a Y-intercept (theoretical Ct to detect one copy) above 38 cycles and an R^2 value (accuracy measurement) greater than 0.97.

Table 3.3.2.1 Quantified plasmid standards: The table below shows the averages of the triplicate standards, their measured Ct, and their input copy number.

Label	Content	Ct	Copy No.
$5x10^1$	Standard	36.074	50
$5x10^2$	Standard	32.649	500
$5x10^3$	Standard	28.485	5000
$5x10^4$	Standard	25.347	50000
$5x10^5$	Standard	21.534	500000

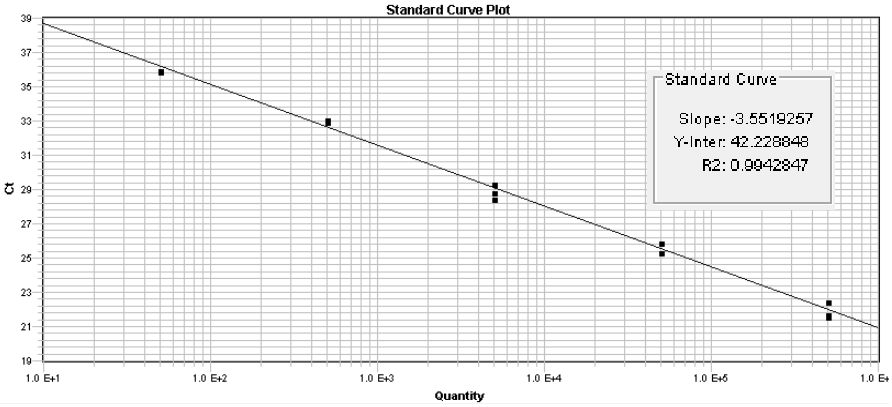


Figure 3.3.2.1 Batch 1 qPCR standard curve: The straight line graph above displays the observed Ct of the first batch of pZAC plasmid standards in relation to their concentration. The gradient of the slope, R^2 value and the Y-intercept were within our acceptable range.

Using the accepted standards curve displayed above, cDNA duplicates from samples A124, A137, A144 and A157 were successfully quantified. Their estimated copy numbers are seen in table 3.3.2.2 below, along with their dissociation curve analysis in figure 3.3.2.2.

Table 3.3.2.2 qPCR Batch 1 quantified cDNA: Below are the Cts and estimated cDNA copy numbers as estimated using the previously described standard curve. All cDNA samples were quantified except for the first replicate from sample A157 which failed to amplify for unknown reasons. The sample's copy number is calculated by multiplying the estimated cDNA copy number by seven, since, from each 40 µl cDNA reaction, one 5 µl aliquot was used for qPCR and the remaining seven aliquots, for pre-nested PCRs.

Label	Ct	Copy No.	Sampled Copy No.
A124-1	32.307	549.28	3844.98
A124-2	33.605	241.65	1691.58
A137-1	35.309	82.17	575.19
A137-2	35.690	64.57	451.98
A144-1	35.305	82.39	576.75
A144-2	33.144	323.49	2264.40
A157-1	N/A	0.00	0.00
A157-2	34.143	171.85	1202.96

As SYBR green-based qPCRs measure general amplification and not the production of a specific amplicon, dissociation analysis is performed after amplification. Briefly, well fluorescence is measured as the temperature is increased from 60 to 95 °C at a rate of 0.2 °C per second. The derivative of the observed fluorescence is plotted against the temperature, forming an amplicon dissociation/melt curve. Nonspecific products, such as primer dimers, melt at relatively low temperatures in comparison to the specific amplicons. These specific products display similar melt curve signatures, since they gave similar compositions (i.e.: similar nucleotide sequence and G-C content). Alternatively amplified products can be distinguished from specific products by their melt curve signature.

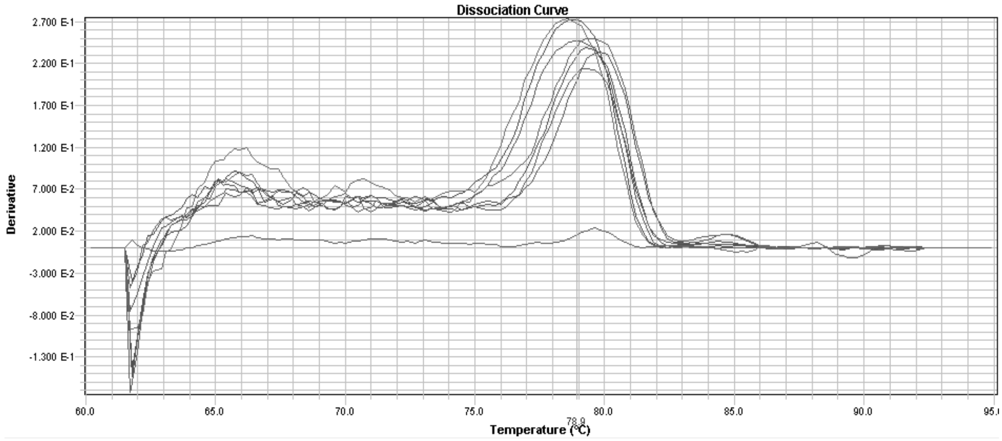


Figure 3.3.2.2 qPCR Batch 1 melt-curve analysis: The line graph above displays the dissociation curve of the amplicons produced from patient cDNA. The green vertical line indicates the average melting temperature for specific amplicons at approximately 79 °C. The faint pink line showing almost no fluorescence change represents the first cDNA replicate from patient A157 which failed to amplify.

The second batch of qPCR results are seen in figures 3.3.2.3 and 3.3.2.4, and in tables 3.3.2.3 and 3.3.2.4 below. This qPCR batch quantified cDNA duplicates from patients A158, A202, and

A207 using the previously described subtype C plasmid (pZAC) standard in triplicate. Standard curve acceptance was based on the previously mentioned criteria.

Table 3.3.2.3 Quantified plasmids standards: The table below tabulates the average of the triplicate standards, their measured Ct and their input copy number.

Label	Content	Ct	Copy No.
$5x10^1$	Standard	34.518	50
$5x10^2$	Standard	31.776	500
$5x10^3$	Standard	27.962	5000
$5x10^4$	Standard	25.032	50000
$5x10^5$	Standard	21.187	500000

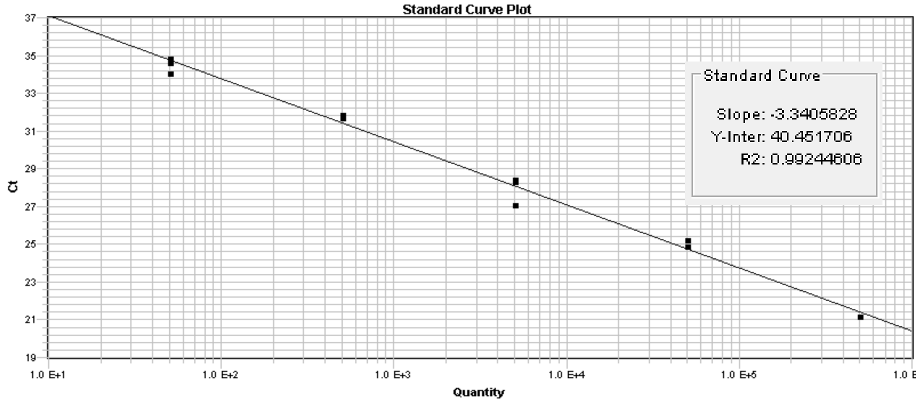


Figure 3.3.2.3 Batch 2 qPCR standard curve: The straight line graph displays the observed Ct of the second batch's pZAC plasmid standards in relation to their concentration. The gradient of the slope, R^2 value and the Y-intercept were within our acceptable range.

The standard curve above was used to estimate the cDNA concentration of samples A158, A202 and A207. Table 3.3.2.4 below collates the cDNA samples' Cts and their relative estimated concentration.

Table 3.3.2.4 qPCR Batch 2 quantified cDNA: The table below shows the Cts and estimated cDNA copy numbers as estimated using the second batch's standard curve shown above. All cDNA samples were successfully quantified. The sample copy number was again calculated by multiplying the cDNA copy number by 7 (see table 3.3.2.2).

Label	Ct	Copy No.	Sampled Copy No.
A158-1	33.106	158.06	1106.43
A158-2	32.442	249.79	1748.55
A202-1	28.977	2722.57	19057.98
A202-2	28.836	2999.55	20996.84
A207-1	29.916	1425.49	9978.46
A207-2	28.776	3127.32	21891.24

The dissociation curves of the amplified cDNA samples are seen below in figure 3.3.2.4. As previously described, our specific amplicon has an average melting temperature of roughly 79 °C (as seen in figure 3.3.2.2). The specific dissociation temperature of each amplicon is determined by its nucleotide composition as G/C rich amplicons will have a higher melting temperature than its analogues. Similarly, amplicons from various viral species will have slightly varied compositions and thus varied dissociation temperatures.

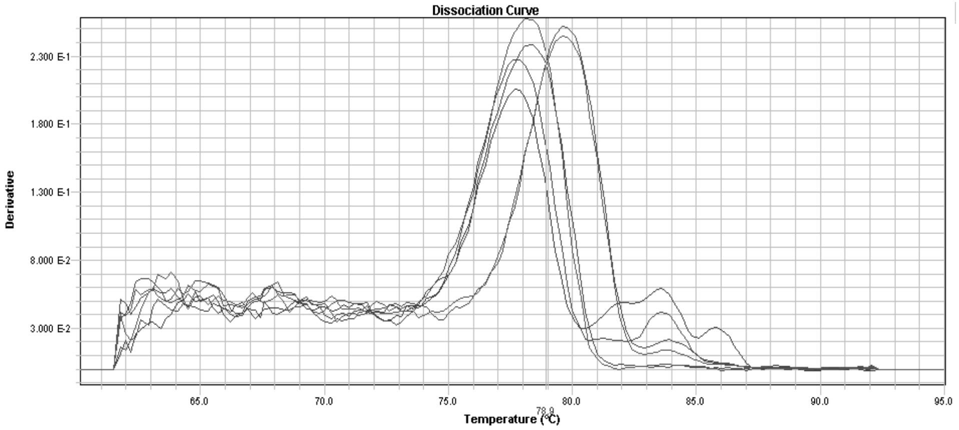


Figure 3.3.2.4 qPCR Batch 2 melt-curve analysis: The line graph above displays the dissociation curve of the amplicons produced from patient-derived cDNA sample. The green vertical line indicates the average melting temperature for specific amplicons at approximately 79 °C. The additional peaks seen at approximately 84 °C suggests the presence of larger, nonspecific amplicons or spurious G/C rich amplicons which could skew the quantification of sample A158.

The final batch of qPCR results are seen in figures 3.3.2.5 and 3.3.2.6, and tables 3.3.2.5 and 3.3.2.6 below. The cDNA duplicates from patients A297, A300, A302, A312, A313, A318, and A326 were quantified in this third batch using the linearized subtype C plasmid as triplicate DNA standards. Once again, standard curve acceptance was based on the previously mentioned criteria (section 3.3.2 above).

Table 3.3.2.5 Quantified plasmid standards: The table below shows the average of the triplicate standards, their measured Ct values and their input copy number.

Label	Content	Ct	Copy No.
5x10^1	Standard	34.793	50
5x10^2	Standard	31.743	500
5x10^3	Standard	28.079	5000
5x10^4	Standard	24.431	50000

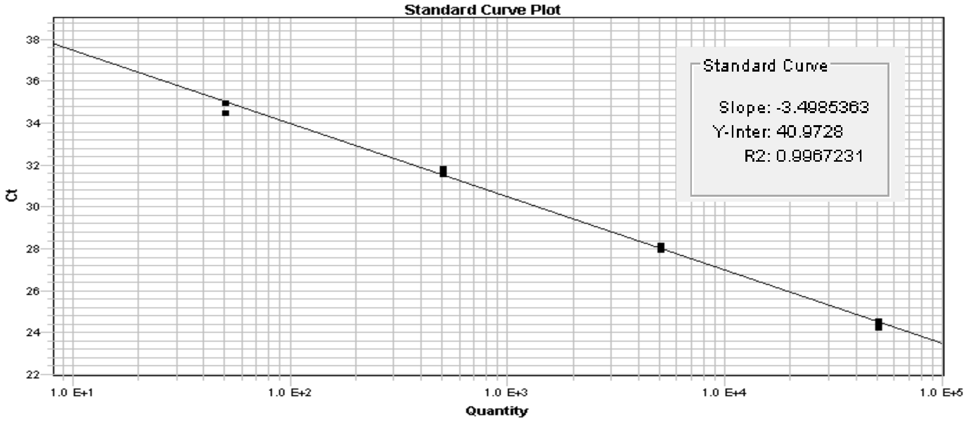


Figure 3.3.2.5 Batch 3 qPCR standard curve: Above is the straight line graph of the observed pZAC plasmid standards' Cts in relation to their concentration. The gradient of the slope, R^2 value and the Y-intercept were within our acceptable range.

The tabulated data below shows the HIV cDNA copy number in samples A297, A300, A302, A312, A313, A318, and A326 duplicates, as estimated using our in-house qPCR (table 3.3.2.6).

Table 3.3.2.6 qPCR Batch 3 quantified cDNA: The table below shows the Cts and estimated cDNA copy numbers as estimated using the standard curve shown above. All cDNA samples were successfully quantified. Sampled copy number is calculated by multiplying the estimated cDNA copy number by seven (as in table 3.3.3.2 above).

Label	Ct	Copy No.	Sampled Copy No.
A297-1	32.915	201.05	1407.35
A297-2	34.216	85.36	597.54
A300-1	36.981	13.83	96.83
A300-2	36.694	16.72	117.02
A302-1	31.789	421.65	2951.53
A302-2	32.207	320.39	2242.72
A312-1	35.100	47.72	334.04
A312-2	34.925	53.54	374.77
A313-1	31.313	576.96	4038.75
A313-2	31.586	482.08	3374.53
A318-1	34.723	61.14	427.98
A318-2	32.540	257.22	1800.55
A326-1	26.949	10196.17	71373.20
A326-2	26.415	14493.81	101456.64

A dissociation curve gives some insight into the specificity of the qPCR reaction and impacts the confidence in the estimated copy number. For example: if a nonspecific peak is seen to be higher or of the same height as the specific peak in the dissociation curve, there is no means of estimating the contribution of the nonspecific fluorescence to that well reaching Ct. Hydrolysis probes are used as an alternative to intercalating dye based qPCR as the probes bind to specific genomic targets and quantify only amplification of that target. Most commercial qPCR assays employ hydrolysis probe technology which is prone to under quantification if the probe target is sufficiently uncomplimentary to the probe. The dissociation curve in figure 3.3.2.6 below shows high, nonspecific peaks contributing to the fluorescence of cDNA duplicates from samples A297, A302 and A312. The source of these nonspecific products are unknown but could be the result of pentadecamers carried over from the reverse transcription. In the subsequent qPCR they prime the cDNA template in random areas, generating nonspecific PCR amplicons.

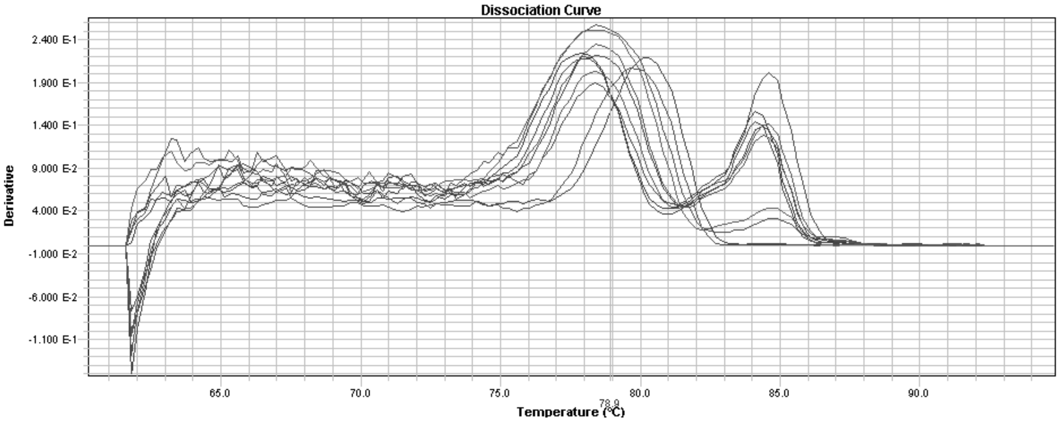


Figure 3.3.2.6 qPCR Batch 3 melt curve analysis: The above line graph displays the dissociation curve of the amplicons produced from patient-derived cDNA samples. The green vertical line indicates the expected melting temperature for specific amplicons at approximately 79 °C. The additional peaks seen at approximately 84 °C suggests the presence of larger, nonspecific amplicons or spurious G/C rich amplicons which could have resulted in overquantification of cDNA duplicates A297, A302 and A312.

3.3.3 Nested and re-PCR

The full volume of each nested PCR was separated in a 1% agarose gels made in sodium borate buffer and stained with 1 x GelStar (Lonza). GelStar is a DNA gel stain that permits DNA visualization using blue light with an orange filter as opposed to ultraviolet light which degrades DNA. Nested PCRs with less than five out of seven positive results were repeated to exclude operator error. However, the repeated nested PCRs remained negative and these samples were excluded from further analysis. Gel figure 3.3.3.1 below displays the results of the successfully amplified samples A297, A300, A312, A313, A318 and A326. The gel figures of the remaining nine samples (A124, A137, A144, A157, A158, A202, A207, A302 and A364) are appended at the end of this chapter (Appendix C).

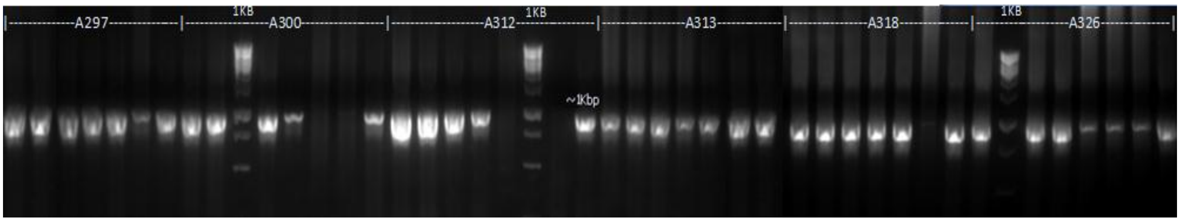


Figure 3.3.3.1 Concatenated nested PCR gels: The above figure depicts the results of four joined electrophoresis gel photographs with excluded samples removed. This figure shows successful amplification of the 1 Kbp nested PCR amplicon from samples A297, A300, A312, A313, A318 and A326. Half of the samples were positive in all seven nested PCRs, A318 was positive for six reactions and A300 and A312, positive for five.

Separated amplicons were immediately gel purified as described in the methods (section 3.2.2.8 above).

3.3.4 Amplicon quantification before sequencing

Once gel extracted, 2 µl of the 15 size-selected, patient-derived sample amplicons were quantified in triplicate on the NanoDrop® ND1000 (Thermo Scientific, Massachusetts, USA). The concentrations and purity measurements are tabulated below (table 3.3.4.1):

Table 3.3.4.1 Concentrations of sequence amplicons: The table below shows the number of positive nested PCRs per sample and the results of the spectrophotometric quantification of the purified amplicons. Preferable 260/280 and 260/230 ratios are 1.8 and 2 respectively, achievable when purifying DNA directly from the PCR product. While our samples displayed favourable 260/280 values, the carried over electrophoresis buffer contributes to the optical density seen at 230nm, resulting in lower than expected 260/230 ratios.

Sample	No of PCR's Positive	Average gel purified product values			
		conc. (ng/µl)	A260	260/280	260/230
A124	07 of 07	33.25	0.66	1.99	1.50
A137	07 of 07	39.89	0.80	1.95	1.59
A144	07 of 07	56.78	1.14	1.92	1.13
A157	07 of 07	49.86	1.00	1.98	1.56
A158	07 of 07	38.73	0.77	1.92	1.24
A202	07 of 07	27.86	0.56	1.86	1.02
A207	06 of 07	37.13	0.74	1.86	1.46
A297	07 of 07	45.45	0.91	1.88	1.65
A300	05 of 07	25.60	0.51	1.86	1.21
A302	07 of 07	20.21	0.40	1.84	1.06
A312	05 of 07	41.41	0.83	1.94	1.65
A313	07 of 07	35.67	0.71	1.85	1.70
A318	06 of 07	34.08	0.68	1.87	1.27
A326	07 of 07	30.17	0.60	1.84	1.41
A364	07 of 07	25.82	0.52	1.91	0.93

3.3.5 Diagnostic drug resistance testing results

Using in-house drug resistance testing (DRT), only two patients had major NNRTI resistance mutations: Patient A157 had K103N and A313 had Y181I. These mutations were also detected in the other sequencing events: K103N was quantified at 73.8% by the PGM (Life Technologies, California, USA), 51.6% by the MiSeq (Illumina, California, USA) and 47.9% by clonal sequencing, while Y181I was seen at 72.4%, 73.3% and 74.4% for the respective platforms.

3.3.6 PGM results summary

Considering the coverage for both the PGM (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) data sets, and the high cDNA yield in most cases, a threshold of 0.5% was used to distinguish minor variants. In addition to the NNRTI mutation detected by Sanger sequencing, the PGM (Life Technologies, California, USA) detected K103N in patient A124 and Y181C in A158. The median read coverage (the median number of times a locus is sampled) on the PGM (Life Technologies, California, USA) was seen at 8 939 (IQR: 4 521 - 12 585) and the median proportion of bases with quality scores of 30 or greater (error ≤ 0.1%) was 43.2% (40.5–46.0%). Table 3.3.6.1 below displays the post quality filtering median coverage per amino acid while figure 3.3.6.1 shows the coverage for all samples across the sequence amplicon.

Table 3.3.6.1 PGM per codon coverage: The table below shows the average and median coverage per codon for each patient sample. The values in the table were calculated after quality filtering and mapping.

Coverage:	Average	Median
A124	11525.03	12284.0
A137	12316.49	11660.5
A144	16534.87	17785.0
A157	4333.01	4263.5
A158	4583.94	4370.0
A202	14194.31	13535.0
A207	4026.27	3699.0
A297	9250.93	9843.5
A300	2484.74	2379.5
A302	6343.35	6358.5
A312	8680.82	7295.0
A313	9197.21	8893.5
A318	7063.44	6095.5
A326	4011.135	3296.0
A364	13389.22	11482.0

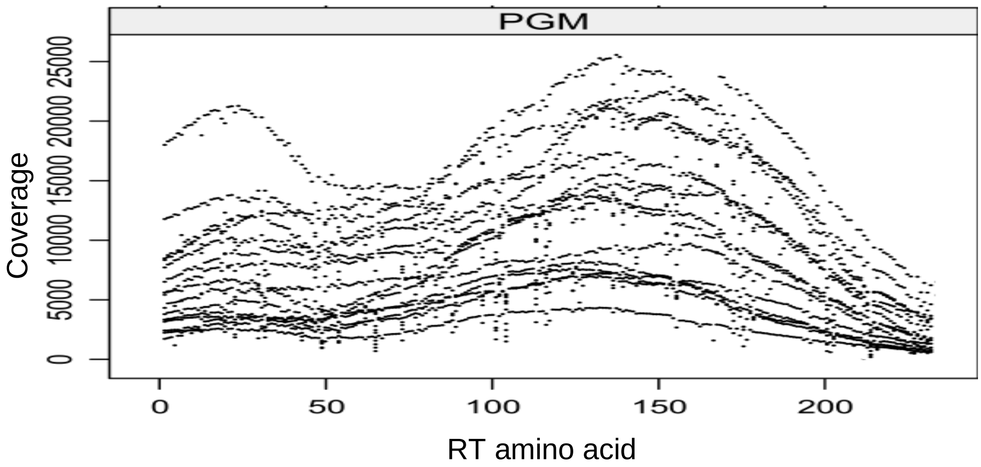


Figure 3.3.6.1 PGM sample coverage: This figure shows the amino acid coverage for each patient. This figure was constructed in R-commander by plotting each codon's coverage values against the first approximately 240 amino acids of the HIV-1 *Reverse Transcriptase* gene.

3.3.7 MiSeq results summary

The median read coverage with MiSeq (Illumina, California, USA) was 74 181 (IQR: 50 173 - 104 089) and the median proportion of bases with <0.1% error was 89.1% (84.7 – 93%). Table 3.3.7.1 below displays the post quality filtering median coverage per amino acid while figure 3.3.7.1 shows the coverage for all samples across the sequence amplicon.

Table 3.3.7.1 MiSeq per codon coverage: This table shows the average and median coverage per codon for each patient sample. The values in the table were calculated after quality filtering and mapping.

Coverage:	Average	Median
A124	33173.74	33037.00
A137	31668.16	36875.00
A144	20905.12	23753.00
A157	106433.03	120014.00
A158	107216.38	122231.50
A202	112606.88	127778.00
A207	84326.64	93773.50
A297	81423.86	86060.00
A300	66938.11	73928.50
A302	55838.76	61950.00
A312	92616.70	103221.50
A313	57013.34	64062.50
A318	105063.56	118473.50
A326	103764.10	108687.50
A364	32921.07	32504.00

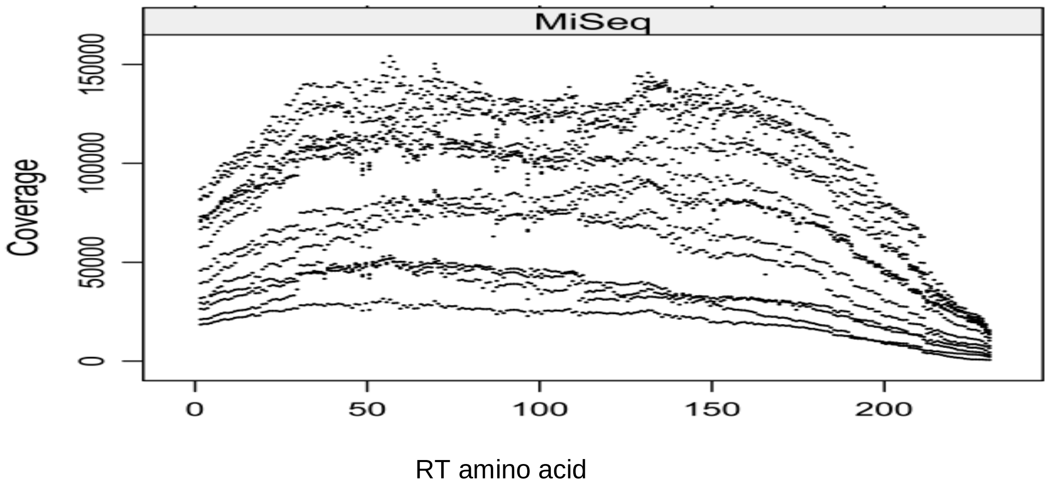


Figure 3.3.7.1 MiSeq sample coverage: The above figure shows the amino acid coverage for each patient of the combined forward and reverse MiSeq reads. This figure, constructed in R-commander, plots codon coverage values against the first approximately 240 amino acids of the HIV-1 *Reverse Transcriptase* gene.

3.3.8 Maximum likelihood tree

The NGS consensus sequences and bulk sequencing, Sanger sequences formed well-supported, monophyletic groups, as seen in figure 3.3.8.1 below. MiSeq (Illumina, California, USA) consensus sequences were in much better agreement with Sanger sequences than PGM's (Life Technologies, California, USA). When considering the chromatograms used to resolve mixed bases in the Sanger sequences, the bulk of Sanger sequences and MiSeq (Illumina, California, USA) consensus sequences were identical, whereas the mean difference between PGM (Life Technologies, California, USA) consensus and Sanger sequences for a given patient were 2.2/630 nucleotides (IQR:1–3, maximum 4).

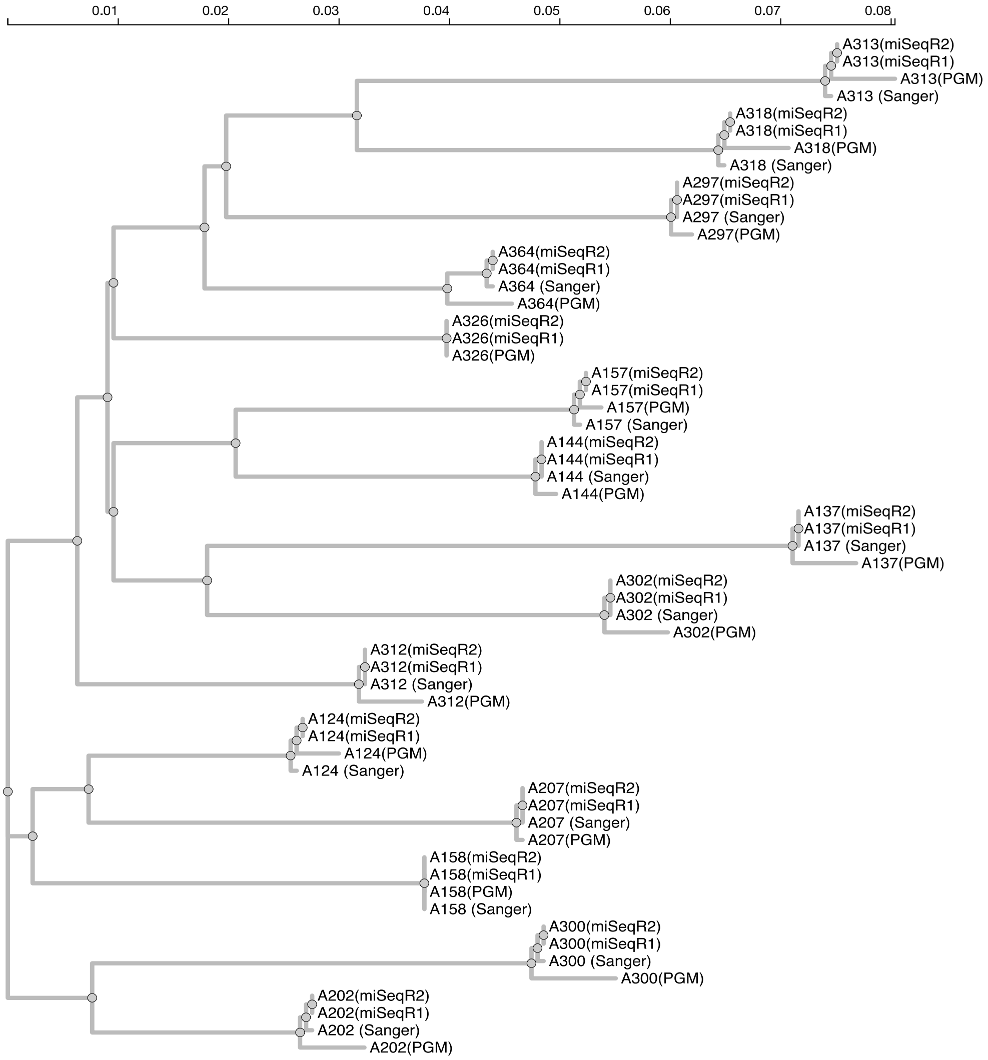


Figure 3.3.8.1 Maximum likelihood tree of inferred evolutionary relationship: A maximum likelihood phylogenetic tree of NGS consensus and Sanger sequences (GTR + CAT model in FastTree). In the tree above, MiSeq R1 and R2 indicate the forward and reverse reads respectively. The scale is expected substitutions per site per unit of time. Each intra-host clade has a bootstrap support of >95%.

3.3.9 Clonal sequencing results

Two hundred and fifty recombinants were sequenced for A124, A157, A158 and 269 recombinants for A313. Clonal sequencing was able to detect DRMs at less than 1% frequencies and confirmed the minor variant population identified in the PGM (Life Technologies, California, USA) sequencing. There was better agreement between the clonal sequencing and the MiSeq (Illumina, California, USA) data than with the clonal sequencing and the PGM (Life Technologies, California, USA) data as shown in table 3.3.10.1 below (at the end of the results section).

3.3.10 Comparative DRT results

Table 3.3.10.1 displays the comparative DRT results across all four sequencing events: our in-house diagnostic DRT, Ion Torrent PGM (Life Technologies, California, USA) amplicons sequencing, MiSeq (Illumina, California, USA) amplicon sequencing, and recombinant sequencing. For both NGS events, the table displays the median read coverage per template and the percentage template sampling ratio. This ratio is calculated by dividing the median read coverage per template by the sampled cDNA copy number provided by the qPCR.

Table 3.3.10.1 Comparative DRT across sequencing events: Four of the five patients shown to have diverse viral populations were included for clonal sequencing (i.e.: A124, A157, A158, and A313 but not A300). Major variant drug resistance mutations were detected by genotyping with an bulk sequencing (Sanger DRM). Major and minor variant drug resistance mutations detected with the PGM are labeled as PGM DRMs and those detected with the MiSeq, MiSeq DRMs. The DRMs respective frequencies, at a threshold of >0.5%, are shown in parenthesis. A364 cDNA was not quantified and thus the percentage template sampling ratio could not be calculated. K65R was detected in all samples; and is commonly detected by deep sequencing in HIV-1 subtype C. Clonal sequencing confirmed the presence of DRMs >1%. Agreement with clonal sequence frequency was better for MiSeq than Ion PGM.

Patient No	Sanger DRM	PGM DRMs	PGM: median coverage (per template)	MiSeq DRMs	MiSeq: median coverage (per template)	Clonal sequencing
A124	None	K65R (1.4%), K103N (32.0%)	5748 (0.9)	K65R (1.0%), K103N (15.7%)	33,037 (5.2)	K65R (0.4%), K103N (11.2%)
A137	None	K65R (1.2%)	11,429 (9.7)	K65R (1.0%)	36,875 (31.2)	NA
A144	None	K65R (2.0%)	17,020 (6.2)	K65R (1.8%)	23,753 (8.7)	NA
A157	K103N	K65R (1.4%), K103N (73.8%), V106M (1.4%), V106A (1.0%), Y181C (4.4%)	4 210 (3.8)	K65R (0.9%), K103N (51.6%), V106M (1.4%), V106A (1.0%), Y181C (4.4%)	120,014 (107.1)	K65R (0.4%), K103N (47.9%), V106M (0.4%), V106A (0.4%), Y181C (0.82%)
A158	None	K65R (2.9%), Y181C (2.0%)	4 190 (1.3)	K65R (1.3%), Y181C (2.6%)	122,232 (37.4)	K65R (1.2%), Y181C (1.2%)
A202	None	K65R (1.9%)	13,178 (0.3)	K65R (0.7%)	127,778 (2.9)	NA
A207	None	K65R (1.2%)	3 677 (0.1)	K65R (1.1%)	93,774 (2.8)	NA
A297	None	K65R (1.3%)	9 781 (4.6)	K65R (1.0%)	86,060 (40.5)	NA
A300	None	K65R (2.5%)	2 387 (9.8)	K65R (1.1%), G190E (0.7%)	73,929 (304.2)	NA
A302	None	K65R (1.5%)	6 155 (1.1)	K65R (1.5%)	61,950 (11.5)	NA
A312	None	K65R (1.4%)	6 925 (8.9)	K65R (1.1%)	103,222 (132.8)	NA
A313	Y181I	K65R (1.2%), Y181I (72.4%)	8 545 (1.1)	K65R (0.7%), Y181I (73.3%)	64,063 (8.5)	K65R (0.7%), Y181I (74.4%)
A318	None	K65R (1.5%)	5 940 (2.5)	K65R (1.1%)	118,474 (50.5)	NA
A326	None	K65R (1.2%)	3 296 (0.02)	K65R (1.1%)	108,688 (0.7)	NA
A364	None	K65R (0.9%)	11,218 (NA)	K65R (0.9%)	39201(NA)	NA

3.4 Discussion

To date, the present study was the only one comparing two modern NGS platforms for DRT of infants failing PMTCT, to our knowledge. More recently, viral quasispecies from 19 pre-therapy patients were sequenced on 454 (454 Life Sciences, Connecticut, USA) and on Illumina's MiSeq (California, USA) (Keys et al. 2015), however, these were adult patients who were not PMTCT exposed and DRM transmission was horizontal. This study employed PCR enrichment and the ligation of sequencing adaptors, which is an alternative method to the conventional fusion primer PCR. Since the commercial ligation kits are costly, the fusion primer method is more feasible for targeted resequencing of small genome targets such as for HIV drug resistance genotyping. However, fusion primers are especially prone to resampling error (discussed in more detail in the next chapter). Nevertheless, using NGS this study improved the DRT in 15 infants who became infected despite being exposed to a PMTCT regimen of maternal and infant AZT and NVP.

Sanger genotyping before cART initiation detected major NNRTI DRMs, K103N and Y181I in A157 and A313, respectively, while no patients had minor or major variant AZT-associated DRMs. Ion Torrent sequencing detected K103N in A124; Y181C and V106A/M in A157; and Y181C in A158 in addition to the Sanger DRMs. The clinical implication of detecting Y181C in A157 would be reduced susceptibility to the second-line NNRTI, etravirine, since the presence of drug-resistant, minor populations have been linked to therapy failure (Li et al. 2011). MiSeq (Illumina, California, USA) sequencing further improved DRT over Ion PGM with the detection of G190E in patient A300. The higher coverage, and higher read quality obtained in MiSeq data may have contributed to this additional DRM detection, when compared to PGM (Life Technologies, California, USA). Both the PGM (Life Technologies, California, USA) and the MiSeq (Illumina, California, USA) had good agreement with clonal sequencing but the overall agreement with clonal sequencing was better with MiSeq (Illumina, California, USA).

At the time of patient sample collection the PMTCT regimen was AZT from 28 weeks of gestation and sdNVP intra-partum, while the neonate received sdNVP and AZT for 1 week and was formula-fed to prevent post-natal transmission. This strategy reduced vertical HIV transmission to less than 10% (Draper & Abdullah 2008) but was later improved in 2010 when infant sdNVP with daily NVP for the first 6 weeks of life and the PMTCT-failure rate decreased to <3% (Barron et al. 2013). In 2013 the WHO option B plus was adopted in the Western Cape, which recommends lifelong cART for pregnant women regardless of CD4 count or disease stage, to further decrease the PTMCT failure rate.

A recent South African study reported 56.8% of children having NNRTI resistance by bulk sequencing after PMTCT exposure; however, the majority of the infants received only NVP (Kuhn et al. 2014). In a large randomized study the addition of ZDV and 3TC for four or seven days to the infants' regimens significantly reduced drug resistance when compared to sdNVP (NVP alone – 7/8 (88%), four days – 4/25 (17%) and seven days – 0/10 (0%)) (McIntyre et al. 2009). Data on the prevalence of minor DRMs after PMTCT remains limited and most studies of minor variant DRMs after PMTCT used allele-specific PCR targeting only a few important mutations (Rowley et al. 2010; McMahon et al. 2013; Boltz et al. 2012). A larger global study (ACTG A5175) not only stressed the value of viral genotyping before therapy initiation (wherever possible), but also demonstrated that subtype C infected patients are at a higher risk of disease progression, regardless of ART (Kantor et al. 2015).

Special precautions were taken to reduce both RT sampling bias and PCR resampling, through the use of random cDNA primers and partitioning the PCR assay into a few parallel reactions. We used random pentadecamers to reduce our RT selection bias and yielded longer cDNA fragments. Other researchers suggest that pentadecamer-mediated RT produces cDNA more efficiently than random hexamers, when used in combination with Invitrogen's SuperScript enzyme (California, USA) (Stangegaard et al. 2006) and at higher concentrations (Nardon et al. 2009). Moreover, pentadecamers have been shown to improve assay limit of detection in quantitative RT-PCR for genomics studies (Ross et al. 2008).

Cabrera et al were aware of the primer-induced RT selection biases and used three RT replicates to increase the number of sampled variants (Cabrera et al. 2006), while others tried to circumvent early primer selection bias by increasing the number of PN-PCR replicates, which they later pooled before nested amplification (Mavigner et al. 2009). In the present study duplicate random pentadecamer primed reverse transcriptions were done with low selection bias, followed by 14 PN-PCRs. Due to selection events occurring early which result in diminished population variants, duplicate reactions were only pooled after pre-nesting. Alternative methods for reduced PCR resampling include limiting dilution PCR (Salgado et al. 2010) and single molecule emulsion PCR (Metzger et al. 2013). However, both of these methods are laborious, either incurring exorbitant reagent costs or require specialized equipment.

More recently, random primer-ID tagging during reverse transcription has been used to allow post-hoc correction of PCR resampling bias (Jabara et al. 2011). Using this approach, the quantity of variants is not calculated from the final proportion of variants (after PCR enrichment) in the sequenced population, but by counting the number of unique species that share the same random primer ID and which represents the original pre-enriched population (Jabara et al. 2011). The primer ID approach improved accuracy of minor variant quantification by reducing both PCR resampling and PCR-induced error (Zhou et al. 2015; Keys et al. 2015). However, it has been shown to substantially underestimate the number of variants in the original population, as it cannot obviate PCR bias introduced by the use of large fusion primers (Boltz et al. 2015). Moreover, the primer ID technique is applicable only for patients with sufficiently high viral loads so as to allow sufficient cDNA recovery. As primer ID requires a highly-specific chromatographic separation of the cDNA and large fusion primers, the method can be technically challenging, since any downstream ID primer contamination would invalidate the results.

Forty-six samples were initially included in our cohort however two thirds of the reactions failed to generate a specific amplicon for five of the seven nested PCR's. This result could possibly be due to a number of confounding factors: the poor sensitivity of the high fidelity pre-nested PCR; the fragmentation of the extracted RNA during sample storage; or the generation of incomplete or truncated cDNA templates. When the number of variants sampled is close to the limit of assay detection, very few variants are selected and post sequencing analysis could overestimate the number of variants sampled. In these instances, many of the observed minor variant mutations are as a result of PCR error and recombination. To this end, samples that failed to amplify in more than five of the seven nested PCRs were excluded.

Another major limitation of 454 (454 Life Sciences, Connecticut, USA) and PGM (Life Technologies, California, USA) sequencing is homopolymer read error (Loman et al. 2012), which has been associated with false positive detection of K65R in HIV-1 subtype C using the 454 platform (454 Life Sciences, Connecticut, USA) (Varghese et al. 2010). On both the 454 (454 Life Sciences, Connecticut, USA) and PGM (Life Technologies, California, USA), nucleotides are sequentially presented to the polymerase and incorporation is measured by monitoring the increase in luminescence on the 454 (454 Life Sciences, Connecticut, USA) or pH increase on the PGM (Life Technologies, California, USA) (Marinier et al. 2015). The fold-increase in these parameters is translated into the number of bases incorporated, resulting in difficulty in distinguishing the incorporation of more than five of the same bases. The frequency of K65R in our study was higher in PGM (Life Technologies, California, USA) reads than MiSeq (Illumina, California, USA), suggesting that homopolymer read error contributed to K65R variant calling.

Nevertheless, K65R was also present, albeit at lower frequency, in MiSeq (Illumina, California, USA) and clonal sequences. Previous reports have indicated that using high fidelity PCR enzymes could limit the proportion of K65R minor variants (Varghese et al. 2010). Despite the use of high fidelity PCR, K65R >0.5% was detected in all infants, which would be unexpected in a tenofovir-naïve cohort (Recordon-Pinson et al. 2012). NVP-containing regimens have been associated with an increased risk of K65R detection (Brenner & Coutsinos 2009; Tang et al. 2013), though the mechanism has not yet been described. Overall, while HIV-1 subtype C has an increased risk of K65R under drug pressure (Brenner & Coutsinos 2009; Coutsinos et al. 2011), the proportion of these K65R mutations that occur due to in vivo rather than in vitro reverse transcription error, is unclear.

3.5 Conclusion

Our study showed that NGS using PGM (Life Technologies, California, USA) or MiSeq (Illumina, California, USA), combined with a bioinformatic ‘pipeline’ could enable the detection of minor variant reverse transcriptase DRMs after PMTCT. The read quality was best for MiSeq (Illumina, California, USA), probably as it is less prone to homopolymer error, and the higher coverage increased the confidence of minor variant calling. As NGS platforms become more affordable they may prove invaluable in the investigation of infants, where antiretroviral therapy for PMTCT failed to prevent transmission, for the following reasons:

- NNRTI minor variant DRMs are associated with subsequent failure on a particular regimen (Paredes et al. 2010; Li et al. 2012),
- PMTCT failure of regimens relying on maternal cART may be associated with an increased risk of complex resistance patterns, and
- DRM detection using NGS, in contrast to allele-specific PCR, allows for detection of all DRMs across a particular sequence.

Laboratories that process PCR-enriched microbial or viral samples (such as HIV) do not require onsite NGS capability but could make use of commercial or academic core facilities that provide NGS on a fee-for-service basis. Nevertheless it is critical that a specimen which is PCR-enriched for NGS should be a representative sample of the source: in this study cDNA was quantified to confirm that the sample contained sufficient species to enable minor variant detection and amplification was separated in parallel PCRs to limit random PCR error. Of equal importance is the post-analytical processing of NGS data to correctly identify haplotypes (variants) from PCR and sequencing error.

Although several published programs are freely available, they often require some programming capability and the integration across different bioinformatics platforms. A versatile and openly available bioinformatics ‘pipeline’ for processing of NGS sequences would be greatly beneficial. This study’s ‘pipeline’ combined the following processing and analysis functions: exclusion of low quality reads, correction for homopolymer errors or random sequencing error and construction of consensus reads and frequent haplotypes (variants). The output of the analysis performed for this study is available. (http://bit.ly/gvz-PMTCT-NGS).

Online access to an open source service that performs quality control and trouble-shooting of next generation sequencing data would be highly valuable in providing increased access to this technology. The recent introduction of Seq2Res and RAMICS allows for relatively easy HIV drug resistance analysis from NGS data from multiple deep sequencing platforms (Wright & Travers 2014). To our knowledge, this is one of the first widely accessible analysis pipelines for HIV-specific NGS analyses, available online at hiv.sanbi.ac.za/tools#/seq. However, the bioinformatic pipeline, used in the present study is available in the public domain at hyphy.org/w/index.php/Main_Page.

3.6 References

Archer, J. et al., 2012. Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PloS one*, 7(11), p.e49602.

Barron, P. et al., 2013. Eliminating mother-to-child HIV transmission in South Africa. *Bulletin of the World Health Organisation*, 91(1), pp.70–74.

Bennett, D.E. et al., 2009. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PloS one*, 4(3), p.e4724.

Boltz, V.F. et al., 2015. Analysis of Resistance Haplotypes Using Primer IDs and Next Gen Sequencing of HIV RNA Methods. *CROI*, p.1. Available at: http://www.croiconference.org/sites/default/files/posters-2015/593.pdf [Accessed July 19, 2015].

Braithwaite, R.S. et al., 2011. Alternative antiretroviral monitoring strategies for HIV-infected patients in east Africa: opportunities to save more lives? *Journal of the International AIDS Society*, 14, p.38.

Brenner, B.G. & Coutsinos, D., 2009. The K65R mutation in HIV-1 reverse transcriptase: genetic barriers, resistance profile and clinical implications. *HIV therapy*, 3(6), pp.583–594.

Cabrera, C. et al., 2006. Genetic evolution of gp41 reveals a highly exclusive relationship between codons 36, 38 and 43 in gp41 under long-term enfuvirtide-containing salvage regimen. *AIDS (London, England)*, 20(16), pp.2075–2080.

Chang, M.W. et al., 2013. Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *Journal of virological methods*, 189(1), pp.232–234.

Chaturbhuj, D.N. et al., 2014. Evaluation of a Cost Effective In-House Method for HIV-1 Drug Resistance Genotyping Using Plasma Samples. *PLoS ONE*, 9(2), p.e87441.

Claassen, M., van Zyl, G.U. & Engenlbrecht, S. 2011. *In-house Genotypic Antiretroviral Resistance Test: Optimisation and Validation for Use in Research and by Mathilda Claassen*. Stellenbosch University. Available at: http://scholar.sun.ac.za/handle/10019.1/42809?show=full.

Coutsinos, D. et al., 2011. A template-dependent dislocation mechanism potentiates K65R reverse transcriptase mutation development in subtype C variants of HIV-1. *PloS one*, 6(5), p.e20208.

Draper, B. & Abdullah, F., 2008. A review of the prevention of mother-to-child transmission programme of the Western Cape provincial government, 2003 - 2004. *South African medical journal*, 98(6), pp.431–434.

Estill, J. et al., 2013. Monitoring of antiretroviral therapy and mortality in HIV programmes in Malawi, South Africa and Zambia: mathematical modelling study. *PloS one*, 8(2), p.e57611.

Hattori, J. et al., 2010. Trends in transmitted drug-resistant HIV-1 and demographic characteristics of newly diagnosed patients: nationwide surveillance from 2003 to 2008 in Japan. *Antiviral research*, 88(1), pp.72–79.

Hunt, G.M. et al., 2014. Concordance between allele-specific PCR and ultra-deep pyrosequencing for the detection of HIV-1 non-nucleoside reverse transcriptase inhibitor resistance mutations. *Journal of virological methods*, 207, pp.182–187.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Jacobs, G.B. et al., 2012. Construction of a High Titer Infectious HIV-1 Subtype C Proviral Clone from South Africa. *Viruses*, 4(9), pp.1830–1843.

Jourdain, G. et al., 2010. Association between detection of HIV-1 DNA resistance mutations by a sensitive assay at initiation of antiretroviral therapy and virologic failure. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 50(10), pp.1397–1404.

Kantor, R. et al., 2015. Pretreatment HIV Drug Resistance and HIV-1 Subtype C Are Independently Associated With Virologic Failure: Results From the Multinational PEARLS (ACTG A5175) Clinical Trial. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 60(10), pp.1541–1549.

Keys, J.R. et al., 2015. Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS research and human retroviruses*, 31(6), pp.658–668.

Kijak, G.H. et al., 2014. Targeted deep sequencing of HIV-1 using the IonTorrentPGM platform. *Journal of virological methods*, 205C, pp.7–16.

Kosakovsky Pond, S.L. et al., 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS computational biology*, 5(11), p.e1000581.

Kuhn, L. et al., 2014. Drug resistance among newly diagnosed HIV-infected children in the era of more efficacious antiretroviral prophylaxis. *AIDS (London, England)*, 28(11), pp.1673–1678.

Li, J.Z. et al., 2013. Impact of minority nonnucleoside reverse transcriptase inhibitor resistance mutations on resistance genotype after virologic failure. *The Journal of infectious diseases*, 207(6), pp.893–897.

Li, J.Z. et al., 2011. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA*, 305(13), pp.1327–1335.

Li, J.Z. et al., 2012. Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS (London, England)*, 26(2), pp.185–192.

Liu, S.L. et al., 1996. HIV quasispecies and resampling. *Science (New York, N.Y.)*, 273(5274), pp.415–416.

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–439.

Marinier, E., Brown, D.G. & McConkey, B.J., 2015. Pollux: platform independent error correction of single and mixed genomes. *BMC bioinformatics*, 16, p.10.

Mavigner, M. et al., 2009. HIV-1 residual viremia correlates with persistent T-cell activation in poor immunological responders to combination antiretroviral therapy. *PloS one*, 4(10), p.e7658.

McIntyre, J.A. et al., 2009. Efficacy of short-course AZT plus 3TC to reduce nevirapine resistance in the prevention of mother-to-child HIV transmission: a randomized clinical trial. *PLoS medicine*, 6(10), p.e1000172.

McMahon, D.K. et al., 2013. Greater suppression of nevirapine resistance with 21- vs 7-day antiretroviral regimens after intrapartum single-dose nevirapine for prevention of mother-to-child transmission of HIV. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 56(7), pp.1044–1051.

Metzger, B.P.H., Gelembiuk, G.W. & Lee, C.E., 2013. Direct sequencing of haplotypes from diploid individuals through a modified emulsion PCR-based single-molecule sequencing approach. *Molecular ecology resources*, 13(1), pp.135–143.

Morey, M. et al., 2013. A glimpse into past, present, and future DNA sequencing. *Molecular genetics and metabolism*, 110(1-2), pp.3–24.

Nardon, E. et al., 2009. Higher random oligo concentration improves reverse transcription yield of cDNA from bioptic tissues and quantitative RT-PCR reliability. *Experimental and molecular pathology*, 87(2), pp.146–151.

Palmer, S. et al., 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *Journal of clinical microbiology*, 43(1), pp.406–413.

Palmer, S. et al., 2012. Short-course Combivir after single-dose nevirapine reduces but does not eliminate the emergence of nevirapine resistance in women. *Antiviral therapy*, 17(2), pp.327–336.

Palumbo, P. et al., 2010. Antiretroviral treatment for children with peripartum nevirapine exposure. *The New England journal of medicine*, 363(16), pp.1510–1520.

Paredes, R. et al., 2010. Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *The Journal of infectious diseases*, 201(5), pp.662–671.

Phillips, A. et al., 2014. Cost-effectiveness of HIV drug resistance testing to inform switching to second line antiretroviral therapy in low income settings. *PloS one*, 9(10), p.e109148.

Ramachandran, S. et al., 2008. End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions. *Journal of virological methods*, 151(2), pp.217–224.

Recordon-Pinson, P. et al., 2012. K65R in subtype C HIV-1 isolates from patients failing on a first-line regimen including d4T or AZT: comparison of Sanger and UDP sequencing data. *PloS one*, 7(5), p.e36549.

Ross, D.M. et al., 2008. Reverse transcription with random pentadecamer primers improves the detection limit of a quantitative PCR assay for BCR-ABL transcripts in chronic myeloid leukemia: implications for defining sensitivity in minimal residual disease. *Clinical chemistry*, 54(9), pp.1568–1571.

Rowley, C.F. et al., 2010. Ultrasensitive detection of minor drug-resistant variants for HIV after nevirapine exposure using allele-specific PCR: clinical significance. *AIDS research and human retroviruses*, 26(3), pp.293–300.

Salazar, M. (CHAVI), Salazar-Gonzalez, J. (CHAVI) & McPherson, D. (CHAVI), 2007. *Standard Operating Procedure for: WHOLE GENOME AMPLIFICATION OF HIV-1 FROM A SINGLE RNA TEMPLAT*, Birmingham. Available at: http://www.uab.edu/medicine/cfar/images/chavi-mbsc-21.pdf.

Salgado, M. et al., 2010. Evolution of the HIV-1 nef gene in HLA-B*57 positive elite suppressors. *Retrovirology*, 7, p.94.

Simen, B.B. et al., 2014. An international multicenter study on HIV-1 drug resistance testing by 454 ultra-deep pyrosequencing. *Journal of virological methods*, 204, pp.31–37.

Sinunu, M.A. et al., 2014. Evaluating the impact of prevention of mother-to-child transmission of HIV in Malawi through immunization clinic-based surveillance. *PloS one*, 9(6), p.e100741.

Stangegaard, M., Dufva, I.H. & Dufva, M., 2006. Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *BioTechniques*, 40(5), pp.649–657.

Sungkanuparph, S. et al., 2012. Emergence of HIV-1 drug resistance mutations among antiretroviral-naive HIV-1-infected patients after rapid scaling up of antiretroviral therapy in Thailand. *Journal of the International AIDS Society*, 15(1), p.12.

Tang, M.W. et al., 2013. Nucleoside reverse transcriptase inhibitor resistance mutations associated with first-line stavudine-containing antiretroviral therapy: programmatic implications for countries phasing out stavudine. *The Journal of infectious diseases*, 207 Suppl , pp.S70–7.

Ton, Q. & Frenkel, L., 2013. HIV drug resistance in mothers and infants following use of antiretrovirals to prevent mother-to-child transmission. *Current HIV research*, 11(2), pp.126–136.

Varghese, V. et al., 2010. Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. *PloS one*, 5(6), p.e10992.

Violari, A. et al., 2012. Nevirapine versus ritonavir-boosted lopinavir for HIV-infected children. *The New England journal of medicine*, 366(25), pp.2380–2389.

Western Cape Provincial Govt., 2012. *Provincial Strategic Plan on HIV / AIDS , STIs and TB*, Cape Town. Available at: https://www.westerncape.gov.za/assets/departments/health/provincial_strategic_plan_on_hiv_aids_stis_tb_2012_-_2016_-_15_june_2012.pdf.

World Health Organisation, 2012. *Antiretroviral treatment as prevention (TASP) of HIV and TB.* Available at: https://www.msh.org/sites/msh.org/files/2012_dec_tasp_brief_email.pdf

Wright, I.A. & Travers, S.A., 2014. RAMICS: trainable, high-speed and biologically relevant alignment of high-throughput sequencing reads to coding DNA. *Nucleic acids research*, 42(13), p.e106.

Zhou, S. et al., 2015. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of virology*. [Cited ahead of print]

Van Zyl, G.U. et al., 2014. Emerging antiretroviral drug resistance in sub-Saharan Africa: novel affordable technologies are needed to provide resistance testing for individual and public health benefits. *AIDS (London, England)*, 28(18), pp.2643–2648.

3.7 Appendix C: Supplemental figures

(An addendum to section 3.3.3 Nested and re-PCR)

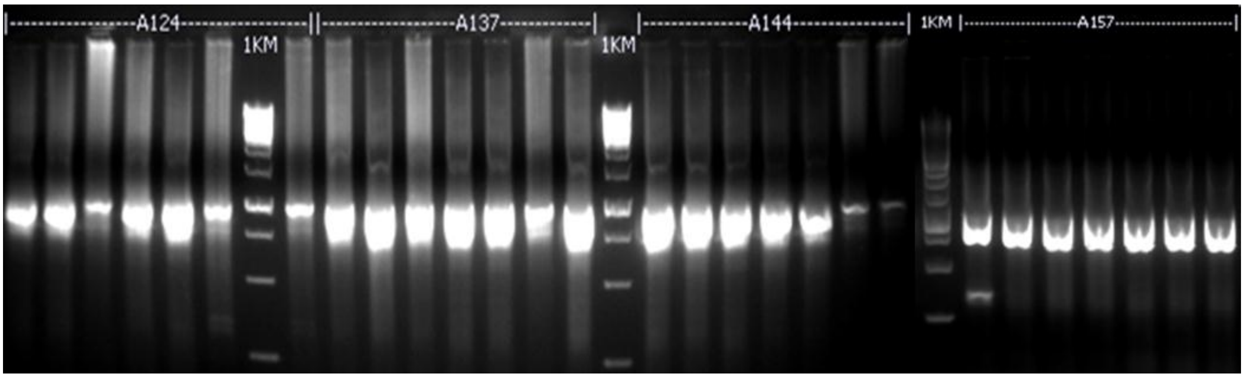


Figure 3.7.1 Concatenated nested PCR gels: The above figure depicts the results of two joined electrophoresis gel photographs with the excluded samples removed. This figure shows successful amplification of the 1 Kbp nested PCR amplicon from samples A214, A137, A144, and A157. All of these samples were positive in all seven nested PCRs, however, A144 shows inefficient amplification in two of its seven reactions.

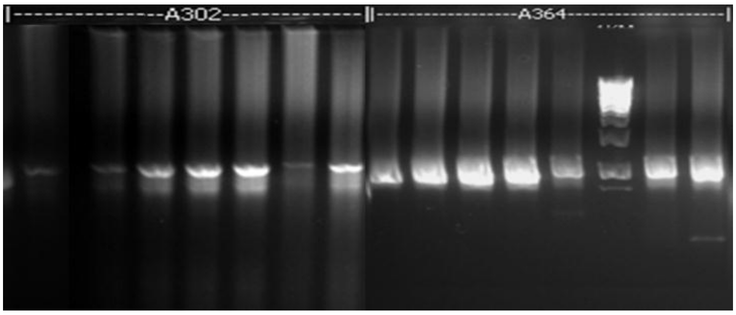


Figure 3.7.2 Concatenated nested PCR gels: The above figure depicts the results of three joined electrophoresis gel photographs with the excluded samples removed. This figure shows successful amplification of the 1 Kbp nested PCR amplicon from samples A158, A202 and A207. Two of these samples, A158 and A207 were positive for six of the seven nested PCRs while A202 shows inefficient amplification in all seven reactions.

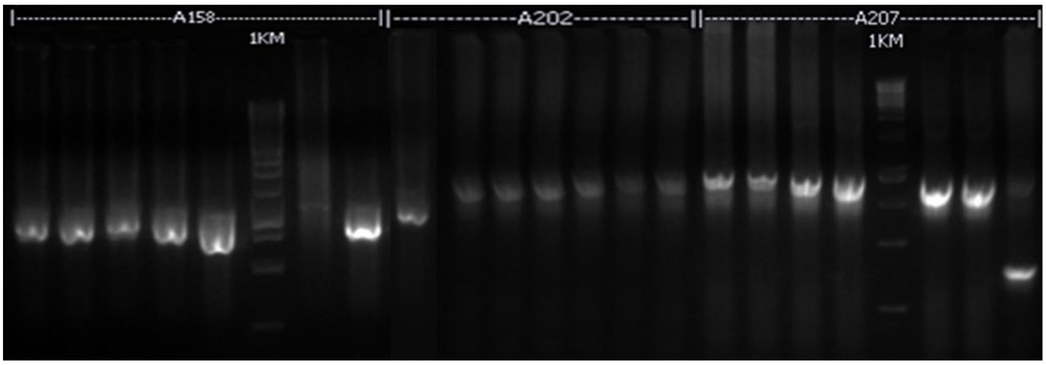
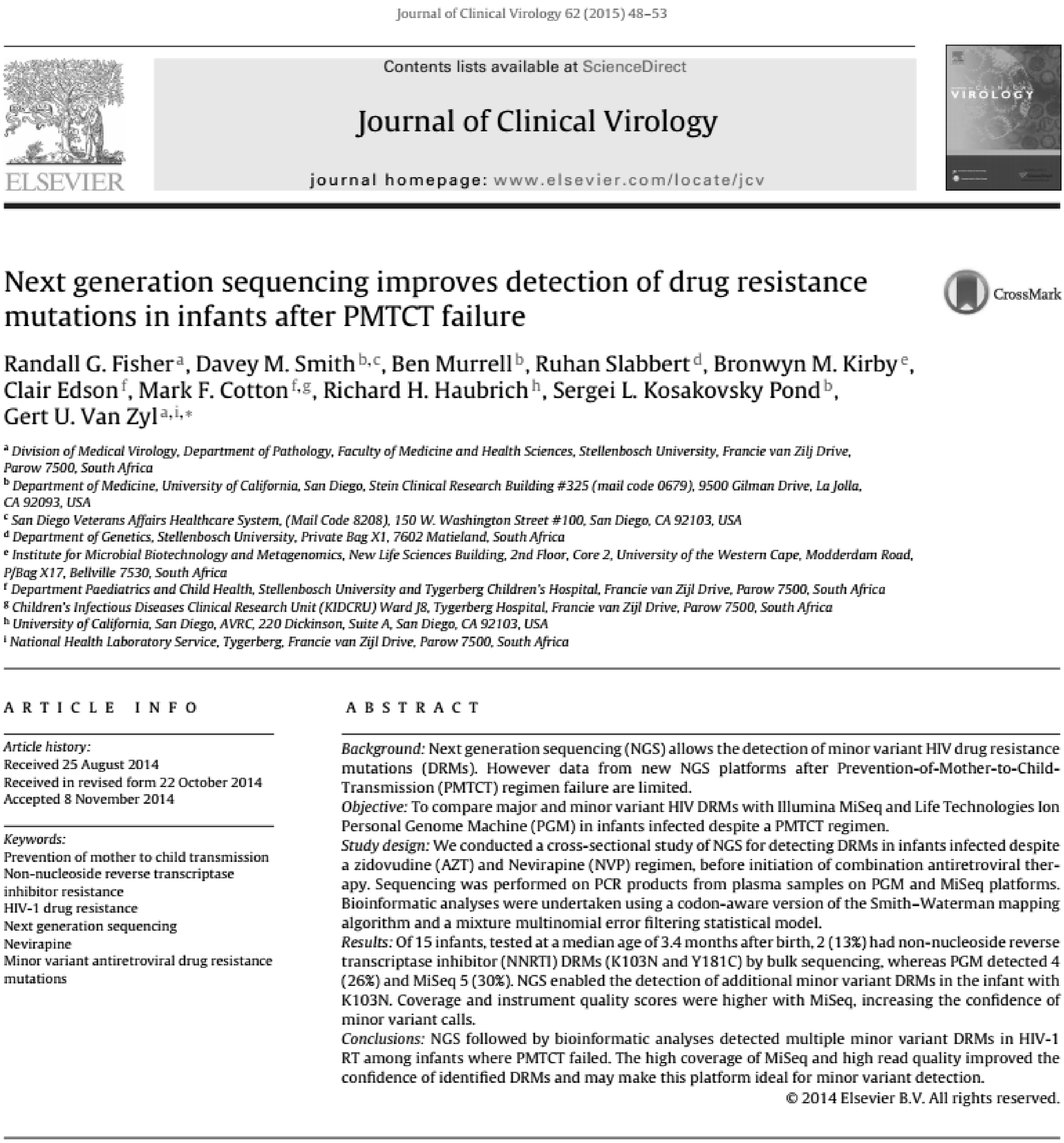


Figure 3.7.3 Concatenated nested PCR gels: The above figure depicts the results of three joined electrophoresis gel photographs with the excluded samples removed. This figure shows successful amplification of the 1 Kbp nested PCR amplicon from samples A302 and A364. Sample A302 showed inefficient amplification in three of its seven reactions while A364 was positive in all seven nested PCRs.

3.8 Published article



Journal of Clinical Virology 62 (2015) 48–53

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Next generation sequencing improves detection of drug resistance mutations in infants after PMTCT failure

Randall G. Fisher[a], Davey M. Smith[b,c], Ben Murrell[b], Ruhan Slabbert[d], Bronwyn M. Kirby[e], Clair Edson[f], Mark F. Cotton[f,g], Richard H. Haubrich[h], Sergei L. Kosakovsky Pond[b], Gert U. Van Zyl[a,i,*]

[a] *Division of Medical Virology, Department of Pathology, Faculty of Medicine and Health Sciences, Stellenbosch University, Francie van Zilj Drive, Parow 7500, South Africa*
[b] *Department of Medicine, University of California, San Diego, Stein Clinical Research Building #325 (mail code 0679), 9500 Gilman Drive, La Jolla, CA 92093, USA*
[c] *San Diego Veterans Affairs Healthcare System, (Mail Code 8208), 150 W. Washington Street #100, San Diego, CA 92103, USA*
[d] *Department of Genetics, Stellenbosch University, Private Bag X1, 7602 Matieland, South Africa*
[e] *Institute for Microbial Biotechnology and Metagenomics, New Life Sciences Building, 2nd Floor, Core 2, University of the Western Cape, Modderdam Road, P/Bag X17, Bellville 7530, South Africa*
[f] *Department Paediatrics and Child Health, Stellenbosch University and Tygerberg Children's Hospital, Francie van Zijl Drive, Parow 7500, South Africa*
[g] *Children's Infectious Diseases Clinical Research Unit (KIDCRU) Ward J8, Tygerberg Hospital, Francie van Zijl Drive, Parow 7500, South Africa*
[h] *University of California, San Diego, AVRC, 220 Dickinson, Suite A, San Diego, CA 92103, USA*
[i] *National Health Laboratory Service, Tygerberg, Francie van Zijl Drive, Parow 7500, South Africa*

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form 22 October 2014
Accepted 8 November 2014

Keywords:
Prevention of mother to child transmission
Non-nucleoside reverse transcriptase inhibitor resistance
HIV-1 drug resistance
Next generation sequencing
Nevirapine
Minor variant antiretroviral drug resistance mutations

ABSTRACT

Background: Next generation sequencing (NGS) allows the detection of minor variant HIV drug resistance mutations (DRMs). However data from new NGS platforms after Prevention-of-Mother-to-Child-Transmission (PMTCT) regimen failure are limited.
Objective: To compare major and minor variant HIV DRMs with Illumina MiSeq and Life Technologies Ion Personal Genome Machine (PGM) in infants infected despite a PMTCT regimen.
Study design: We conducted a cross-sectional study of NGS for detecting DRMs in infants infected despite a zidovudine (AZT) and Nevirapine (NVP) regimen, before initiation of combination antiretroviral therapy. Sequencing was performed on PCR products from plasma samples on PGM and MiSeq platforms. Bioinformatic analyses were undertaken using a codon-aware version of the Smith–Waterman mapping algorithm and a mixture multinomial error filtering statistical model.
Results: Of 15 infants, tested at a median age of 3.4 months after birth, 2 (13%) had non-nucleoside reverse transcriptase inhibitor (NNRTI) DRMs (K103N and Y181C) by bulk sequencing, whereas PGM detected 4 (26%) and MiSeq 5 (30%). NGS enabled the detection of additional minor variant DRMs in the infant with K103N. Coverage and instrument quality scores were higher with MiSeq, increasing the confidence of minor variant calls.
Conclusions: NGS followed by bioinformatic analyses detected multiple minor variant DRMs in HIV-1 RT among infants where PMTCT failed. The high coverage of MiSeq and high read quality improved the confidence of identified DRMs and may make this platform ideal for minor variant detection.

The full article is available online at:

http://www.sciencedirect.com/science/article/pii/S1386653214004223/pdfft?md5=052f63d83191342d7dca344443efd6b8&pid=1-s2.0-S1386653214004223-main.pdf

4 PCR sampling error associated with the use of fusion primers ("founder effect" resampling error)

4.1 Background

4.1.1 Founder effect on population diversity

The founder effect described in population genetics is an evolutionary bottleneck that results in diminished population diversity. A typical example is when a group of migrants who are not a representative of their origin population, travel to a new geographical area where they settle and multiply (Wright 1942). The genetic diversity in the subsequent offspring population is limited compared to the population of origin. The same principle applies to molecular biology when a relative small fraction of a highly diverse sample is enriched through exponential amplification (e.g. PCR) and characterised in high resolution by next generation sequencing (NGS). This phenomenon is better known as PCR resampling error (Liu et al. 1996) and consists of biased priming of particular species and random sampling error of low frequency events (minor variants). Figure 4.1.1.1 below demonstrates the mechanics of the two PCR-enrichment associated errors when compared to representative sampling.

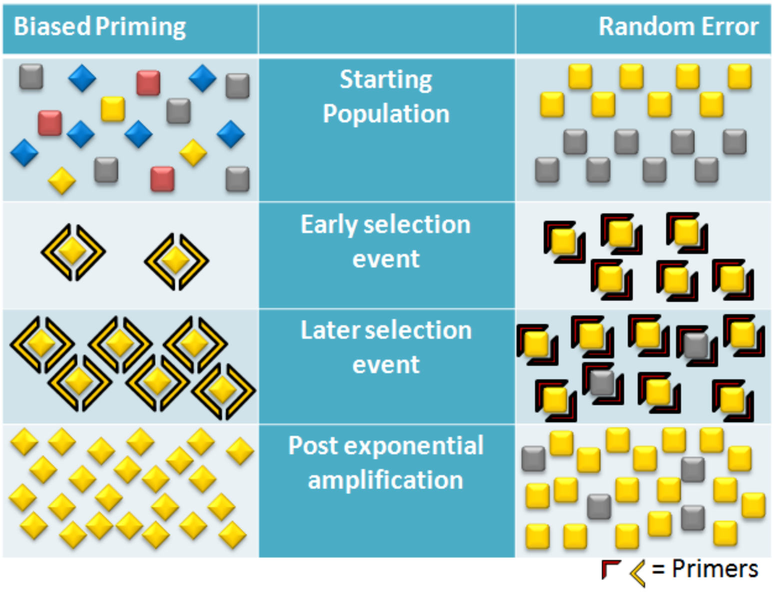


Figure 4.1.1.1 Effect of bias on population diversity: In column one of the figure above, the starting population is diverse however, amplification with stringent primers diminishes the sampled diversity, post-enrichment. In column three, the starting population contains an equal distribution of two variants. The variant randomly selected during the first few enrichment cycles (yellow), is eventually over-represented in the post-amplification population.

4.1.2 Biased priming and PCR amplification

A fundamental principle of PCR is the ability of the designed primers to specifically amplify a particular target sequence. Ideally, once the forward and reverse binding sites are primed, exponential amplification of the DNA sequence flanked by the primers occurs, resulting in millions of copies of the desired target after 35 PCR cycles. Since amplification is exponential, any early preference for a particular template would result in bias: In case of a good match between the primer and template a high proportion of primers would bind during the annealing phase and a high proportion of templates would be extended. In cases of mismatches a lower proportion of templates would be primed either due to a lower primer-template annealing temperature or due to inefficient extension in cases of mismatches in bases close to the primer 3’ end. Any such inefficient priming would result in poor efficiency of amplification during the first few cycles. Due to the nature of exponential amplification, any template-related difference in efficiency during early PCR cycles would result in a difference in template yield at the end of the PCR reaction.

4.1.3 Random sampling error

A second PCR enrichment-associated error resulting in misrepresentation of the original sample population is random. Primers randomly prime particular templates earlier than others and any random difference in time to initial priming and extension is amplified by exponential amplification: the so-called “primer resampling error” (Liu et al. 1996).

4.1.4 Primer error and random error misrepresent the original population

The unequal representation of particular templates is either due to biased priming or random resampling error is conceptually comparable to the “founder effect” described above. Consequentially, the proportion of minor variants at the end of a PCR reaction, as determined by deep sequencing, would not accurately represent the template distribution in the original sample. In the previous chapter (chapter 3), primer induced selection bias is partially avoided by using primers binding in conserved genomic areas, however, this cannot obviate bias in highly diverse viruses such as HIV.

4.1.5 Fusion primers and the founder effect

A particular scenario where this “founder effect” may be more pronounced is in the case of fusion primers being used to generate amplicons for NGS. Genome Sequencer (GS) Junior (454 Life Sciences, Connecticut, USA) amplicon sequencing requires the use of a sequencing adaptor, a sample-specific molecular identifier (MID) and a template-specific region. Incorporation of sequencing adaptors and MIDs in the primers obviates the need to ligate them after PCR, simplifies workflows and reduces costs. Amplicon sequencing using fusion primers is well-established for HIV-1 drug resistance genotyping (Garcia-Diaz et al. 2013; Mohamed et al. 2014; Simen et al. 2014; Fisher et al. 2012). However, the primer requirements result in abnormally long primers with high melting temperatures (Tms).

During the first round of PCR amplification only the 3’ template-specific fusion primer domains bind to the template. After the first PCR cycle, fusion primers harbouring the 5’ MID and sequencing adaptor would have been incorporated into the template. Due to the higher Tm of these PCR-generated templates, at temperatures close to the Tm of the specific primer, PCR derived templates from the first few rounds of amplification would be preferentially amplified compared to templates for which amplification initiates in subsequent cycles. Resampling is practically relevant as, due to the risk of primer dimer formation, high temperature cycling conditions are used for fusion primer PCR annealing. Any primer-template mismatches that delay amplification would therefore be more pronounced with fusion primers, resulting in a pronounced unequal enrichment or a “founder effect”.

4.1.6 Identifying original templates with random tags before PCR enrichment

Recently, two cDNA random primer ID approaches have enabled bioinformaticists to identify each sampled RNA species, compensating for PCR resampling. Both strategies reverse transcribe RNA with a gene-specific cDNA primer that contains a 5’ PCR primer-binding site, followed by eight to ten random bases that are used to identify individual cDNA species. This primer is seen in figure 4.1.6.1 below along with the enrichment strategy. In the first strategy, an eight base random ID tag is used to identify a particular cDNA species. Then, PCR is used to add the sequence adaptors and molecular identifiers (MIDs) to the amplicon by priming the 3’ PCR primer-binding site and an upstream gene-specific region (Jabara et al. 2011). The second method uses a 10 random-base ID tag and 22-mer pre-nested PCR primers with 5’ uracils. These uracils are later digested to produce sticky-ends to which single-stranded sequence adaptors and MIDs are ligated. The single-stranded adaptor is complemented by a polymerase, forming a functional amplicon (Boltz et al. 2015). Once amplicons have been sequenced, sequences are condensed based on their individual IDs and a consensus is generated for each sampled RNA species.

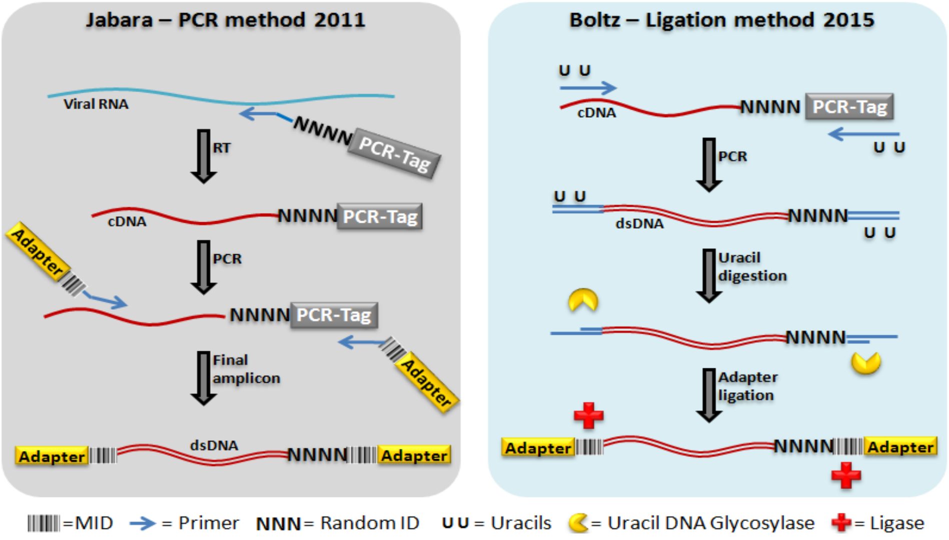


Figure 4.1.6.1 Random base ID tag: In the Jabara method (left), cDNA synthesis uses a primer consisting of a gene-specific region, an eight random-base ID and a PCR binding tag (3' to 5' direction). Each cDNA species has a unique tag. This primer is incorporated into the cDNA. The tagged cDNA is PCR amplified with primers complementary to the 3' PCR tag and a 5' genomic region. These PCR primers contain 5' patient MIDs and the MiSeq adaptors. For the Boltz method (right), cDNA is generated with a gene-specific primer with a 10 random-base ID and a PCR binding tag (3' to 5'). cDNA is then amplified with 22-mer uracil-containing primers that bind to the incorporated PCR tag on the 3' end and a 5' genomic region. Uracil DNA glycosylase digests the uracil bases at the end of the incorporated primers and the single-stranded adaptors are ligated to the amplicon's sticky-ends. Finally, the adaptor compliment is synthesised by a high fidelity polymerase.

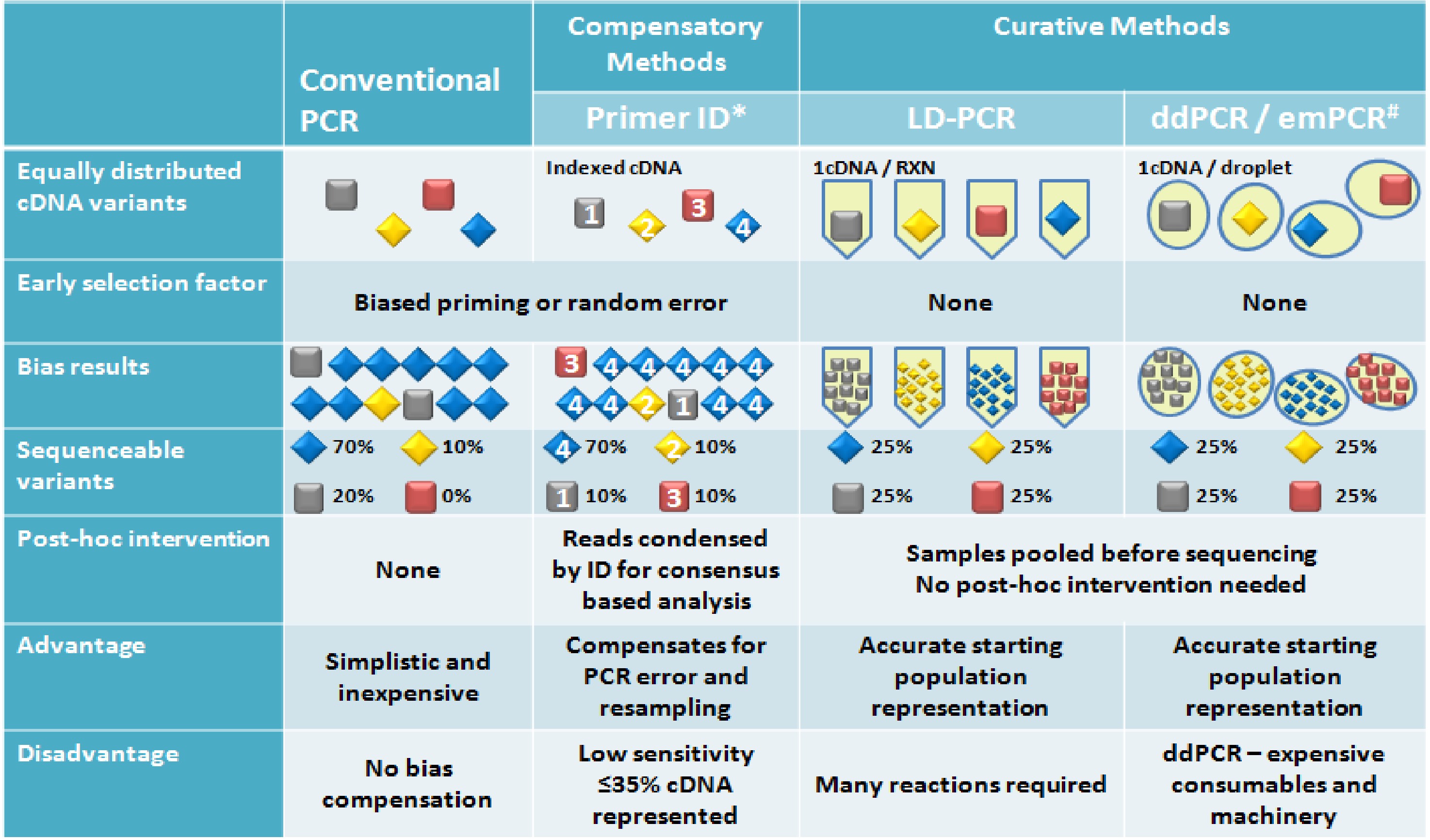
These primer ID methods allows bioinformatic correction resampling (as demonstrated in figure 4.1.7.1 below), but have proven to be insensitive and only clinically applicable to patients presenting with high viral loads. Empirically, the older methods sample between 2 and 20% (Keys et al. 2015; Zhou et al. 2015; Brodin et al. 2015) and the newer method (Boltz et al. 2015), samples roughly 35% of the available RNA species. In addition, the primer ID approach induces biased resampling and does not always correct for PCR substitutions (Brodin et al. 2015). Curative alternatives, such as limiting dilution PCR (LD-PCR) and partitioned PCR (e.g.: nano-droplet partitioning as in digital droplet PCR (ddPCR)) as opposed to the corrective random primer ID sequence approach, would allow representational enrichment of the original template.

4.1.7 Representational enrichment to overcome PCR resampling

LD-PCR has been used for single genome amplification (Salazar-Gonzalez et al. 2008) and requires a predetermined cDNA copy number and cDNA dilution to a single-genome-per-PCR resolution. Essentially, selection bias is eliminated since there is no binding competition between PCR primers and numerous cDNA templates, as illustrated in figure 4.1.7.1 below. While this methodology efficiently eliminates the effects of PCR recombination, resampling and selection biasing, it proves laborious and expensive when considering the sheer number of reactions required.

Another viable alternative to LD-PCR is ddPCR. This technology diminishes the PCR founder effect through partitioning templates in individual nanolitre reactions which diffusion constrains templates and primers (as depicted in figure 4.1.7.1 below). These conditions may increase the efficiency of PCR in the presence of template-primer-mismatches. However, the limited nano-reaction volumes result in early PCR plateau of templates that are primed first and amplified with high efficiency. Commercial ddPCR platforms offer the advantage of standardised reaction droplet volumes and are available from BioRad in the QX200 ddPCR system (California, USA) or the RainDance Technologies droplet PCR platform (Massachusetts, USA).

Both platforms generate one nanolitre PCRs however, only the RainDance Technologies platform (Massachusetts, USA) accommodates a high fidelity PCR for upstream NGS sample preparations since the QX200 is dedicated for absolute quantification. Also common to both ddPCR platforms is the expensive, obligatory equipment (i.e.: RainDance Technologies RDT1000 droplet generator or the BioRad QX200 droplet generator) and the exorbitant cost of consumables, such as the disposable microfluidics cartridges. These factors limit the accessibility of the platforms in resource-limited settings.



	Conventional PCR	Compensatory Methods	Curative Methods	
		Primer ID*	LD-PCR	ddPCR / emPCR#
Equally distributed cDNA variants	⬜ 🔶 🟥 🔷	Indexed cDNA 1 2 3 4	1cDNA / RXN ⬜ 🔶 🟥 🔷	1cDNA / droplet ⬜ 🔶 🔷 🟥
Early selection factor	Biased priming or random error		None	None
Bias results	⬜ 🔷 🔷 🔷 🔷 🔷 🔷 🔷 🔶 ⬜ 🔷 🔷	3 4 4 4 4 4 4 4 2 1 4 4	⬜ 🔶 🔷 🟥	⬜ 🔶 🔷 🟥
Sequenceable variants	🔷 70% 🔶 10% ⬜ 20% 🟥 0%	4 70% 2 10% 1 10% 3 10%	🔷 25% 🔶 25% ⬜ 25% 🟥 25%	🔷 25% 🔶 25% ⬜ 25% 🟥 25%
Post-hoc intervention	None	Reads condensed by ID for consensus based analysis	Samples pooled before sequencing No post-hoc intervention needed	
Advantage	Simplistic and inexpensive	Compensates for PCR error and resampling	Accurate starting population representation	Accurate starting population representation
Disadvantage	No bias compensation	Low sensitivity ≤35% cDNA represented	Many reactions required	ddPCR – expensive consumables and machinery

* cDNA is indexed with a random ID tag during reverse transcription. # em-PCR (emulsion PCR) is achievable with basic laboratory equipment

Figure 4.1.7.1 Compensatory and curative approaches: The figure above compares conventional PCR to the compensatory and curative methods for obviating PCR sampling error. The random primer ID method, currently the only compensatory method, does not correct the sampling bias but post-hoc allows for sequences derived from the same template to be condensed into one consensus sequence. By generating a consensus, any PCR and sequencing errors can be corrected, based on the nucleotide prevalence at a given locus. Alternatively, a curative approach to enrichment error is nano-partitioning in minute individual reactions, each containing one template only.

4.1.8 Water-in-oil emulsion PCR (em-PCR)

Partitioning PCRs into “nano-reactors” obviates not only biased priming and random sampling error but also PCR recombination. An inexpensive and robust means of generating minuscule, diffusion constrained reactions is by PCR emulsification in oil. Various methods of PCR emulsification have previously been published (Shao et al. 2011; Nakano et al. 2003; Hori et al. 2007), the easiest of which was described by Schütze and colleagues in 2011 (Schutze et al. 2011). Briefly, a 50 µl PCR is emulsified in three volumes of a pre-chilled oil and surfactant mixture, by vortexing the aqueous and oil faction for three minutes at 10 °C (achieved by placing an inexpensive Vortex-genie II (Scientific Industries, New York, USA) in a cold room). After enrichment, the emulsion was broken with an organic solvent and the DNA, recovered.

Similar to the nano-reaction depletion phenomenon seen in ddPCR, the fento or atto-litre reactions generated by this emulsion PCR are also subject to depletion, possibly sooner than the larger nano-reactions. Remedially, Nakano and colleagues proposed emulsion dissolution to replenish template-containing droplets with empty droplets (Nakano et al. 2003). Presumably, the initial emulsified PCR cycles are sufficient to mutate the mismatched templates (selected and amplified as a consequence of partitioning) into templates that are perfectly matched to the same species of fusion primer, in subsequent PCR cycles.

4.1.9 Study rationale

While fusion primers remain a cost effective means of performing 454 NGS targeted resequencing, we set out to measure the founder effect induced when using these unusually long primers. We also applied the described em-PCR methodology to a high fidelity PCR, primed with the 454 fusion primers that are known to be highly selective. The em-PCR would allow us to “handicap” the best-matched variants (by being volume constrained and reaching PCR plateau early), while permitting the mismatched variants to “catch-up”. Once the template has been successfully PCR-mutated (probably after seven cycles), we replenish the droplets with negative droplets for efficient processivity.

In short, we synthesised three plasmids with patient-derived HIV *reverse transcriptase* sequences, containing variable primer binding targets sequences. The plasmids were characterised by Sanger sequencing and each plasmid was quantified with a ddPCR. To assess the effect of early selection events on the relative proportion after PCR enrichment, three mixtures (of the three plasmid species) were made and consisted of 1%; 9% and 90% of each plasmid respectively. These mixtures were then PCR-amplified in triplicate, using three fusion primer sets (each including a different MID) for each mixture. Thus, a total of nine primer sets with different MIDs was used. For each plasmid mixture, the same set of three fusion primers was used for both em-PCR and bulk or open-PCR (o-PCR), resulting in a total of 18 reactions (nine em-PCRs and nine o-PCRs). The em-PCR and o-PCR products were then size selected and purified before being deep-sequenced in separate runs on the GS Junior platform (454 Life Sciences, Connecticut, USA). Post-sequencing, homopolymer sequencing error was corrected using a codon-aware sequence mapping software.

4.2 Methods

The sampling strategy used in our study was to amplify three known plasmid mixtures, each with three fusion primer sets, through both open (o-PCR) and emulsion PCR (em-PCR). To achieve this goal, specific objectives had to be met. These objectives were as follows:

- Synthesise, characterise and accurately quantify the three plasmid species that would form our three mixed populations.
- Design and test GS Junior fusion primers through both o-PCR and em-PCR.
- Optimize a high fidelity em-PCR to reduce the initial sampling bias and produce sufficient product to be sequenced.
- Perform post-sequencing bioinformatic analysis to map all GS Junior reads to one of three plasmid-derived “reference” sequences.

The methods that follow describe the plasmid synthesis, sequencing, quantification and the plasmid pooling strategy to form the three sample populations. Also described below are the fusion primer design, the o-PCR and em-PCR optimisation, the amplicon size selection and purification and the post-sequencing data analysis.

4.2.1 Plasmids

4.2.1.1 Ligation

Amplicons sequenced in this study were derived from a mixture of three recombinant plasmid species. Each plasmid construct consisted of the pGEM®-T Easy Vector “back-bone” (Promega, Wisconsin, USA) and a ~930 bp, patient-derived HIV-1 subtype C, *reverse transcriptase* insert. These inserts were generated as part of another study and the associated methods can be seen in the previous chapter (chapter 3). Once generated, amplicons were separated in a 2% agarose gel before being excised and gel extracted. 150 to 250 ng of these gel purified inserts were ligated into 100 ng of the pGEM®-T Easy Vector (Promega, Wisconsin, USA) using in-house optimised reaction mixtures and conditions seen below. The ligation reaction was validated using the relevant reaction controls that included a ligation positive control insert and a background/no-insert control. Each reaction was prepared individually (i.e.: no mastermix was prepared) in a 200 µl thin-walled PCR tube and all thermal incubations were carried out in ABI 9700 thermocycler (Applied Biosystems, Massachusetts, USA), with the heat-lid activated at 105 °C.

	Reagent	Final concentration (in 20 µl)
1.	2 x Rapid ligation reaction buffer	1 x
2.	T4 DNA Ligase (3 U/µl) (Promega)	3 U/rxn
3.	Control DNA insert (4ng) or DNA amplicon	Various
4.	10 µl Nuclease free water (N.F.H_2O)	1 x
5.	pGEM T-Easy vector (50 ng/µl)	100 ng

Ligation reactions were then incubated at 25 °C for 2 hrs, followed by 4 °C for 16 hrs, 75 °C for 5 min and an indefinite hold at 4 °C. On completion, recombinants were immediately transformed using an in-house optimised transformation protocol shown below.

4.2.1.2 Culturing recombinant plasmids

The recombinant plasmids were transformed using the following protocol, along with a transformation efficiency control plasmid:

- Pre-chill 1.5 ml tubes and SOC media while thawing 250 µl JM109 chemically competent bacteria (Promega, Wisconsin, USA) on ice.
- Add 5 µl of the ligation constructs to the pre-chilled tubes.
- Add 50 µl JM109 (Promega, Wisconsin, USA) cells to each transformation tube. Mix once by slow up and down pipetting.
- Incubate tubes on ice for 1 min and heat shock at exactly 42 °C for 50 sec.
- Incubate reactions on ice for 2 min before adding 950 µl cold SOC media.

The reaction was then incubated at 37 °C for 1.5 to 2 hrs with gentle shaking before being spread on semisolid ampicillin-selective LB media culture plates, for IPTG-X-Gal mediated blue/white colony selection. Single, white colonies were inoculated into 4 ml of sterile LB liquid media with 100 µg/ml ampicillin. After an overnight shaking incubation at 37 °C, plasmids were isolate using Promega's PureYield™ Plasmid Miniprep system (Wisconsin, USA) and eluted in warm (50 °C) TE buffer.

4.2.1.3 Quantification and genotyping recombinants

The dsDNA BR assay kit (Life Technologies, California, USA) was used according to the manufacturers' specifications to quantify the plasmid species in triplicate, on the Qubit Fluorometer 2.0 (Life Technologies, California, USA). The Qubit (Life Technologies, California, USA) measures the florescence emanating from double-stranded DNA diluted in the assay buffer, which contains an intercalating fluorescent DNA dye. An intercalating dye assay allows for DNA quantification without quantifying background, non-specific, co-purified agents (e.g.: salts and proteins). When using spectrophotometric quantification, the spectral overlap of these co-purified agents cause false, over-quantification. The average of the three measured concentrations was used to add 150 to 250 ng of the purified pGEM$^{\circledR}$ constructs to an in-house optimised BigDye$^{\circledR}$ Direct Cycle (V3.1) sequencing reaction (Applied Biosystems, Massachusetts, USA), primed with standard M13 forward or reverse oligos.

	Reagent	Final concentration (in 10 µl)
1.	5 x Terminator sequencing reaction buffer	1.5 x
2.	BigDye$^{\circledR}$ Terminator reaction mix	1 x
3.	M13 Fwd or Rev Primer	5 pMoles
4.	pGEM DNA construct	150 to 250 ng

Sequencing PCRs were performed using the following in-house-optimised reaction conditions in a 9700 thermocycler:

95 °C for 10 sec
25 x { 95 °C for 5 sec
72 °C for 4 min }
4 °C soak

On completion, the sequencing PCR reaction was purified using an in-house-optimised BigDye$^{\circledR}$ X-terminator clean-up (Applied Biosystems, Massachusetts, USA), seen below. The SAM solution™ (Applied Biosystems, Massachusetts, USA) used in the purification reaction was heated to room temperature before use.

	Reagent	Volume added to Mastermix
1.	SAM™ solution	49 µl
2.	X-terminator solution	11 µl

Fifty-five microlitres of the above reaction mixture was added to each sequencing PCR in a 96 well plate, before the wells were sealed and the reaction, vortexed for 30 to 45 min. The purified sequencing PCR product was then centrifuged at 1 000 xg for 2 min to pellet the X-terminator nano-particles and the sequencing reaction was read on an ABI Genetic Analyser 3130xL (Applied Biosystems, Massachusetts, USA). Chromatograms from the sequencing reactions were imported into Geneious (V6.0) (Biomatters, Auckland, New Zealand) and blasted against the National Centre for Bioinformatics Information (NCBI) nucleotide database to determine specificity (available online at: http://blast.ncbi.nlm.nih.gov/Blast.cgi). Three plasmid species were selected for further use (i.e.: P2.38, P3.32 and P4.38) based on their diversity between regions 2 559 to 2 931 (HXB2). The

3 015 bp plasmid copy numbers were estimated using a molecular weight calculator (available online at http://www.bioinformatics.org/sms2/dna_mw.html) and a copy number calculator (available online at http://cels.uri.edu/gsc/cndna.html). Plasmid stocks were then diluted to approximately 1 x 10^3 and 1 x 10^2 c/µl before absolute quantification using BioRad's QX200 (California, USA) digital droplet PCR system.

4.2.1.4 Absolute quantification of recombinant plasmids

At this stage in our study, we become aware of the Qubit's (Life Technologies, California, USA) tendency to under-quantify DNA concentration due to differential intercalation of the assay dye. As we were already aware of the NanoDrop's (Thermo Scientific, Massachusetts, USA) inclination to over quantification due to contaminant spectral overlap, we decided to quantified the plasmids using BioRad's QX200 (California, USA) digital droplet real time PCR (ddPCR) system and the primers listed in table 4.2.1.4.1 below. Digital droplet PCR allows for accurate quantification of DNA without the necessity of DNA standards.

Table 4.2.1.4.1 ddPCR absolute quantification primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward primer	F3_F_RealT	GGGCAACATAGAGCAAAAATAGARGA	3135 to 3160
Reverse primer	F3_R_RealT	GGAGTTCATACCCCATCCAAAGAAATG	3226 to 3252 (RC)

(RC) = Reverse complement

The three selected plasmid species were quantified using BioRad's EvaGreen Digital PCR SuperMix (California, USA) according to the manufacturer's specifications as seen below.

	Reagent	Final concentration (in 25 µl)
1.	2 x Bio-Rad EvaGreen SuperMix	1 x
2.	F3_F_RealT FWD Primer	100 nM
3.	F3_R_RealT REV Primer	100 nM
4.	5 µl plasmid DNA (P2.38, P3.32 or P4.38)	Supposedly 50 and 500 copies
5.	3 µl N.F.H_2O	1x

The mastermix of each plasmids species was made for four reactions, of which only three, technical replicates were quantified. A no-template control (NTC) was included with nuclease free water (N.F.H_2O) supplementing the 20 µl reaction volume. Of each reaction mastermix, 20 µl was pipetted into the sample wells of the microfluidics DG8 droplet generator cartridge (figure 4.2.1.4.1). Seventy microliters of the EvaGreen droplet generator oil was added to each of the oil reservoirs in the DG8 cartridge (figure 4.2.1.4.2) and the cartridge was loaded into the holder (figure 4.2.1.4.3). Once the droplet generator gasket was placed on the cartridge holder, the assembly was loaded into the droplet generator (figure 4.2.1.4.4). (The images shown below are available online at http://www.bio-rad.com/en-za/category/digital-pcr)

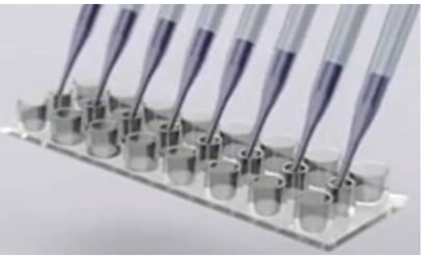
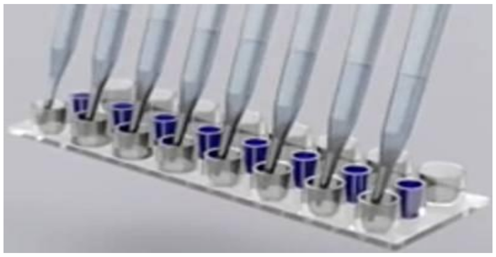


Figure 4.2.1.4.1 Sample loading: 20 µl sample mastermix is loaded into sample wells of the DG8 cartridge using a multichannel pipette. Extra care is taken to prevent air bubbles from forming at the bottom of the loading well as this would hinder the embedded microfluidics from forming the emulsion droplets.

Figure 4.2.1.4.2 Oil loading (right): 70 µl emulsion oil is loaded into the oil reservoir. Once again, the operator must be cautious to prevent air bubbles from collecting at the bottom of the wells.



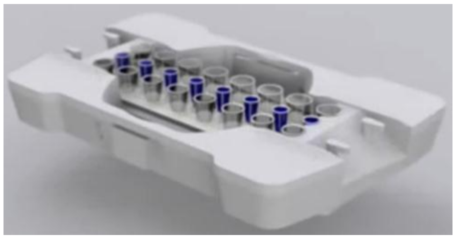
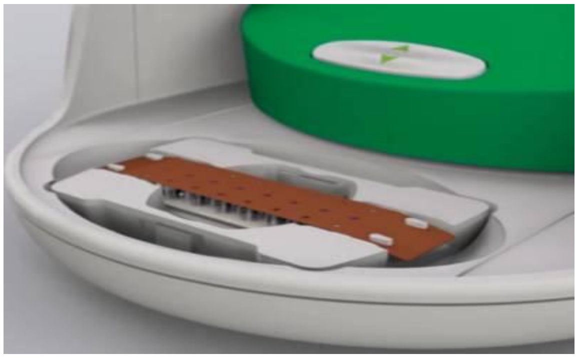


Figure 4.2.1.4.3 Cartridge holder assembly (left): The DG8 cartridge is loaded into the holder which clips closed to secure the cartridge in place before the rubber gasket is hooked over the edges of the cartridge holder and the assembly is loaded into the droplet generator (as seen in figure 4 below).

Figure 4.2.1.4.4 Assembly loading (right): The assembly is loaded into the droplet generator and once the assembly is correctly in place, the green "sample loaded" light is engaged. The droplet generator is closed and the microfluidics are engaged.



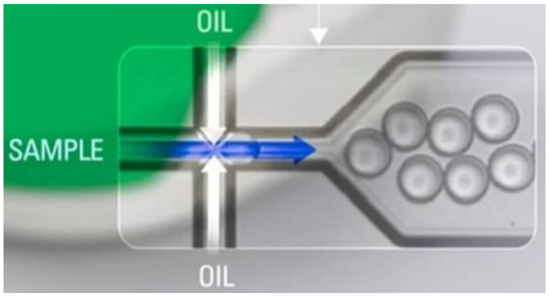


Figure 4.2.1.4.5 Droplet generation (left): By pressurising the air in the oil and sample wells, the droplet generator creates 20 000 one nanolitre PCR reaction, as depicted in figure 5.

The droplet generator then uses the microfluidics embedded in the DG8 cartridge to generate ~20 000 droplet reactions (figure 4.2.1.4.5) which are then carefully loaded into an Eppendorf (Hamburg, Germany), twin-tech, semi-skirted, 96 well plate. Once heat sealed with pierceable foil, these samples are amplified through 40 cycles of PCR and a final signal stabilisation step of 90 °C for 5 min. Thermocycling conditions were as follows:

95 °C for 5 min

40 x { 95 °C for 30 sec
60 °C for 30 sec }

4 °C for 5 min

90 °C for 5 min (signal stabilisation)

4 °C soak

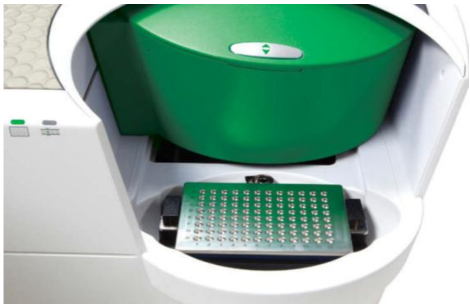


Figure 4.2.1.4.6 Sample reading: After amplification, the plate is loaded into the droplet reader and the probe-guide cover-plate is secured in place.

On completion, the twin-tech plate was loaded into the droplet reader machine, making sure to clamp the plate down correctly with the plate holder lid (figure 4.2.1.4.6 above). Data acquisition on the droplet reader (BioRad, California, USA) requires that the wells are defined with respect to the expected florescence (i.e.: EvaGreen, VIC or FAM probe based detection) in the graphical user interface software, QuantaSoft (BioRad, California, USA). The software also allows the user to define the contents of the experiment in each well.

The droplet reader reads as many droplets as possible in a similar manner to a flowcytometer. Positive discrimination is based on the difference in the amount of observed florescence between droplets with or without an amplified PCR target. As is the case with EvaGreen reactions, “negative droplet” exhibit baseline florescence, comparable to that observed in no-template controls (NTCs), while “positive droplets” display an approximate three to four-fold increased florescence.

When the samples are analysed or “read” on the droplet reader, two distinct populations are visible, provided that the reaction was not over saturated with amplifiable template (i.e.: all droplets would be positive). In the resultant 1 dimensional data plot, positive droplets display high florescence amplitude readings and negative droplets show base-line florescent. QuantaSoft (BioRad, California, USA) analysis software employs Poisson probability statistics to estimate the number of target copies in the starting 20µl reaction, using the ratio of positive to negative droplets and the total number of droplets sampled.

The results of the absolute quantification experiments and the employed dilution factors were used to calculate and correct the concentrations of the three plasmid species stocks. These three plasmid species were reconstituted at a corrected 2x 10^4 c/µl in elution buffer before being pooled into three plasmid mixtures each consisting of different proportions of the individual plasmids.

4.2.1.5 Plasmid pooling

Employing the absolute quantification results, the three plasmid species stocks were diluted appropriately and combined in the mixtures tabulated in the results section below (table 4.3.6.1). These plasmid mixtures were constituted at 10-fold the required concentration and formed the synthetic mixture which would be PCR amplified using the fusion primers which include molecular identifiers (MIDs) and GS Junior sequencing adaptors.

4.2.2 Fusion primers

GS fusion primers were based on the previously published Varghese et al. (2010) primer sets, to amplify a 417 bp region of HIV-1 (Varghese et al. 2010). These Varghese primers were subtype C-specific with sequencing adaptors and MIDs for the FLX Titanium 454 platform (454 Life Sciences, Connecticut, USA), while ours contained adaptors and MIDs for the GS Junior sequencing platform (454 Life Sciences, Connecticut, USA), and the target-specific regions were chosen based on the South African subtype C consensus sequence. The fusion primer gene-specific regions were a perfect match for plasmid 2.38; had one forward and one reverse mismatch to plasmid 3.32; and three forward mismatches to plasmid 4.38 (the most fatal of which was in position -3). The primer specifics can be seen in table 4.2.2.1 below, while the primer schematics are listed in section 4.3.8 of the results section.

Table 4.2.2.1 Modified GS Junior fusion primer specifics: Underlined text notates a primer to template mismatch and **bolded** text indicates a lethal mismatch. The * notates an annealing temperature calculated using the neighbour joining method available online at https://www.exiqon.com/ls/Pages/ExiqonTMPredictionTool.aspx.

MID to plasmid mix matching		**Primer to template matching**		
		Origin	**FWD Sequence (Tm* 56-58°C)**	**REV Sequence (Tm* 59°C)**
MIDs	**Plasmid mix**	Primer	TGCACAYTAAATTTTCCAATTAG	ACTAGGTATGGTGAATGCAG
13, 14, 60	90%P2, 9%P3, 1%P4 (MIX1)	Plasmid 2	TGCACAYTAAATTTTCCAATTAG	ACTAGGTATGGTGAATGCAS
5, 7, 20	90%P3, 9%P4, 1%P2 (MIX2)	Plasmid 3	TGCACAYTGAATTTTCCAATTAG	ACTAGGTATGCTGAATGCAG
10, 17, 24	90%P4, 9%P2, 1%P3 (MIX3)	Plasmid 4	TGCACAYTGAATTTCCCAAT**A**AG	ACTAGGTATGGTGAATGCAG

4.2.3 Fusion primer PCR optimisation

The fusion primer PCR was initially tested in the presence of Epicentre's MasterAmp PCR enhancer containing Betaine (Wisconsin, USA) which reduced the occurrence of primer dimers, when compared to the native high fidelity PCR, with and without added dimethyl sulfoxide (DMSO). The assay was further optimised with respect to primer and magnesium chloride ($MgCl_2$) concentration, and finally, annealing temperature. These steps dictated an optimised reaction using 1x MasterAmp PCR enhancer (Epicentre, Wisconsin, USA), a final $MgCl_2$ and fusion primer concentration of 2.25 mM and

100 nM respectively, and an annealing temperature of 55 °C. The following gel figures were used as end-point evaluation of the efficacy of the optimisation steps implemented. As described in the methods, the fusion primer PCR was optimised with regards to PCR additives, $MgCl_2$ and primer concentration and annealing temperature.

4.2.3.1 PCR additives

The three primer species were chosen as optimisation candidates as they represent the three fusion primer PCR outcomes observed in the standard reaction:

- MID 3 fusion primers produced a specific amplicon without any dimer.
- MID 13 efficiently produced both a specific product and a primer dimer.
- MID 20 efficiently produced a sequence amplicon but less efficiently produces a primer dimer.

Figure 4.2.3.1.1 below depicts the effect of various PCR additives on the fusion-primer-initiated amplicon production.

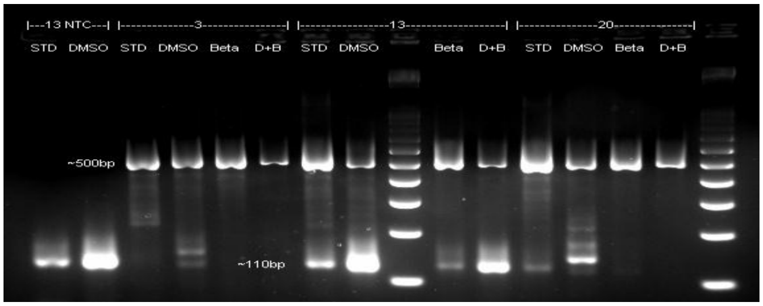


Figure 4.2.3.1.1 PCR additives DMSO and betaine: In the gel figure above, the first row shows MID fusion primer set used in the reaction while the second, the reaction additives: STD notates the native reaction; DMSO notates a reaction with 1 mM DMSO (final conc.); Beta labels a reaction with added MasterAmp betaine-containing PCR enhancer; and D+B labels reactions with both MasterAmp and DMSO. The addition of DMSO biased the reaction towards primer dimerization, evident by the brighter bands visible at ~120 bp, when compared to the standard reaction.

Conversely, the addition of the Betaine-containing PCR enhancer biased the reactions towards the sequence amplicon in reactions with both dimers and amplicons (MID 20 and MID 3) and improved amplicon production in reactions without dimer (MID 3). The addition of both the PCR enhancer and DMSO reduced the occurrences of primer dimers at the expense of specific amplicon production. These results warranted the use of the PCR enhancer without the addition of DMSO.

4.2.3.2 $MgCl_2$ concentration optimisation

MID 13 was selected for $MgCl_2$ optimisation as it gave the best indication as to the reactions tendency towards dimerization. Figure 4.2.3.2.1 below displays the results of the $MgCl_2$ titration reactions. While the higher concentrations of $MgCl_2$ resulted in efficient amplicon production (lanes eight to 10), it also favoured the production of non-specific products, such as primer dimers, seen at aproximately 120 bp. For this purpose, a final $MgCl_2$ concentration of 2.25 mM was employed dispite the possible bias towards non-specific products, which we would exclude through amplicon size selection.

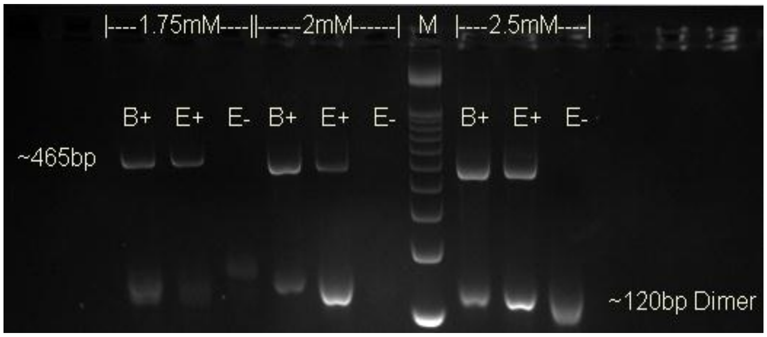


Figure 4.2.3.2.1 $MgCl_2$ optimisation with MID 13: $MgCl_2$ was titrated at 1.75 mM (lanes one to three), 2.0 mM (lanes four to six) and 2.5 mM (lanes eight to 10) in both o-PCR and em-PCR as depicted above. At the time of optimisation, we refered to open PCRs as "bulk PCR" and labled these reactions as "B+" while "E+"and "E-" notated positive and NTC emulsion reactions, respectively.

4.2.3.3 Fusion primer concentration optimisation

The manufacturer-specified primer concentration for the Expand High Fidelity^{PLUS} PCR system ranges from 200 to 400 nM (Roche Diagnostics, Basel, Switzerland). We attempted to reduce primer dimer formations by titrating fusion primers to lower concentrations, as seen in figure 4.2.3.3.1 below. The gel figure below depicts the results of a range of NTC reactions amplified with 200, 100 and 50 nM primer concentrations.

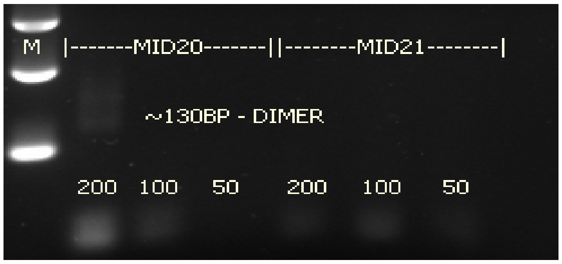


Figure 4.2.3.3.1 Fusion primer titration: MIDs 20 and 21 were used to test the effect of primer concentration on dimer formation. Since the fusion primers were ~65 bases in length, faint band visible at almost twice this size were presumed to be a dimer. In the 200 nM fusion primer concentration reaction (MID 20), a primer dimer is seen at ~130 bp (lane two from the left), but is absent in the 100 nM reaction.

For our assay, a final primer concentration of 100 nM was used since it was the highest tested concentration that did not produce a dimer in the absence of amplifiable template. While MID 21 showed no dimerization at all primer concentrations, it was not used for further experiments as it failed to amplify plasmid template in both open and emulsion PCRs. MID 20 dimerized at 200 nM but not at 100 nM and was used for further experiments.

4.2.3.4 Fusion primer annealing temperature optimisation

The nearest-neighbour Tm prediction method was used to calculate the appropriate annealing temperature for the template-specific region of the fusion primers. The optimum annealing temperature however, had to be empirically determined. Gel figure 4.2.3.4.1 below depicts a sample of the optimisation assay wherein primers were annealed at five and ten degrees below their 65 °C melting temperature however, the full range of fusion primers were tested.

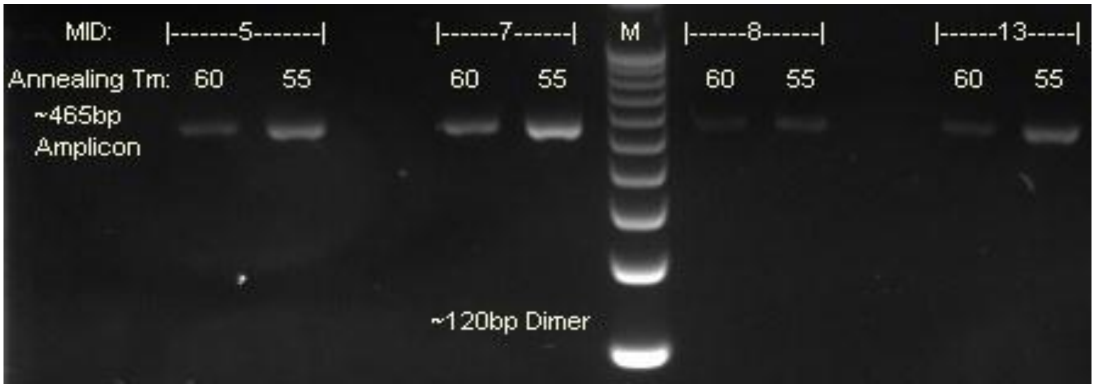


Figure 4.2.3.4.1 Fusion primer Tm optimisation: Using MIDs 5, 7, 8 and 13, the assay annealing temperature was broadly optimised at five and ten degrees below the calculated melting temperature of the fusion primer. Brighter bands are observed when using the 55 °C annealing temperature which suggests better efficiency with the less stringent annealing parameters.

The figure above demonstrated more efficient amplification using an annealing temperature 10 °C below the primer melting temperature. A lower annealing temperature is associated with less stringent primer binding, a higher primer to template mismatch tolerance and thus a lower selection bias. To this end, an annealing temperature of 55 °C was used for further experiments as it displays more efficient amplification and a better mismatch tolerance.

4.2.3.5 Additional em-PCR optimisation

In preliminary experiments using the em-PCR as described by Schütze (Schutze et al. 2011), amplification for the full reaction duration resulted in poor product yield due to reagent depletion in droplets that contain both primers and amplifiable template, as is visible in the figure 4.2.3.5.1 below.

Both o-PCRs and em-PCRs were allowed to complete 50 cycles, and on completion, the emulsion reaction was broken through centrifugation and separated in a 2% agarose gel, alongside the o-PCR. Remedially, emulsification was limited to the first seven cycles, as PCR sampling bias is probably most affected by early PCR events, before PCR mediated template mutagenesis (PCR template is predominantly primer derived after a few cycles). After these emulsified cycles, the emulsion was broken through centrifugation, to replenish substrate-depleted, template-containing droplets with unamplified, negative droplet substrates. In a study conducted by Nakano in 2003, 13 emulsified cycles were used to saturate larger emulsion droplets before the emulsion was broken to replenish PCR substrates (Nakano et al. 2003).

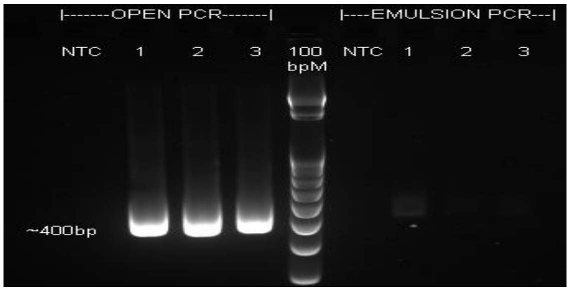


Figure 4.2.3.5.1 Inefficient emulsified amplification: The gel figure to the left displays the o-PCR NTC in lane one on the left, followed by triplicate positive reactions, displaying a 465 bp amplicon in lanes two to four. The em-PCR products can be seen to the right of the o-PCRs (after the 100 bp marker) and loaded in the same format. In lanes seven to nine, the em-PCRs display faint amplicons and poor amplification efficiency.

Our final em-PCR reaction consisted of an initial seven cycles of emulsified PCR, allowing initial partitioned amplification (to prevent competition with other templates) in diffusion constrained droplets (to enhance the efficiency of amplification). These initial cycles were expected to facilitate PCR-mediated mutagenesis of the template to match the fusion primer. After seven cycles, the large majority of PCR product would be expected to be primer derived, and after this, to counteract the reaction depletion in the minute vesicles, the emulsion was broken through centrifugation and the PCR cycling then continued for an additional 43 PCR cycles. Gel figure 4.2.3.5.2 below depicts the fully optimised fusion primer PCR executed on the three, previously-described plasmid mixtures, implementing the replenished em-PCR methods.

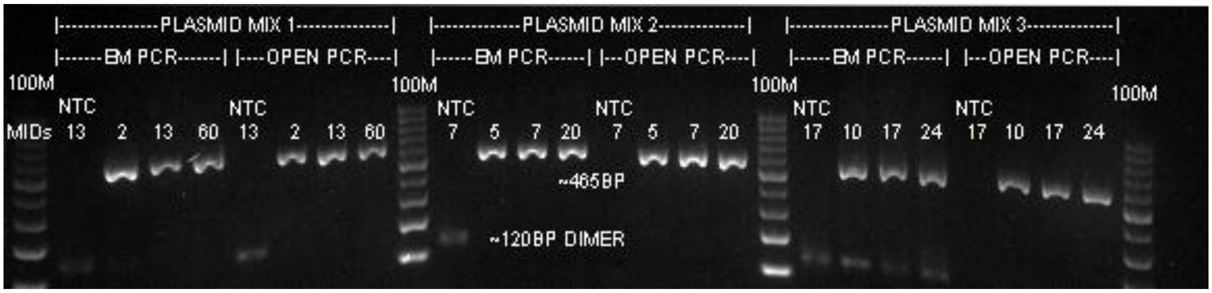


Figure 4.2.3.5.2 Optimised fusion primer open and em-PCR: Successful amplification of the three plasmid mixtures is seen above. Both the em-PCRs and the o-PCRs generated the approximately 465 bp sequence amplicon while faint, primer dimers (roughly 120 bp) are seen in the NTCs and the em-PCR for plasmid mixture 3.

4.2.4 Amplicon generation

4.2.4.1 Reaction mixture

Each of the plasmid mixtures was amplified in triplicate through both em-PCR and o-PCR, using fusion primers with a different MID for each of the replicates. Prior to each experiment, emulsion oil was prepared, vortexed and stored on ice until required. The emulsion oil consisted of 73% Tegosoft DEC (Evonik Industries, Essen, Germany), 23% mineral oil (Sigma Aldrich, Missouri, USA) and 7% Abil WE09 (Evonik Industries, Essen, Germany) in a final volume of 200 µl, of which 150 µl was used per reaction. The PCR mastermix was prepared using the Roche Diagnostics' Expand High Fidelity^{PLUS} (Basel, Switzerland) reagents at the concentration indicated below:

	Reagent	Final concentration (in 50 µl)
1.	5 x Reaction buffer	1x
2.	dNTP Mix (10 mM of each)	200 µM
3.	$MgCl_2$	2.25 mM
4.	Expand HiFi ^{Plus} Polymerase	2.625 U/rxn
5.	5 µl Plasmid mixture	2 000 c/rxn
6.	22 µl Buffer EB (Qiagen, Netherlands)	1 x

To reduce the occurrence of primer dimers, forward and reverse fusion primers were pipetted into separate areas of the reaction tubes and were only combined when the reaction mastermix was added. In addition, fusion primers were titrated to 100 nM as opposed to the manufacturer prescribed 200 nM final concentration. As indicated by the reaction recipe above, each plasmid mixture was added to its respective mastermix in the DNA loading area however, this only occurred once the o-PCR and em-PCR no-template controls (NTCs) had been removed. Elution buffer was added to the NTC reactions in the DNA loading area, before the plasmid mixtures were opened, to control for possible environmental contamination.

4.2.4.2 o-PCR preparation

Once the plasmid mixture had been added to the reaction mix, the "spiked" mastermix was added to the already aliquotted MIDs fusion primers to result in a 50 µl reaction volume. For the o-PCRs, reactions were prepared in 200 µl, thin walled PCR tubes while em-PCRs were first prepared in 1.5 ml, low DNA-binding Eppendorf (Hamburg, Germany) tubes which were later aliquotted into two 100 µl PCRs, once emulsified.

4.2.4.3 em-PCR preparation

For emulsification, 150 µl of the chilled emulsion oil was layered above the now 50 µl PCR, resulting in a 3:1 oil to aqueous-PCR ratio. These 1.5 ml tubes were then pulse vortexed five times using a Heidolph REAX top vortex mixer (Schwabach, Germany) at maximum speed, to generate an opaque emulsion. The emulsified reaction was distributed into two 100 µl reactions in single 200 µl, thin walled PCR tubes. All the prepared reactions were placed into a Veriti thermocycler (Applied Biosystems, Massachusetts, USA) and thermo-cycled using the following conditions:

94 °C for 2 min
7 x { 94 °C for 30 sec
55 °C for 30 sec
72 °C for 30 sec }
72 °C for 10 min

At the completion of the first seven cycles, the temperature is maintained at 72 °C for 10 min, allowing sufficient time for all reactions to be taken out of the cycler and placed on ice, while the em-PCRs are centrifuged for 2 min at 16300 xg. This centrifugation step breaks the emulsions droplets, replenishing the depleted template-containing droplets with the reaction mixture contained in droplets that contained no amplifiable template. The reactions were then returned to the thermocycler and the remainder of the 72 °C incubation was skipped. The following conditions were used for the remaining cycling:

43 x { 94 °C for 30 sec
55 °C for 30 sec
72 °C for 30 sec }
72 °C for 7 min
4 °C soak

Open-PCRs were stored at 4 °C until post-amplification clean-up while the em-PCRs were cleaned using isobutanol (Sigma Aldrich, Missouri, USA) and the following amplicon recovery protocol: 100 µl isobutanol was added to each of the 100 µl emulsion reactions before vortexing. Once homogenous, reactions were centrifuged at 10 000 xg for 1min to separate the aqueous and oil factions. In a well-lit post-amplification area, the aqueous faction was pipette out from underneath the oil faction using an extra length, 10 µl filter pipette tips (QSP, California, USA) and the appropriate pipette. Great care was taken not to aspirate the emulsion oil during the amplicon recovery process wherein the two em-PCR aliquots were combined to form an approximately 38 µl aqueous faction.

4.2.4.4 Gel electrophoresis

Once cleaned, 1 µl of both the em-PCR amplicons and the un-cleaned o-PCR amplicons were separated on a 2% agarose gel stained with 1x GR Green nucleic acid stain (Labgene, Châtel-Saint-Denis, Switzerland), along with a 100 bp DNA ladder (Lonza, Basel, Switzerland). These gels were electrophoresis at 80 V for 50 min and were photographed on completion.

4.2.4.5 Additional precautions taken

- Surfaces and pipettes were flushed with 20% household bleach (0.7% Sodium Hypochlorite) and 70% ethanol before and after each reaction run.
- Once PCRs were completed, reaction NTCs were separated from positive reaction to avoid any possible post-amplification contamination. These NTCs were only used later for the reaction validation through nested PCR.
- Each plasmid mixture was processed on a different day and each batch of em-PCR amplicons were recovered in a different post-amplification area to prevent reaction cross contamination.
- A different set of tips and pipettes were used in every post-amplification clean-up areas.
- PCR efficiency is influenced not only by the template-specific sequence but by the whole fusion primer (based on our own empiric data and published literature). Therefore for comparison, the same primer set – identical primers with identical MIDs were used for the em-PCR and o-PCR reactions to amplify the same plasmid mixtures. Hence em-PCR and o-PCR sequencing were performed in separate 454 sequencing runs.
- Post-amplification em-PCR and o-PCRs were never opened in the same area since cross contamination between these amplicon batches would invalidate the entire experiment.
- A fourth (additional) MID fusion primer set was included in every run to substitute any failed reactions in a specific run.

4.2.5 Amplicon and reaction validation

While NTCs yielding no specific band would be considered uncontaminated, for our purposes, very low copy number contamination would resemble a minor variant in downstream GS Junior sequencing. For this purpose, in our experiment, a reaction was only considered valid if the NTCs remained un-amplified following a nested PCR. In addition, 454's sequence-by-synthesis technology dictates that amplicons are only considered viable for sequencing if they are amplifiable. Considering the above criteria, amplicons and NTCs were validated using a nested PCR amplifying a 165 bp region within the proposed pre-nested sequence amplicon. Nested or validation PCRs were prepared using Promega's GoTaq® HotSart PCR system (Wisconsin, USA) with the primers listed in the table 4.2.5.1 below and the reaction protocol that follows.

Table 4.2.5.1 Validation PCR primers

Description	Name	Sequence (5'-3')	HXB2 binding
Forward primer	AK11	GTACCAGTAAAATTAAARCCAG	2571 to 2592
Reverse primer	30.PROT.1	GCAAATACTGGAGTATTGTATGGATTTTCAGG	2734 to 2706 (RC)

(RC) = Reverse complement

GoTaq® HotSart (Promega, Wisconsin, USA) validation PCRs were prepared using the listed reagents at their described concentration in a final reaction volume of 50 µl.

Reagent	Final concentration (in 50 µl)
1. 5 x Reaction buffer	1 x
2. AK11 FWD Primer	200 nM
3. 30.PROT.1 REV Primer	200 nM
4. dNTP Mix (10mM of each)	200 µM
5. $MgCl_2$	2.5 mM
6. GoTaq® HotStart DNA Polymerase	0.625 U/rxn
7. 0.5 µl Pre-nested amplicon / NTC	Unknown
8. 31.25 µl N.F.H_2O	1 x

The following thermocycling conditions were used on an Applied Biosystems 9700 thermocycler (Massachusetts, USA). On completion, products were once again separated on a 2% agarose gel as previously described.

95 °C for 2 min

35 x
- 95 °C for 30 sec
- 50 °C for 30 sec
- 72 °C for 30 sec

72 °C for 5 min

4 °C soak

4.2.6 Amplicon size selection and purification

Once individual runs were validated, amplicons were size selected using Invitrogen's E-Gels SizeSelect 2% system (California, USA) and recovered in nuclease-free water. Due to the large volumes required to effectively recover the amplicons from the E-Gel, size selected products were concentrated through in-house optimised ethanol precipitation described below. To prevent cross contamination, em-PCRs and o-PCRs were size selected and precipitated on separate days and in separate laboratory areas. Precipitated amplicons were resuspended in 25 µl elution buffer before being spectrophotometrically quantified and sent to the Institute for Microbial Biotechnology and Metagenomics (IMBM) at the University of the Western Cape (UWC) where they were prepared for GS Junior (454 Life Sciences, Connecticut, USA) sequencing.

4.2.6.1 Ethanol precipitation of size selected amplicons

Freshly prepared pH 5.5 sodium acetate was added to each sample to a 0.3 molarity, along with 20 µg molecular grade glycogen (Roche Diagnostics, Basel, Switzerland). Samples were briefly vortexed and maintained on wet ice before the addition of two to three volumes of ice cold 95% GR grade ethanol (Sigma Aldrich, Missouri, USA). After mixing by inversion, the DNA was pelleted at 23 100 xg for 20 min in a 4 °C centrifuge. Once completed, the supernatant was discarded revealing a clear, glossy DNA pellet that was then washed with ice cold 80% GR grade ethanol. The DNA was again pelleted at 20 000xg for 15 minutes before the supernatant was discarded and the pellet, resuspended in 25 µl warm elution buffer.

4.2.7 GS Junior library preparation

Once purified, amplicons sets were sent to the Institute for Molecular Biology and Metagenomics for GS Junior (454 Life Sciences, Connecticut, USA) library preparation and sequencing. Briefly, amplicons were quantified using the PicoGreen intercalating dye assay according to the manufacturer's specifications, before equimolar pooling. For enrichment and deep sequencing, the GS Junior Titanium emPCR Lib-A version two kit was used according to the manufacturers' specifications (Roche Diagnostics, Basel, Switzerland).

4.2.8 Bioinformatic analysis

When sequencing was completed, the GS junior data was available in standard flowgram format (.sff) which was sent to our collaborators at the South African National Bioinformatics Institute (SANBI) for analysis. Also provided to the bioinformaticist was the M13 sequencing results of the three plasmid species comprising the various plasmid mixtures. Once de-multiplexed by MID barcodes, our collaborators (doctors Simon Travers and Imogen Wright) mapped the GS Junior reads using RAMICS, a codon-aware sequence mapping tool (available online at: http://hiv.sanbi.ac.za/tools#/ramics). For each plasmid mixture, the MID reads were summed together and the number of reads mapping to the each of the three species contained in the mixture, was tabulated.

4.2.9 Statistical analysis

To gauge significant differences in sampling error, a binomial probability test was conducted using R. (version 3.1.0.), available online at: http://www.R-project.org (Gentleman & Ihaka 1997).

4.3 Results

4.3.1 Recombinant plasmid synthesis

Successful recombination and transformation was validated by the presence of white and blue colonies on the LB agar plates post overnight 37 °C incubation and by the expected outcomes of the ligation and transformation controls. The successful growth of cultures inoculated with single colonies again validated effective transformation through by observing ampicillin resistance.

4.3.2 Recombinant sequencing

Pair-wise-aligned forward and reverse M13 sequences confirmed the presence of the ligated amplicons in the three recombinants, previously validated by the observed white colonies on culture plates. The contiguous sequences can be seen in Appendix D.

4.3.3 Absolute quantification of plasmid 2.38

Figures 4.3.3.1 and 4.3.3.2 below depict the results of the ddPCR quantification of the plasmid 2.38 using the F3 primers and the QuantaSoft (BioRad, California, USA) analysis software. The scatter plot seen below (figure 4.3.3.1) is referred to as a one dimensional data plot (or a 1D plot) and depicts the observed fluorescence amplitude in relation to the number of ddPCR droplets read. QuantaSoft (BioRad, California, USA) automatically distinguishes between positive and negative droplets based on the varied fluorescence amplitude units (FAU) seen between these two populations. For the purpose of this experiment, a user-defined threshold of 7 107 FAU was used to distinguish positivity in all samples.

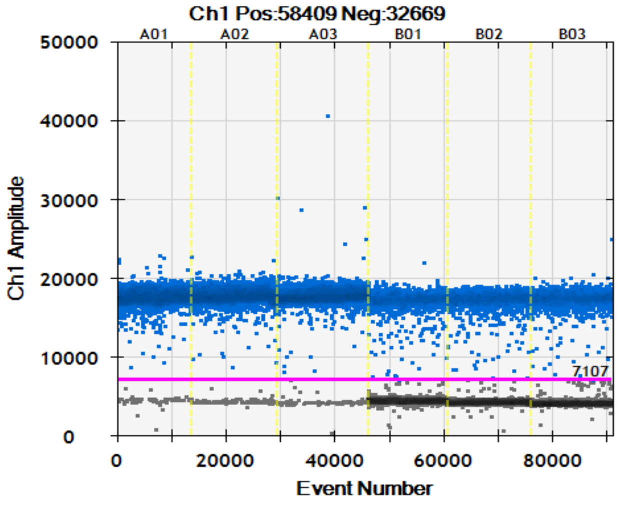


Figure 4.3.3.1 Plasmid 2.38 1D data plot: The 1D plot to the left displays the positive droplets population (blue dots) at approximately 18 000 FAU and the negative droplets population (black dots) at 5 000 FAU, separated by the 7 107 FAU user-define threshold. The first three columns correspond to the 500 copies input and the last three, the 50 copies input reactions.

In figure 4.3.3.1 above, the greatly diminished negative population seen in the 500 copies input reactions is indicative of an oversaturated ddPCR. Similar to conventional quantitative PCR (qPCR), oversaturation results in an inaccurate starting copy number estimation. The bar graph in figure 4.3.3.2 below displays the absolute count of positive and negative droplets in relation to the total number droplets, as sampled in the replicates of the plasmid 2.38 ddPCR experiments.

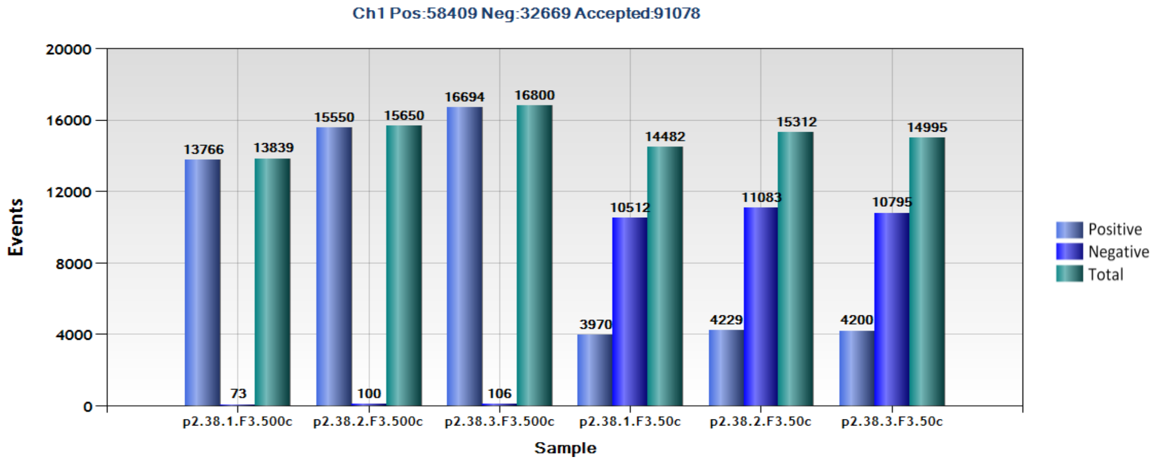


Figure 4.3.3.2 Plasmid 2.38 positivity count: The figure above displays the absolute number of positive and negative droplets in relation to the total number of droplets sampled for each reaction. Reactions containing an estimated 500 copies displayed a high ratio of positive to negative droplets while the 50 copy input reactions showed a 1:2.6 ratio. Also, 50 copy input reactions triplicates show good agreement.

In figure 4.3.3.2 above, the first nine bars on the left indicate the oversaturated 500 copies input, ddPCR triplicates with a high positive to negative ratio. The nine bars on the left correspond to the 50 copies input reaction triplicates and depict an approximate 2.6:1 ratio of negative to positive reaction droplets. For accurate Poisson statistics a lower ratio of positive to negative droplets is favourable since, in oversaturated reaction with high population rations, one droplet may contain more than one template plasmid copy.

4.3.4 Absolute quantification of plasmid 3.32

The following figures (4.3.4.1 and 4.3.4.2) depict the results of the ddPCR quantification of the plasmid 3.32 using the F3 primers and the QuantaSoft analysis software (BioRad, California, USA). The 1D plot below (figure 4.3.4.1) indicates a similar trend as observed in the plasmid 2.32 ddPCR, above. Oversaturation is once again depicted in the 500 copies input reactions (columns C01, C02 and C03), displaying a dense positive droplet population at approximately 18 000 FAU and a sparse negative population at 5 000 FAU.

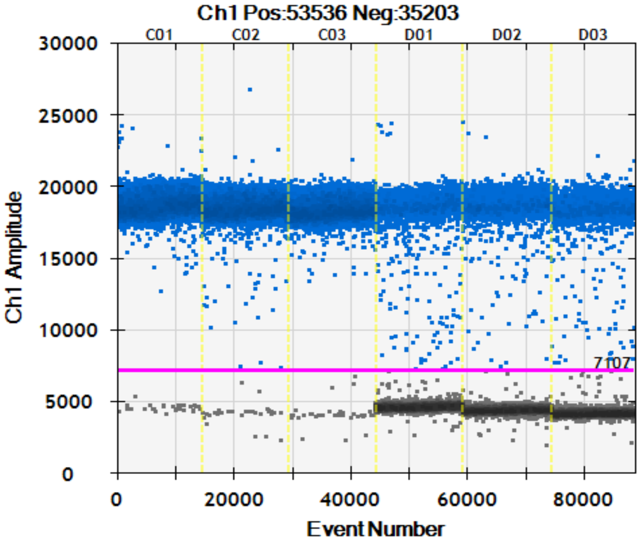


Figure 4.3.4.1 Plasmid 3.32 1D data plot: Figure 4.3.4.1 displays the 1D data plot for plasmid 3.32 ddPCR triplicates. 500 copies input reactions are in columns C01 to C03 and 50 copies input reactions, in columns D01 to D03. Both sets of triplicates display both positive (blue dots at ~18 000 FAU) and negative droplet (black dots at 5 000 FAU) populations, separated by the user-defined positivity threshold (pink line at 7 107 FAU).

Similarly to plasmid 2.38 ddPCR, plasmid 3.32's bar graph in figure 4.3.4.2 below depicts the number of positive and negative droplets in both sets of triplicate reactions (i.e.: 500 and 50 copies input), in relation to the total number of droplets sampled.

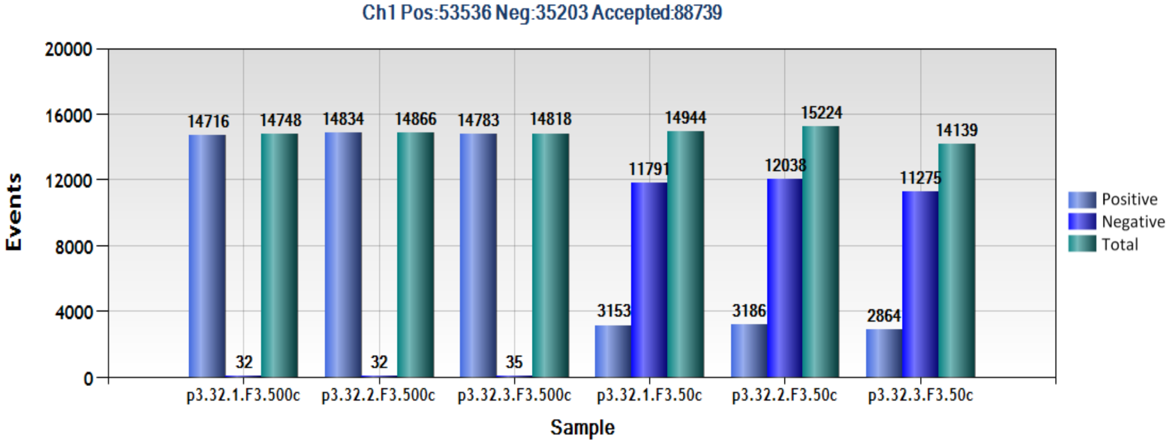


Figure 4.3.4.2 Plasmid 3.32 positivity count: Figure 4.3.4.2 displays the total number of positive, negative and the total number of droplets sampled for each reaction. Reactions containing an estimated 500 copies displayed a high ratio of positive to negative droplets while the 50 copy input reactions showed a 1:3.8 ratio. Once again, the reactions triplicates show good agreement.

In figure 4.3.4.2 above, the nine bars on the left correspond to the three oversaturated 500 copies input ddPCRs, the nine bars on the right display an approximate 3.8:1 negative to positive reaction droplet ratio for the 50 copies input reactions.

4.3.5 Absolute quantification of plasmid 3.32

The following figures (4.3.5.1 and 4.3.5.2) depict the results of the ddPCR quantification of plasmid 4.38 using the F3 primers and the QuantaSoft analysis software (BioRad, California, USA). A similar trend was observed in plasmid 4.38 reactions as in the other two ddPCRs.

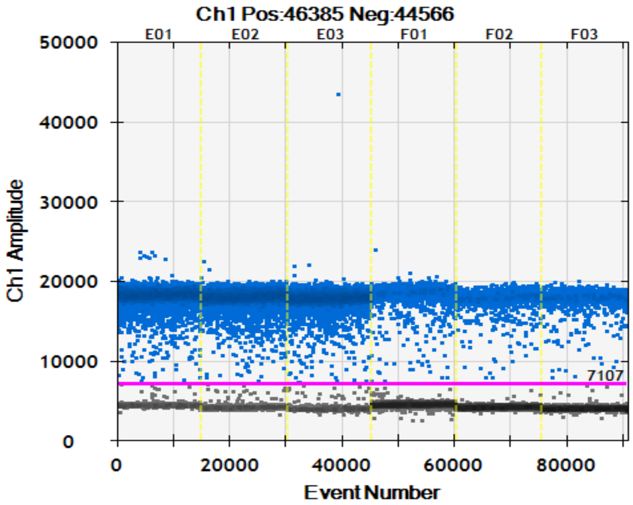


Figure 4.3.5.1 Plasmid 4.38 1D data plot: The 1D plot depicts a large positive population (blue dots) in relation to the negative droplet population (black dots) in the 500 copies input reactions (columns E01 to E03) and the converse in the 50 copies input reaction (F01 to F03). The 500 copies input reactions are however not oversaturated, as a sizable negative population is clearly visible at 5 000 FAU, while positive droplets are seen at roughly 18 000 FAU, and are separated by the user defined threshold at 7 107 FAU.

The "raining effect" observed in ddPCR indicates the presence of weakly positive droplets, resulting from an inefficient PCR. This inefficiency is possibly due to poor primer-to-template matching or, as in SYBR green based ddPCRs such as this one, primer dimers. This raining effect is more evident in the case of plasmid 4.38 ddPCRs than in the previous assays since the forward primer used in this reaction has a -2 mismatch to the plasmid, at the primer's 5' end. In this instance, poorly amplified droplets may result in a false copy number estimation since positivity is determined by a droplet's FAU in relation to the threshold and not by the detection of an unquenched fluorophore. Nevertheless, using the arbitrary 7 107 FAU threshold, the figure 4.3.5.2 below depicts the number of positive and negative droplets in both batches of triplicates (i.e.: 500 and 50 copies input), in relation to the total number of droplets sampled.

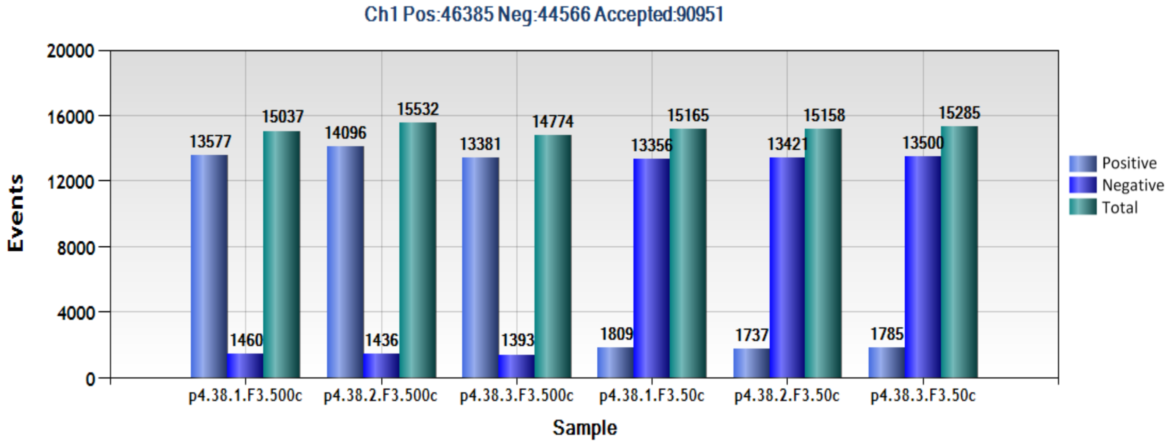


Figure 4.3.5.2 Plasmid 4.38 positivity count: Figure 4.3.5.2 above displays the number of positive, negative and the total number of droplets sampled for each of the triplicate ddPCRs. Reactions containing an estimated 500 copies displayed a high ratio of positive to negative droplets (approximately 1:0.105) while the 50 copy input reactions showed a 1:7.56 ratio. The reactions triplicates show good agreement.

Considering the absolute quantification of plasmid species 4.38 represented in figure 4.3.5.2: the nine bars on the left are indicative of nearly saturated ddPCRs while the nine on the right display a 7.6:1 negative to positive droplet ration.

4.3.6 Summation of ddPCR data

The ratio of positive to negative droplets derived from the above assays, along with the total number of droplets sampled, were used to calculate the probable starting reaction copy number displayed in table 4.3.6.1 below. The Poisson probability model is capable of estimating the initial concentration when considering a high ratio of positive to negative events. The inference is however more accurate with middle to low ratios, as there is a lower probability of two positive events occurring at one unit of measurement (e.g.: two plasmids in one droplet). Aware of the limitations of copy number estimating when over saturated, the higher dilutions data (50 copies input) was used to calculate the inferred plasmid stock concentration.

Table 4.3.6.1 Absolute quantification of plasmid species: Table 4.3.6.1 below displays the expected reaction concentrations (column two) derived by dividing the input copy number by the total ddPCR volume (rounded to the second decimal). The Poisson probability statistics results in column six shows the calculated reaction concentration. The average copy number of the 50 copies input (2.5 c/µl) reaction triplicates was used to redefine the concentration of the plasmid stocks.

Plasmid species	Expected Conc. (copies/µl)	Replicate	Positive droplets	Negative droplets	Calculated Conc. (copies/µl)
P2.38	25	1	13766	73	6107
		2	15550	100	5940
		3	16694	106	5960
	Averages		**15337**	**93**	**6002.33**
	2.5	1	3970	10512	377
		2	4229	11083	380
		3	4200	10795	387
	Averages		**4133**	**10797**	**381.33**
P3.32	25	1	14716	32	7220
		2	14834	32	7220
		3	14783	35	7120
	Averages		**14778**	**33**	**7186.67**
	25	1	3153	11791	279
		2	3186	12038	276
		3	2864	11275	266
	Averages		**3068**	**11701**	**273.67**
P4.38	25	1	13577	1460	2740
		2	14096	1436	2800
		3	13381	1393	2780
	Averages		**13685**	**1430**	**2773.33**
	2.5	1	1809	13356	149
		2	1737	13421	143
		3	1785	13500	146
	Averages		**1777**	**13426**	**146.00**

By substituting the assayed concentrations, the sample volume (5 µl) and the reaction volume (20 µl) into the formula: Conc. ^{stock} = [(Conc. ^{assayed}) x (Vol. ^{assayed})] ÷ (Vol. ^{stock}), we were able to determine the correct 1x 10^{1} plasmid stock concentration as shown below.

For plasmid 2.38: Conc. ^{stock} = [(381.33 c/µl) x (20 µl)] ÷ (5 µl); ∴ Conc. ^{stock} = 1525.32 c/µl

For plasmid 3.32: Conc. ^{stock} = [(273.67 c/µl) x (20 µl)] ÷ (5 µl); ∴ Conc. ^{stock} = 1094.68 c/µl

For plasmid 4.38: Conc. ^{stock} = [(146 c/µl) x (20 µl)] ÷ (5 µl); ∴ Conc. ^{stock} = 584 c/µl

Multiplying the corrected concentration by 100 allowed us to extrapolate the concentration of the 1x10^3 plasmid stocks, which were diluted to a concentration of 2x10^4c/µl in a 100µl final volume as indicated in table 4.3.6.2 below.

Table 4.3.6.2 Copy number correction by dilution: Table 4.3.6.2 below shows the estimated copy number of each of the plasmid stocks and the dilutions employed to correct the stock copy numbers.

	p2.38 10^3 c/µl	p3.32 10^3 c/µl	p4.38 10^3 c/µl
Current Conc.	152 532 c/µl	109 468 c/µl	58400 c/µl
Sample	13.11µl	18.27µl	34.25µl
Elution Buffer	86.89µl	81.73µl	65.75µl
Final Conc.	20 000 c/µl	20 000 c/µl	20 000 c/µl

c/µl=copies/microlitre

The corrected plasmid stocks were used to construct three, synthetic plasmid populations as described below. These plasmid species were later amplified using an in-house optimised fusion primer PCR.

4.3.7 Synthetic populations

The plasmid stocks were recombined to prepare the three plasmid mixtures as described in the method. The pooling strategy resulted in three synthetic populations listed in table 4.3.7.1 below.

Table 4.3.7.1 Synthetic populations

Mixture	Plasmid 2.38		Plasmid 3.32		Plasmid 4.38		Total
	Contribution	Copy No	Contribution	Copy No	Contribution	Copy No	Copy No
1	**90%**	18 000	**9%**	1 800	**1%**	200	20 000
2	**1%**	200	**90%**	18 000	**9%**	1 800	20 000
3	**9%**	1 800	**1%**	200	**90%**	18 000	20 000

4.3.8 Fusion primers

The modified fusion primers contained all the necessities for GS Junior sequencing. The forward primers consisted of an “Adaptor A” sequence, a “key” sequence, a ten-base-long MID and finally, the target gene-specific sequence. Similarly, reverse primers contained an “Adaptor B” sequence, a “key” sequence, an MID and a target gene-specific sequence, reverse complimented to the 3’ end of our target amplicon. These primers schematics can be seen in table 4.3.8.1 and 4.3.8.2 below.

Table 4.3.8.1 Forward fusion primers: In table 4.3.8.1 below, the "Adaptor A" sequence is *italicised*, the "Key" sequence is underlined and the MID is in **bold**, while the template-specific sequence is seen on the 3' end. The third column (P. mix) numerates the plasmid mixture sampled by that specific fusion primer.

Primer name	Sequence (5'-3') *Adaptor A* – Key – MID – Gene-specific sequence	P. mix
RTA-F-MID-13	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**CATAGTAGTG**-TGCACAYTAAATTTTCCAATTAG	**1**
RTA-F-MID-14	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**CGAGAGATAC**-TGCACAYTAAATTTTCCAATTAG	**1**
RTA-F-MID-60	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**CTACGCTCTA**-TGCACAYTAAATTTTCCAATTAG	**1**
RTA-F-MID-5	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**ATCAGACACG**-TGCACAYTAAATTTTCCAATTAG	**2**
RTA-F-MID-7	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**CGTGTCTCTA**-TGCACAYTAAATTTTCCAATTAG	**2**
RTA-F-MID-20	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**ACGACTACAG**-TGCACAYTAAATTTTCCAATTAG	**2**
RTA-F-MID-10	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**TCTCTATGCG**-TGCACAYTAAATTTTCCAATTAG	**3**
RTA-F-MID-17	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**CGTCTAGTAC**-TGCACAYTAAATTTTCCAATTAG	**3**
RTA-F-MID-24	*CGTATCGCCTCCCTCGCGCCA*-TCAG-**TAGAGACGAG**-TGCACAYTAAATTTTCCAATTAG	**3**

Table 4.3.8.2 Reverse fusion primers: In table 4.3.8.2, the "Adaptor B" sequence is *italicised*, the "Key" sequence is underlined and the MID is in **bold**, while the reverse-complemented template-specific sequence is seen on the 3' end. The third column (P.mix) once again numerates the plasmid mixture sampled by that specific fusion primer.

Primer name	Sequence (5'-3') *Adaptor B* – Key – MID – Gene-specific sequence (RC)	P. mix
RT-A-R-MID-13	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**CATAGTAGTG**-ACTAGGTATGGTGAATGCAG	**1**
RT-A-R-MID-14	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**CGAGAGATAC**-ACTAGGTATGGTGAATGCAG	**1**
RT-A-R-MID-60	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**CTACGCTCTA**-ACTAGGTATGGTGAATGCAG	**1**
RT-A-R-MID-5	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**ATCAGACACG**-ACTAGGTATGGTGAATGCAG	**2**
RT-A-R-MID-7	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**CGTGTCTCTA**-ACTAGGTATGGTGAATGCAG	**2**
RT-A-R-MID-20	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**ACGACTACAG**-ACTAGGTATGGTGAATGCAG	**2**
RT-A-R-MID-10	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**TCTCTATGCG**-ACTAGGTATGGTGAATGCAG	**3**
RT-A-R-MID-17	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**CGTCTAGTAC**-ACTAGGTATGGTGAATGCAG	**3**
RT-A-R-MID-24	*CTATGCGCCTTGCCAGCCCGC*-TCAG-**TAGAGACGAG**-ACTAGGTATGGTGAATGCAG	**3**

4.3.9 Sequencing amplicon generation

Once the open and em-PCRs had concluded, em-PCRs were cleaned using isobutanol and the previously described. A 1 µl aliquot of each reaction was diluted in 1.2x concentrated DNA loading dye. This mixture was then separated in a 2% agarose gel stained with 1x GR Green nucleic acid stain, at a constant 90 V for 50 min, along with a 100 bp DNA ladder. The resultant gel figures are shown below.

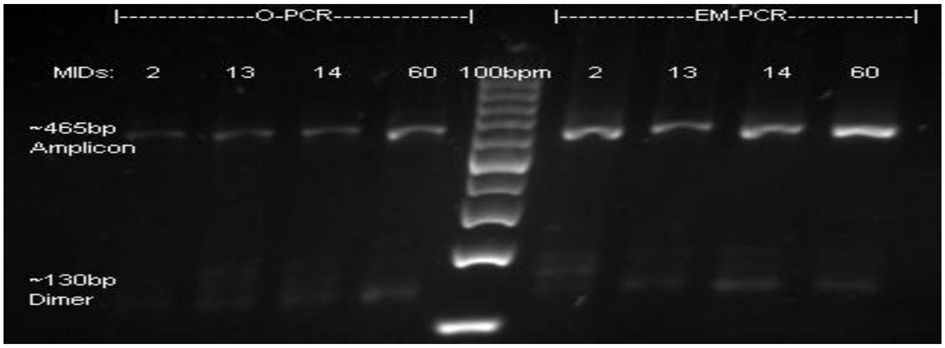

|--------------O-PCR--------------|
|--------------EM-PCR-------------|
MIDs: 2 13 14 60 100bpm 2 13 14 60
~465bp Amplicon
~130bp Dimer

Figure 4.3.9.1 Plasmid mix 1 amplified: Plasmid mix 1 was amplified through both open and emulsion PCR. The figure above displays 1 µl of the separated PCR product showing the characteristic sequence amplicon (approximately 465 bp) while faint primer dimers are also visible at roughly 130 bp. The faint band sequence amplicon seen in the MID2 o-PCR indicates poor reaction efficiency.

Figure 4.3.9.1 above depicts the successful production of the ~465 bp amplicon from plasmid mixture 1, using fusion primer sets 2, 13, 14 and 60 through both o-PCR and em-PCR. Despite all the efforts, a primer dimer is visible in most em-PCRs at approximately 130 bp. Dimerisation is however, to be expected, since the diffusion- constrained reaction in a minute droplet is expected to improve both the efficiency of the specific PCR as well as non-specific primer dimerization in droplets without template. While the NTC reactions appear negative by the absence of a specific band at approximately 465 bp, they were later screened with a nested PCR as confirmation.

Similarly to mixture 1 amplicons, amplicons generated from plasmid mixtures 2 and 3 can be seen in figures 4.3.9.2 and 4.3.9.3 below. While mixture 2 was sampled with fusion primer sets 5, 7, 20 and 24, plasmid mixture 3 was amplified with fusion primer sets 4, 10, 17, and 24.

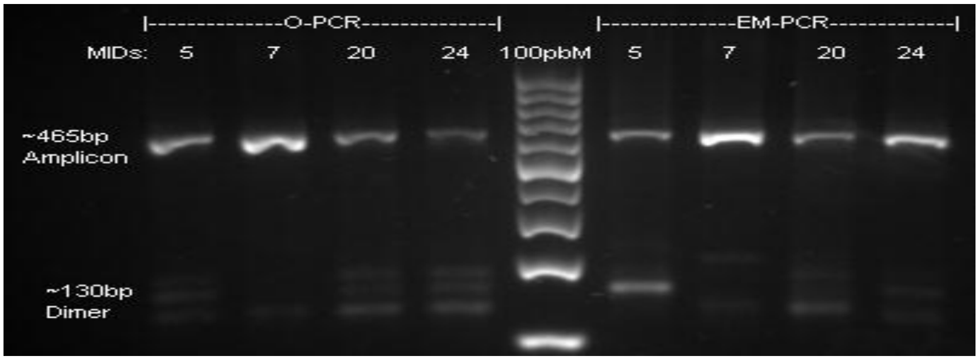

|--------------O-PCR--------------|
|--------------EM-PCR-------------|
MIDs: 5 7 20 24 100pbM 5 7 20 24
~465bp Amplicon
~130bp Dimer

Figure 4.3.9.2 Plasmid mix 2 amplified: Plasmid mix 2 was amplified through both open and emulsion PCR. The figure above displays 1 µl of the separated PCR product showing the characteristic sequence amplicon (approximately 465 bp). Faint primer dimers are also visible at roughly 130 bp despite our best efforts to bias the reaction towards the specific amplicon. The faint band sequence amplicon seen in the MID 24 o-PCR indicates poor reaction efficiency.

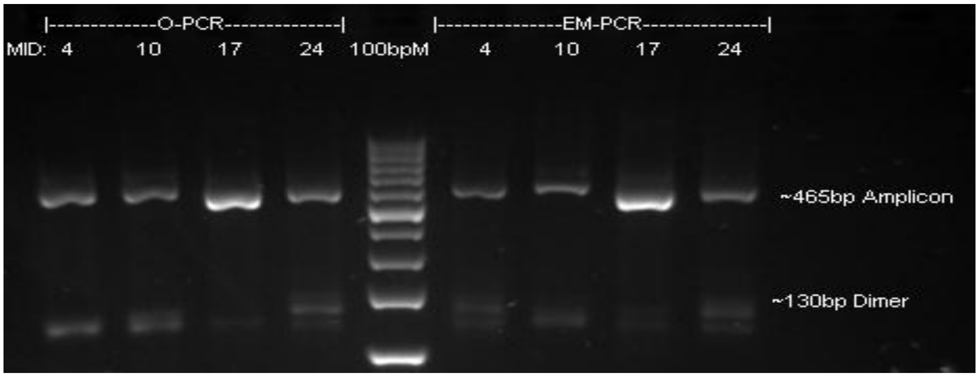

|-------------O-PCR--------------|
|---------------EM-PCR---------------|
MID: 4
10
17
24
100bpM
4
10
17
24
~465bp Amplicon
~130bp Dimer

Figure 4.3.9.3 Plasmid mix 3 amplified: Plasmid mix 3 was amplified through both open and emulsion PCR. The figure above displays 1 µl of the separated PCR product showing the sequence amplicon (approximately 465 bp). Again, faint primer dimers are seen at roughly 130 bp despite our best efforts. The faint band sequence amplicon seen in the MID 4 em-PCR indicates poor reaction efficiency.

The inclusion of a fourth fusion primer set in every open and em-PCR was precautionary to compensate for any failed reaction or failed amplicon recovery, post amplicon size selection.

4.3.10 Post amplicon generation – reaction validation

A nested PCR was used to confirm the viability or 'amplifiability' of the generated amplicons for sequencing-by-synthesis. The gel figure below depicts the 165 bp nest amplicon, generated from the amplicons prepared for deep sequencing from the three plasmid mixtures.

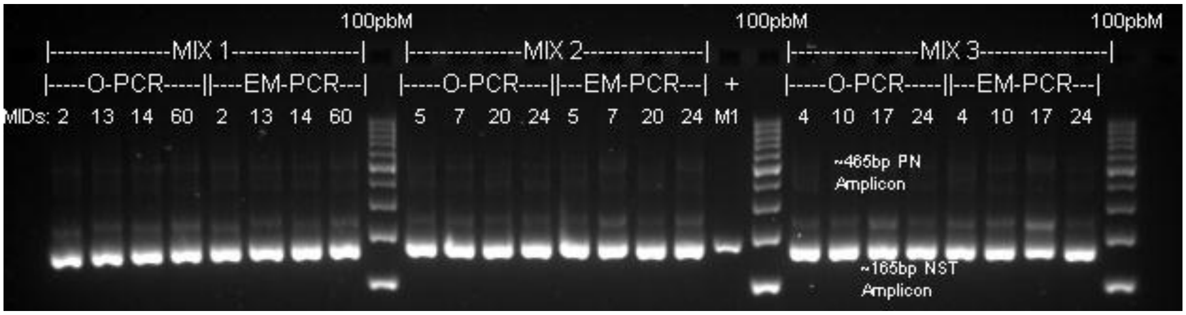

100pbM
100pbM
100pbM
|---------------MIX 1----------------|
|--------------MIX 2---------------|
|----------------MIX 3----------------|
|-----O-PCR-----||----EM-PCR---|
|-----O-PCR----||---EM-PCR---|
+
|-----O-PCR-----||---EM-PCR---|
MIDs: 2 13 14 60 2 13 14 60
5 7 20 24 5 7 20 24 M1
4 10 17 24 4 10 17 24
~465bp PN Amplicon
~165bp NST Amplicon

Figure 4.3.10.1 Sequence amplicon validation: Similar to the positive control (lane 11 from the right), all amplicon validation reactions generated the specific 165 bp nested amplicon. Also visible in nested wells are faint bands containing products from the pre-nested reaction (e.g.: the sequence amplicon at roughly 465 bp). The amplification of the specific product validates the amplifiability of these sequence amplicons.

Also visible in figure 4.3.10.1 are the sequence amplicons carried over from the pre-nested reaction at approximately 465 bp. This band is however, not indicative of reaction contamination since the pre-nested amplicon is not present in the positive control and are easily distinguishable from a specific product as seen in the positive control well 11 (from the right) above.

The controls validation reactions gel can be seen in figure 4.3.10.2 below. The negative nested NTC reactions validate the amplicon generation reaction as being free of contaminants while the positive nested reaction bands at 165 bp, validate the screening reaction. While faint, non-specific bands are visible in the pre-nested reactions, these are not considered contaminants since they were not amplified in the validation reaction.

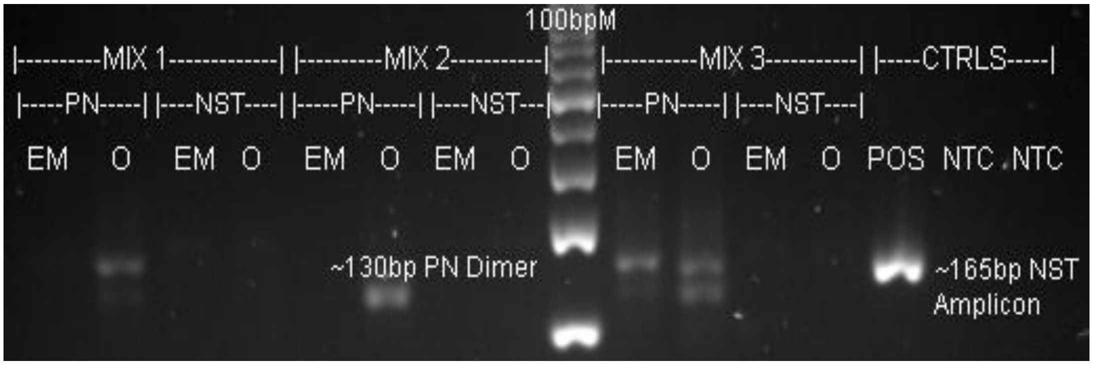

100bpM
|---------MIX 1------------| |---------MIX 2----------|
|----------MIX 3----------| |-----CTRLS-----|
|-----PN-----| |----NST----| |-----PN-----| |---NST---|
|-----PN-----| |----NST----|
EM O EM O EM O EM O EM O EM O POS NTC NTC
~130bp PN Dimer
~165bp NST Amplicon

Figure 4.3.10.2 Reaction control validation: Figure 4.3.10.2 displays the successful validation of the sequence amplicon PCR, as none of the NTC were amplifiable using the short fragment, nested PCR (NST). Primer dimers are visible in most of the pre-nested (PN) o-PCR and em-PCR NTCs (ladled O and EM above) at roughly 130 bp. All validation experiments were deemed viable by the correct functioning of the reaction positive and NTC controls.

4.3.11 Amplicon quantification post ethanol precipitation

On completion of the ethanol precipitation assay, 2 µl of the purified amplicons were quantified using a NanoDrop ND1000 (Thermo Scientific, Massachusetts, USA) according to the manufacturer's specifications. The reconstruction of the absorbance data for the o-PCR ethanol precipitation is seen on the line graph in figure 4.3.11.1 below.

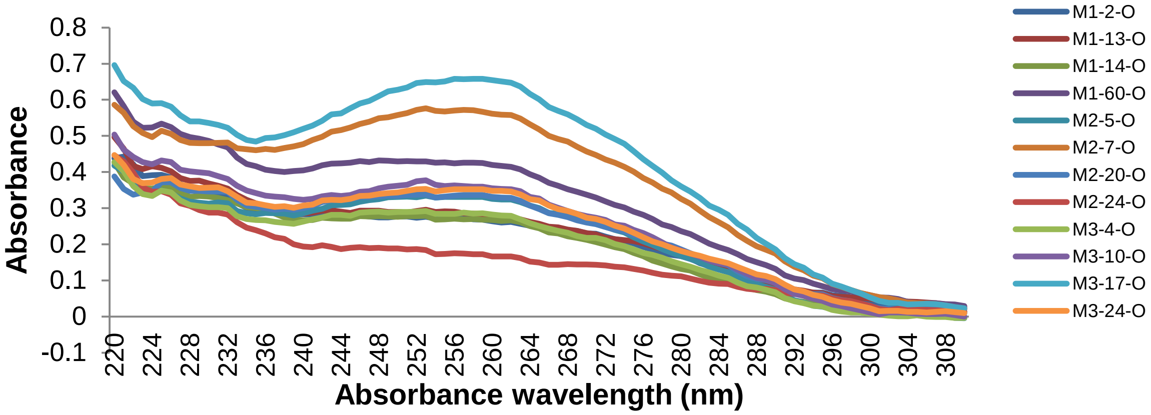


Figure 4.3.11.1 Absorbance at 230, 260 280 nm of o-PCR amplicons: The peak observed at 260 nm represents the presence of precipitated DNA in the quantified sample while the absorbance at 280 nm indicates the presence of the nucleic acid carrier, glycogen, which was added to the precipitation reaction. The absorbance observed 230 nm suggests the co-precipitation of the chaotropic agent, sodium acetate, also added to the precipitation experiment.

The line graph below (figure 4.3.11.2) displays the illustrated absorbance data of the em-PCR amplicons as measured on the NanoDrop (Thermo Scientific, Massachusetts, USA). A similar trend as in the o-PCR data is visible at absorbance wavelengths 230, 260 and 280 nm.

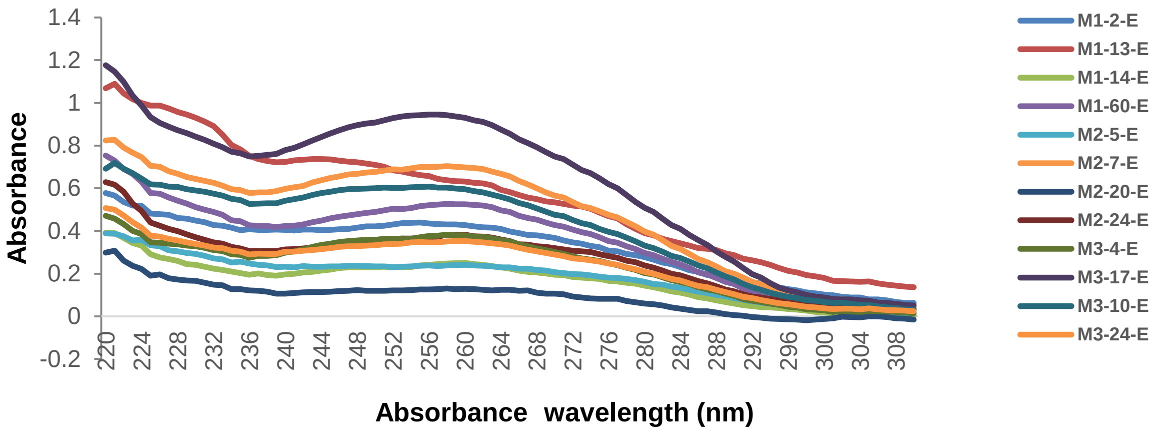


Figure 4.3.11.2 Absorbance at 230, 260 280 nm of em-PCR amplicons: As described above, the peak seen at 260 nm represents the presence of precipitated DNA, while the absorbance at 280 nm indicates the presence of glycogen and the absorbance at 230 nm displays the co-precipitation of sodium acetate.

The NanoDrop (Thermo Scientific, Massachusetts, USA) uses the absorbance values displayed above to generate the tabulated concentration and purity values seen in table 4.3.11.1 below.

Table 4.3.11.1 Quantification of size-selected sequence amplicons: The table above shows the absorbance values for the size-selected and purified o-PCR and em-PCR sequence amplicons. In column one (left) are the three plasmid mixtures that were amplified using the corresponding MID fusion primers, adjacently listed in column two.

	MID	Open PCR			Emulsion PCR		
		ng/µl	260/280	260/230	ng/µl	260/280	260/230
Plasmid Mixture 1	2	13.18	1.58	0.78	21.4	1.54	0.96
	13	13.88	1.56	0.75	31.61	1.62	0.68
	14	13.36	2.02	0.8	12.5	1.73	1.04
	60	21.02	1.78	0.86	26.24	1.77	1.03
Plasmid Mixture 2	5	16.3	1.95	1.04	12.04	1.49	0.83
	7	28.07	1.72	1.17	34.89	1.76	1.08
	20	16.55	1.78	0.96	6.44	2.13	0.78
	24	8.31	1.49	0.58	19.14	1.6	1.03
Plasmid Mixture 3	4	14.11	1.94	0.93	19	1.87	1.16
	10	17.75	1.94	0.89	46.5	1.83	1.11
	17	32.72	1.82	1.22	29.8	1.8	1.01
	24	17.39	1.92	0.97	17.66	1.69	1.04

Preferable 260/280 and 260/230 ratios are 1.8 and 2 respectively, and are achievable when precipitating larger volumes of DNA. The addition of the chaotropic salts to the precipitation reaction is essential in providing a hydrophobic environment required for DNA to migrate out of solution (Green & Sambrook 2012), and contributes to the optical density seen at 230 nm. Glycogen, a non-essential additive, acts as a nucleic acid carrier molecule by binding and concentrating multiple free DNA strands. While glycogen improves DNA recover in precipitation reactions, it contributes to the 280 nm absorbance. The spectrophotometric sample purity is thus obscured by the addition of these necessary agents.

On the basis of purity and concentration, amplicons with MIDs 13, 14 and 60 were selected to represent plasmid mixture 1, MIDs 5, 7 and 20 were chosen for plasmid mix 2, and MIDs 10, 17 and 24 were chosen for plasmid mixture 3.

4.3.12 Bioinformatic data analysis

Since the same MIDs were used in both the o-PCR and em-PCR, o-PCR amplicons were sequenced in one GS Junior (454 Life Sciences, Connecticut, USA) run and em-PCR amplicons, another. Reads were trimmed based on their quality scores and de-multiplexed by MID. Each read was mapped to one of the three plasmid sequences and the mapping counts from the three MIDs used to sample each plasmid mixture, were summed in tables 4.3.12.1 below.

Table 4.3.12.1 Mapped GS Junior reads: The table above displays the summarised results of the number of o-PCR and em-PCR GS Junior reads that mapped to a particular plasmid species. The three plasmid species are listed in column one while the plasmid mixtures are numerated in row one. Adjacent to the o-PCR mapping totals is the prevalence percentage (Prev) of the plasmid species (listed in column one) in the plasmid mixture listed in row one, above. Also under the numerated plasmid mixture and adjacent to the o-PCR mapping totals is the em-PCR mapping totals. Below each mapping total is the variant prevalence percentage in relation to the total number of reads.

	Plasmid mix 1			Plasmid mix 2			Plasmid mix 3		
	Prev	o-PCR (Prev)	em-PCR (Prev)	Prev	o-PCR (Prev)	em-PCR (Prev)	Prev	o-PCR (Prev)	em-PCR (Prev)
P2.32	**90%**	21036 (99.97%)	30786 (99.94%)	**1%**	1641 (11.41%)	1155 (5.42%)	**9%**	16070 (99.82%)	13768 (63.39%)
P3.38	**9%**	4 (0.02%)	17 (0.06%)*	**90%**	12742 (88.59%)	20051 (94.02%)*	**1%**	24 (0.15%)	7930 (36.51%)*
P4.32	**1%**	3 (0.01%)	3 (0.01%)	**9%**	0 (0.00%)	120 (0.56%)*	**90%**	5 (0.03%)	21 (0.10%)*

* indicates a binomial probability p value <0.01

From the table above, there is a clear sampling bias towards the plasmid 2.32, the species that is perfectly matched to the fusion primers. Even when the poorly matched template constitutes 9 or 10-fold more of the sampled population, as in plasmid mixtures 2 and 3 respectively, plasmid 4.28

remains vastly under-represented. Emulsification improved the sampling accuracy of mismatched templates except where plasmid 4.32 was at 1%, however, seven cycles of emulsion PCR was insufficient to correct for sampling bias. The improved sampling of plasmid 3.38 in the emulsion reaction with plasmid mixture 3 result from MID 24's inexplicable affinity for plasmid 3.38, and highlights the unpredictable nature of fusion primers. Samples in which em-PCR significantly favoured a variant relative to o-PCR are indicated with * (binomial probability p value < 0.01).

4.4 Discussion

In this study, we aimed to accurately quantify PCR associated random resampling error and bias induced by the use of fusion primers to enrich a sample for next generation sequencing. By modifying a previously published em-PCR method (Schutze et al. 2011), we tested a means of circumventing this bias using a high fidelity, emulsion PCR. The deep sequencing results obtained from the open PCR suggest that differential amplification in favour of perfectly primer-to-template matched variants greatly skews variant prevalence after PCR enrichment. Deliberately mismatched-to-template variants were vastly under-represented, despite contributing 90% of the starting population.

We attempted to reduce PCR sampling error by using seven cycles of em-PCR. As these tiny droplets soon expend the available PCR reagents, amplification within the droplet would plateau. These reagent-depleted PCR droplets were replenished with reagents from negative droplets (lacking amplifiable template), by breaking the emulsion through centrifugation. A previous study conducted by Nakano (Nakano et al. 2003), produced larger PCR droplets (two to 10 µm); this method made use of 13 cycles of emulsion PCR before the PCR substrates in template-containing droplets, were depleted, resulting in PCR plateau. Nakano then replenished the positive reactions by breaking the emulsion and continued to enrich specific template through an additional 30 PCR cycles. We did not know the optimal emulsified cycle number but chose seven cycles based on the much smaller droplet size in our reaction which would result in an earlier PCR plateau. The seven cycles we used were evidently insufficient to fully overcome the lower yield of mismatched compared to perfectly-matched templates. Templates with a good match to the primers still largely out-competed the poorly matched template 4.32.

Our original hypothesis was that droplet partitioned PCR would overcome mismatches through improving efficiency despite mismatches. However, our data suggest that the benefit of droplet PCR with reference to poorly matched templates is largely an effect of constraining efficient reactions through the small droplet size (and early PCR plateau) relative to the inefficient non-specific reaction. It is not clear what number of cycles would be optimal, moreover the previous ddPCR experiment and raining effect with mixture 4.32 suggests that even a full complement of 40 PCR cycles would not fully overcome inefficient PCR due to many mismatches. Although the difference in PCR yield for the poorly matched template was small, it was significantly improved and seemed to have a big effect on the yield of the plasmid 3.38 amplicons (with only two mismatches) when the perfectly matched template was not the most abundant. Overall, sampling of mismatched templates was marginally improved.

A standard polymerase chain reaction (PCR) relies on primer sequences that bind to complementary sequences in the target sequence, allowing the polymerase to extend the primed genomic regions. Template-specific primers, by nature, are a great source of selection bias as they are designed to discriminate specific genomic region. As described by Kanagawa in 2003, amplification of genomic variants with mismatches to the primer species is largely inefficient. Remedially, they recommend the use of degenerate bases to compensate for diversity (Kanagawa 2003). In the case of highly diverse viruses, degeneracy may nevertheless not yield any primer species with a perfect match to the template. Moreover, it decreases the ratio of primers that are perfect template complements. Degenerate primers may also increase the risk of non-specific binding to other than the desired targets resulting in non-specific amplification.

In a letter to the editor, Shan-Lu Liu of the University of Washington highlights the dilemma of random sampling error through resampling that occurs when using PCR enrichment to sample minority variants. Liu describes how PCR could randomly amplify particular species in the population and concludes that this can be avoided by PCR amplification of individual species. Single-template amplification is however costly as limiting dilutions and multiple PCRs are required (Liu et al. 1996).

A recent study however, found similar bias with single genome amplification and sequencing to PCR followed by cloning (Jordan et al. 2010). Nevertheless the high accuracy of these methods has allowed the robust characterization of HIV evolution in various compartments (Salazar-Gonzalez et al. 2011), the study of transmitted populations (Danaviah et al. 2015) and intra patient virus evolution in individuals on long-term suppressive cART (Kearney et al. 2014). NGS potentially offers a more efficient process to sequence minor variants. The high loci coverage of NGS potentially allows for a high resolution. This depth of sampling enables the quantification of minor variants at very low frequencies provided that the PCR mediated enrichment process resulted in an accurate representation of these variants at the end of the enrichment process.

A real world example of the mentioned situation is when using deep sequencing for HIV drug-resistant variant detection. High fidelity PCR is required and is the most cost effective means of enriching resistance-associated genes. As previously publicised, sampling bias is induced when using specific primers to sample a diverse population (Kanagawa 2003; Polz & Cavanaugh 1998). NGS at high template coverage (deep sequencing) has been used to detect minor variant drug-resistant populations before commencement of therapy (Simen et al. 2009). When non-nucleoside reverse transcriptase inhibitor (NNRTI)-resistant, minority populations are detected before therapy (using by AS-PCR), their presence is predictive of NNRTI regimen failure (Li et al. 2013; Li et al. 2011; Halvas et al. 2010). Using NGS to identify minority, drug- resistant populations could thus inform therapy choices however, its value in many other clinical settings is still not established.

The limited application of deep sequencing is probably largely due to technical limitations. When sampling low frequency variants, resampling error could contribute to random inaccurate representation of these variants in the sequenced population, whereas in case of a poor primer template match a particular PCR may result in a biased underrepresentation of these variants. Lastly sequencing induced errors due to homopolymer read error affects NGS platforms such as 454 (454 Life Sciences, Connecticut, USA) or Ion Torrent (Life Technologies, California, USA), whereas Illumina platforms (California, USA) are been prone to biased sequencing dependent on AT:GC ratio (Loman et al. 2012). Partitioning of PCR in a myriad of fento or atto-litre individual reactions such as an emulsified PCR could potentially overcome both the random resampling error and biased amplification.

In 2011, Cassandra Jabara of the University of North Carolina published a study that addressed PCR resampling and presented RNA random primer ID tagging as a solution (Jabara et al. 2011). In their study, Jabara synthesized cDNA during a reverse transcription step using an HIV-specific primer that consists of a reverse complemented genome-specific region, an eight random-base RNA ID tag, and a PCR tag. After cDNA synthesis, a primer complimentary to the PCR tag and a 90-base-long forward fusion primer were used to enrich sequenceable amplicons, each containing a unique eight-base random ID sequence. Post-sequencing and filtering, template sequences with identical IDs are condensed into one consensus (representing an original template), allowing one to determine the exact number of original RNA variants sampled. However, although this method allows accurate quantification of variants that are sequenced, it requires a high concentration of template, it is associated with PCR bias and results in a large loss of cDNA species that produce sequences; which limits the usefulness for detection of low frequency variants (Brodin et al. 2015).

In response to these problems, researchers from the National Cancer Institute (Boltz et al. 2015) improved the RNA ID tagging protocol. Briefly, RNA is reverse transcribed using cDNA primer with a 10-random-base tagging approach, similar to Jabara's method. However, instead of 90-mer fusion primers with 5' adaptor, 22-mer primers containing uracils at the 5' end are used for PCR amplification. Subsequently, 14 base adenine overhangs are generated by digesting uracils with uracil N-Glycosylase. This allows highly efficient adaptor and barcode ligation. Similar to the Jabara method, reads with the same randomly generated IDs (10 bases in this case) are consolidated into a consensus sequence as representing a single cDNA species. This modified assay obviates the need to use large fusion primers and has been shown to reduce bias and increase the number of cDNA templates sampled (Boltz et al. 2015). While this modified random tagging approach improves one's ability to correct the sampling error post-sequencing, few other methods, except the cumbersome single genome sequencing approach, have the potential of correcting sampling error at the source.

The alternative approach to template tagging, which could overcome PCR enrichment error, is partitioning PCR into minute reactions. Emulsion PCR is commercially available from RainDance Technologies (Massachusetts, USA) and BioRad (California, USA) and addresses both random sampling error and primer biased sampling through efficiency constraint reactions. Both digital droplet PCR platforms offer the advantage of standardised droplet volumes with identical intra-droplet selection bias, since varied droplet volumes may also induce biased amplification as seen in figure 4.4.1. The QX200 ddPCR system (BioRad, California, USA) was used in our study to accurately quantify the three plasmid species used for our assay and shows great repeatability within triplicates, as it was designed for quantification. The BioRad droplet generator (California, USA) uses proprietary emulsion oil and reaction mixtures produces 1 nanolitre reactions by the use of embedded microfluidics and differential air pressure.

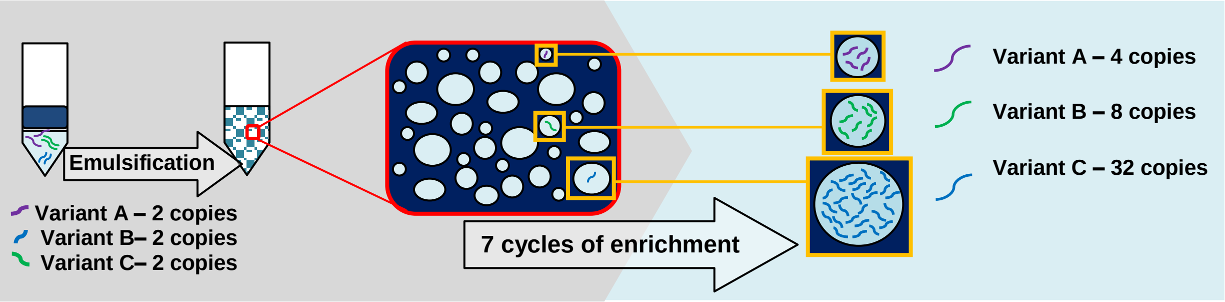


Figure 4.4.1 Biased enrichment in un-equilibrated droplets: The figure above demonstrates the biased enrichment of variant C as a result of the starting template being randomly assigned a larger droplet than variants A and B, at the start of the em-PCR. Before em-PCR enrichment (on the left), all variants had equal prevalence however, since variant A was in a tiny emulsion droplet, it was only enriched two-fold while Variants B and C are resampled more often before their droplets are reagent depleted (on the right).

QX200 technology (BioRad, California, USA) is however, not expandable to next generation library preparation because of its limited sample input volumes and low fidelity. In contrast, RainDance technologies (Massachusetts, USA) offer ddPCR platform that is expandable to next generation sequencing amplicon library preparation by means of an adaptor ligation protocol. The RainDance system (Massachusetts, USA) generates as many as five million reaction droplets from a 50 µl PCR and boasts the highest ddPCR sensitivity, owing to its high droplet count. However, as with the BioRad's QX200 (California, USA), generating these 10 picoliter reactions requires expensive, single-use consumables and an expensive droplet generator.

The methods we present here are widely implementable in any resource limited laboratory with three basic pieces of equipment (i.e.: a vortex, a micro-centrifuge and a thermocycler). The use of diffusion constraint reactions could abrogate the decreased efficiency associated with primer-mismatched variants, but it is not clear if it could overcome poor yield due to primers with multiple mismatches. It may however require multiple experiments and NGS runs to find optimal conditions which may would result in a PCR enriched population that represent the original variant proportions.

Our method modified that of Schütze (Schutze et al. 2011) by using a high fidelity polymerase to allow deep sequencing and by including a depleted droplet replenishing step, adapted from Kanagawa (Kanagawa 2003). It should be noted though, that Schütze's methods were a simplification of a BEAMing protocol (so named on the basis of the four principal components: beads, emulsion, amplification, and magnetics), which was used to enrich magnetic-bead-bound amplicons through em-PCR (Diehl et al. 2006). BEAMing mediates enrichment of amplicons with the required 3' or 5' adaptor sequences and forms the foundation of bead-based enrichment in most next generation library preparation protocols.

4.5 Conclusion

In the interests of simplicity and cost, we used an approach that could not control droplet volumes. The variation in droplet size would result in different droplets reaching PCR plateau after a varying number of cycles which complicates assay optimization. Other aspects of our em-PCR requiring further optimisation include:

- The number of cycles used to saturate the droplets – Provided that the droplet volumes are equilibrated, saturation of positive droplets is advantageous as it would result in equal representation of all the PCR mutated variants, after the emulsion is broken. As was evident from the em-PCR mapped GS Junior analysis, seven emulsified cycles were insufficient to handicap the perfectly matched templates and allow the poorly matched templates to catch up.
- The addition of bovine serum albumin (BSA) – To further improve processivity, BSA addition is advised as it prevents the polymerase from getting trapped and denatured in the water-oil interface (Williams et al. 2006)
- Increased polymerase concentration to increase productivity – Shao (2011) demonstrated a 2.5-fold increase in em-PCR efficiency when using 2.5-fold the polymerase concentration recommended for the open PCR (Shao et al. 2011).

Despite the limitations of our method, it clearly demonstrated that PCR enrichment could misrepresent the relative abundance of the variants in the sampled population, when using next generation sequencing. Although our emulsified reactions resulted in a statistically significant increase in mismatched variant representation in most reactions, except when sampling 1% of the template with three mismatches, it only marginally improved sampling error. Once further developed, emulsified PCR could be a possible solution the observed “founder effect” quantified in this study. While conventional PCR with shorter primers might not induce as radical a bias as is the case of fusion primers, the effect of terminal mismatches is possibly similar when sampling highly diverse sample populations, such as HIV.

Once fully optimised, emulsification as a means of partitioning templates during PCR enrichment may offer an affordable solution to PCR resampling error which is less cumbersome than the primer ID approach and less costly than commercial digital droplet PCR for NGS experiments. A simple PCR enrichment process without significant sampling error would expand the application of NGS in the fields of intra patient evolution, molecular epidemiology of diverse infectious agents in host populations, virus discovery or any other application where a highly diverse population is sampled with specific primers.

4.6 References

Boltz, V.F. et al., 2015. Analysis of Resistance Haplotypes Using Primer IDs and Next Gen Sequencing of HIV RNA Methods. *CROI*, p.1. Available at: http://www.croiconference.org/sites/default/files/posters-2015/593.pdf [Accessed July 19, 2015].

Brodin, J. et al., 2015. Challenges with using primer IDs to improve accuracy of next generation sequencing. *PloS one*, 10(3), p.e0119123.

Danaviah, S. et al., 2015. Evidence of long-lived founder virus in mother-to-child HIV transmission. *PloS one*, 10(3), p.e0120389.

Diehl, F. et al., 2006. BEAMing: single-molecule PCR on microparticles in water-in-oil emulsions. *Nature methods*, 3(7), pp.551–559.

Fisher, R. et al., 2012. Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. *Journal of virology*, 86(11), pp.6231–6237.

Garcia-Diaz, A. et al., 2013. Evaluation of the Roche prototype 454 HIV-1 ultradeep sequencing drug resistance assay in a routine diagnostic laboratory. *Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology*, 58(2), pp.468–473.

Gentleman, Robert & Ihaka, Ross (Department of Statistics, University of Auckland). 1997. R-Project. Available at: http://www.r-project.org [Accessed July 18, 2015].

Green, M.R. & Sambrook, J., 2012. *Molecular cloning: a laboratory manual* 4th ed., New Yourk: Cold Spring Harbor Laboratory Press.

Halvas, E.K. et al., 2010. Low frequency nonnucleoside reverse-transcriptase inhibitor-resistant variants contribute to failure of efavirenz-containing regimens in treatment- experienced patients. *The Journal of infectious diseases*, 201(5), pp.672–680.

Hori, M., Fukano, H. & Suzuki, Y., 2007. Uniform amplification of multiple DNAs by emulsion PCR. *Biochemical and biophysical research communications*, 352(2), pp.323–328.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Jordan, M.R. et al., 2010. Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. *Journal of virological methods*, 168(1-2), pp.114–120.

Kanagawa, T., 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of bioscience and bioengineering*, 96(4), pp.317–323.

Kearney, M.F. et al., 2014. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS pathogens*, 10(3), p.e1004010.

Keys, J.R. et al., 2015. Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS research and human retroviruses*, 31(6), pp.658–668.

Li, J.Z. et al., 2013. Impact of minority nonnucleoside reverse transcriptase inhibitor resistance mutations on resistance genotype after virologic failure. *The Journal of infectious diseases*, 207(6), pp.893–897.

Li, J.Z. et al., 2011. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA*, 305(13), pp.1327–1335.

Liu, S.L. et al., 1996. HIV quasispecies and resampling. *Science (New York, N.Y.)*, 273(5274), pp.415–416.

Loman, N.J. et al., 2012. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5), pp.434–439.

Mohamed, S. et al., 2014. Comparison of ultra-deep versus Sanger sequencing detection of minority mutations on the HIV-1 drug resistance interpretations after virological failure. *AIDS (London, England)*, 28(9), pp.1315–1324.

Nakano, M. et al., 2003. Single-molecule PCR using water-in-oil emulsion. *Journal of biotechnology*, 102(2), pp.117–124.

Polz, M.F. & Cavanaugh, C.M., 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology*, 64(10), pp.3724–3730.

Salazar-Gonzalez, J.F. et al., 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *Journal of virology*, 82(8), pp.3952–3970.

Salazar-Gonzalez, J.F. et al., 2011. Origin and evolution of HIV-1 in breast milk determined by single-genome amplification and sequencing. *Journal of virology*, 85(6), pp.2751–2763.

Schutze, T. et al., 2011. A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. *Analytical biochemistry*, 410(1), pp.155–157.

Shao, K. et al., 2011. Emulsion PCR: a high efficient way of PCR amplification of random DNA libraries in aptamer selection. *PloS one*, 6(9), p.e24910.

Simen, B.B. et al., 2009. Low-Abundance Drug-Resistant Viral Variants in Chronically HIV-Infected, Antiretroviral Treatment – Naive Patients Significantly Impact Treatment Outcomes. *Journalof infectious diseases,* 199(5), pp. 693-701.

Simen, B.B. et al., 2014. An international multicenter study on HIV-1 drug resistance testing by 454 ultra-deep pyrosequencing. *Journal of virological methods*, 204, pp.31–37.

Varghese, V. et al., 2010. Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. *PloS one*, 5(6), p.e10992.

Williams, R. et al., 2006. Amplification of complex gene libraries by emulsion PCR. *Nature methods*, 3(7), pp.545–550.

Wright, S., 1942. Statistical genetics and evolution. *Bulletin of the American Mathematical Society*, 48(4), pp.223–247.

Zhou, S. et al., 2015. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of virology*. [Cited ahead of print].

4.7 Appendix D

(Additional information for section 4.3.2 Recombinant sequencing)

Below are the contiguous Sanger sequences for the three plasmid species used in this study. Forward and reverse primer binding regions are in **bold** while mismatches to the fusion primers are underlined.

Plasmid 2.38

CAGATGATACAGTATTAGAAGATATAAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGAGGAA
TTGGAGGATTTATCAAAGTAAGACAGTATGATCAAATACCTATAGAAATTTGTGGAAAGAAGACTA
TAGGTACAGTATTAATAGGACCTACACCGGTCAACATAATTGGAAGAAATATGTTGACTCAGCTTG
GA**>TGCACACTAAATTTTCCAATTAG>**TCCTATTAAAACTGTACCAGTAAAATTAAAGCCAGGAAT
GGATGGCCCAAAGGTCAAACAATGGCCATTGACAGAAGAAAAAATAAAAGCATTAACAGCTATTT
GTGATGAAATGGAAAAGGAAGGGAAAATTACAAAAATTGGGCCTGAAAATCCATACAACACTCCA
GTATTTGCTATAAAAAAGAAGGACAGTACTAAGTGGAGAAAATTAGTAGATTTCAGGGAACTCAAT
AAAAGAACTCAAGACTTTTGGGAAGTTCAACTAGGAATACCACACCCAGCAGGGTTAAAAAAGAA
AAAATCAGTGACAGTGCTGGATGTAGGGGATGCATATTTTTCAGTTCCTTTAGATGAAAACTTCAG
GAAATATA**<STGCATTCACCATACCTAGT<**ATAAATAATAAAACACCRGGGATTAGGTATCAATAT
AATGTGCTGCCACAGGGATGGAAAGGATCACCAGCAATATTCCAGAGTAGCATGACAAGAATTTT
AGAGCCTTTTAGGGCACAGAATCCAGACATAGTTATCATTCAATATATGGATGACTTGTATGTAGG
ATCTGACTTAGAAATAGGGCAGCATAGAGCAAAAATAGAGGAGTTAAGGGAACATTTATTGAAAT
GGGGATTTACCACACCAGACAAAAAGCATCAAAAAGAACCCCCATTTCTTKGGATGGGGTWTGAA
CTCCATCSTGACAAATGG

Plasmid 3.32

CAGATGATACAGTATTAGAAGAAATAAATTTGCCAGGAAGATGGAAACCAAAAATGATAGGAGGA
ATTGGAGGTTTTATCAAAGTAAGACAATATGAACAAATAGTTATAGAAATTTGTGGAAAAAAGGCT
ATAGGTTCAGTATTAGTAGGACCTACACCTGTCAACATAATTGGAAGAAATATGTTGACTCAGCTT
GGA**>TGCACATTGAATTTTCCAATTAG>**TCCTATTGAAACTATACCAGTAAAATTAAAGCCAGGAA
TGGATGGCCCAAAAGTTAAACAATGGCCATTGACAGAAGAAAAAATAAAAGCACTAACAGCGATT
TGTGAAGAGATGGAAAAGGAAGGAAAAATTACAAAAATTGGGCCCGAAAATCCATACAACACTCC
AGTATTTGCCATAAAAAAGAAGGACAGTACTAAATGGAGAAAATTAGTAGATTTCAGGGAACTCAA
TAAAAGAACTCGAGACTTTTGGGAAGTTCAATTAGGGATACCACATCCAGCAGGATTAAAAAAGA

AAAAATCAGTGACAGTGCTGGATGTGGGGGATGCATATTTTTCAGTTCCTTTAGATGAAGGCTTTA
GGAAATATA**<CTGCATTCACAATACCTAGT<**ATAAACAATGAAACACCAGGAATTAGATATCAATA
CAATGTACTTCCACAGGGATGGAAAGGATCACCAGCAATATTCCAGAGTAGCATGACAAAAATCT
TAGAACCCTTTAGGACACAAAATCCAGACATAGTTATCTATCAATATATGGATGACTTGTATGTAG
GATCTGACTTAGAAATAGGGCAACATAGAGCAAAAATAGAGGAGTTAAGAGAACATTTATTGAGAT
GGGGATTTACCACACCAGACAAGAAGCATCAGAAAGAACCCCCATTTCTGTGGATGGGGTATGA
ACTCCATCCTGACAAATGG

Plasmid 4.38

CAGATGATACAGTATTAGAGGATATAAATTTGCCAGGAAAATGGAAACCAAAAATGATAGGAGGA
ATTGGAGGTTTTATCAAAGTAAGACAGTATGATCAAGTACATATAGAAATTTGTGGAAAAAAGGCT
ATAGGTACAGTATTAGTGGGACCTACACCCGTCAACATAATTGGAAGAAATATGTTGACTCAGATT
GGA**>TGCACATTGAATTTCCCAATAAG>**TTCCATTGAAACTGTACCAGTAAAATTAAAGCCAGGAA
TGGATGGCCCAAAGGTTAAACAATGGCCATTGACAGAAGAGAAAATAAAAGCATTAACAGCAATT
TGTGAAGAAATGGAAAAGGAAGGAAAAATTACAAAAATTGGGCCTGAAAATCCATATAACACTCCA
GTATTTGCCATAAAAAAGAAGGATAGTACTAAGTGGAGAAAATTAGTAGATTTCAGGGAACTCAAT
AAGAGAACTCAAGACTTTTGGGAAGTTCAATTAGGAATACCACACCCAGCAGGGTTAAAAAAGAA
AAAATCAGTAACAGTACTGGATGTGGGGGATGCATACTTCCCAGTTCCTTTAGATGAAAACTTCA
GGAAATATA**<CTGCATTCACCATACCTAGT<**ATAAACAATGAAACACCAGGGATTAGATATTAATA
TAATGTGCTGCCACAGGGATGGAAAGGATCACCAGCCATATTCCAGTGTAGCATGACAAAAATCT
TAGAGCCCTTTAGAACACAAAATCCAGAAATGGTCATCTATCAATATATGGATGACTTGTATGTAG
GGTCTGACTTAGAAATAGGGCAACACAGAGCAAAAATAGAGAAGTTAAGAGACCATCTATTAAGG
TGGGGATTTACCACACCAGATAAGAAACATCAGAAAGAACCCCCATTTCTTTGGATGGGGTATGA
ACTCCATCCTGACAAATGG

5 Conclusion

5.1 Study summary

Next generation sequencing (NGS) is a useful tool for identifying minor HIV variants with drug resistant mutations however, more in depth studies are needed to determine the clinical implications of the presence of these populations (Li 2013). We set out to test the utility of various NGS platforms as alternatives to our in-house human immunodeficiency virus (HIV) drug resistance testing (DRT), which employs PCR and Sanger sequencing (referred to here as bulk sequencing). We selected NGS rather than allele-specific PCRs and oligonucleotide ligation assays because NGS read lengths offer the benefit of mutation linkage study. In addition, NGS offers potential cost savings, accessible through the use of molecular identifiers (MIDs) and sample pooling. The high template coverage due to massive parallel sequencing on an NGS platform permits the detection of minority, drug-resistant virus populations. When detected in therapy-naïve patients, these drug-resistant minority variants may increase the risk of antiretroviral therapy failure of non-nucleoside reverse transcriptase inhibitor (NNRTI)-based combination antiretroviral therapy (cART) failure (Li et al. 2011).

The primary aim for chapters two and three was to detect minority, drug-resistant viral populations in clinical settings for which there is limited published data on these viral variants. In chapter two, we included first-line, NNRTI-regimen-experienced patients, at the stage of failing a second-line, lopinavir (LPV) boosted with low dose ritonavir (LPV/r)-based cART, with HIV RNA loads >500 copies per millilitre of blood. Most of these patients had poor drug exposure (likely poor adherence) as evident from low plasma and hair LPV concentrations (van Zyl et al. 2011). In this cohort, the prevalence of major protease inhibitor (PI) resistance mutations, as detected by bulk sequencing, was 7% (Fisher et al. 2012), similar to that observed by other researchers investigating cohorts failing PI-based regimens (Wallis et al. 2010; El-Khatib et al. 2010). However, we did not know if minority, drug resistance mutations (DRMs) associated with the first-line, or the current PI regimen, would be detectable.

Overlapping amplicons were generated from the products of a low fidelity, two-step reverse transcription-pre-nested PCR, sampling viral RNA isolated after second-line therapy failure. These pre-nested products were also used for bulk sequencing, while the fusion primer amplicons were sequenced on the FLX Titanium NGS platform (454 Life Sciences, Connecticut, USA).

NGS improved DRT by identifying minority, drug-resistant viral populations in five of the seven patients included in this study. In addition, minority populations with first-line cART DRMs were detected, although no major PI-resistant viral populations were identified. We observed discordance in the prevalence of certain DRMs when comparing the DRM quantities (as determined by NGS) to

DRMs identified by in-house DRT, when using an expected bulk sequencing threshold of 20% (Palmer et al. 2005). This disagreement is possibly due to biased priming of specific templates, followed by PCR resampling. Also, K65R was identified at low frequencies in all patients. This finding may have been PCR-induced since our pre-nested PCR used low fidelity and *reverse transcriptase* codon 65 is found in a homopolymer run.

Our study of PI resistance had several limitations:

- We did not make use of a high fidelity PCR system for the first round of amplification, which could have contributed to PCR-induced error.
- Our target sequences did not include recently described domains in *gag* and *envelope* which may be associated with PI-failure (Rabi et al. 2013; Sutherland et al. 2015).
- Due to cost considerations and limited sample availability, we did not perform NGS at the time of first-line failure in these patients.

The lack of detectable majority variant drug resistance in this cohort may be attributed to the rapid waning of PI-resistant variants in the absence of sustained drug pressure (in patients with poor or intermittent adherence). Under reduced selection pressure, PI-resistant variants wane due to the high fitness price of PI resistance mutations (Rosenbloom et al. 2012).

The 15 infants sampled in chapter three were exposed to the prevention of mother-to-child HIV transmission (PMTCT) regimen but otherwise treatment-naïve. Initial drug resistance in infants (a combination of transmitted and PMTCT induced resistance) is challenging to identify in PMTCT failures since some DRMs may wane below the sensitivity threshold of bulk sequencing, at the time of testing (Rowley et al. 2010). However, minority, drug-resistant populations present at baseline (before therapy initiation) may affect therapy response (Metzner et al. 2009; Jourdain et al. 2010; Paredes et al. 2010; Li et al. 2012; Li et al. 2013).

Infant samples included in this study had previously been assayed with bulk sequencing and the residual plasma samples were used for NGS. To include all cDNA and reduce random enrichment error, viral *protease* and *reverse transcriptase* were enriched through 14 parallel pre-nested and seven parallel nested PCRs. Quantified cDNA assisted in determining the limit of minor variant detection while clonal sequencing (cloning and Sanger sequencing) validated the NGS results.

Both the Ion Torrent Personal Genome Machine™ (PGM) (Life Technologies, California, USA) and the Solexa/Illumina MiSeq (California, USA) improved the DRM detection in all patients, below the sampling thresholds of bulk sequencing and clonal sequencing. Bulk sequencing was validated by both NGS platforms but there was better agreement between the MiSeq and the clonal sequencing. This finding was probably due to homopolymer reading error on the PGM, which resulted in spurious insertions or deletions and contributed to the loss of coverage at homopolymer sites.

Considering the study findings, the Solexa/Illumina MiSeq (California, USA) may be ideal for minor variant detection in PMTCT-exposed infants, who were initiated onto cART. In addition to the longer read lengths attainable with bidirectional sequencing (2x 300 bp), the superior homopolymer sequencing makes the MiSeq (Illumina, California, USA) a suitable candidate for HIV DRT. Moreover, the MiSeq (Illumina, California, USA) generates up to 15 gigabases of sequence data which can allow for the pooled testing of a large number of samples in one sequencing run. To our knowledge, this was the first study to perform and validate comparative NGS of PMTCT failures and the only the third study to sequence HIV using the PGM (Life Technologies, California, USA).

In chapter four we attempt to quantify and overcome PCR resampling and the primer-induced selection bias brought about by fusion primers. An amplicon sequencing approach, using fusion primers, focuses NGS coverage on the region of interest and is thus most efficient for deep sequencing to detect rare variants. Moreover, this approach obviates the need to ligate sequencing adaptors as they are included in the fusion primers. As the commercial ligation kits are expensive, this can bring about substantial cost reduction. When amplicons are generated for the 454 GS Junior (454 Life Sciences, Connecticut, USA), the MID fusion primers may be longer than 60 bases and introduce a strong primer-induced selection bias. An alternative approach to next generation deep sequencing to detect rare variants is the use of limiting dilution PCR (LD-PCR). This method prevents PCR selection bias as individual species are amplified separately and sequenced separately. A more elegant solution to LD-PCR could be to reduce biased enrichment prior to NGS by partitioning templates in droplets, through emulsion PCR (em-PCR).

We compared two enrichment approaches for amplicon sequencing: A standard or “open PCR” and a PCR where the reagents were partitioned in small droplets by a water-oil emulsion. Both the emulsion (em-PCR) and open PCRs were with the same high fidelity enzyme mastermix and fusion primers containing sequencing adaptors for 454 amplicon sequencing. When we kept the mastermix emulsified for the full duration of the PCR reaction, there were no visible PCR products. This was probably because the picoliter droplets reached PCR plateau early, which reduced processivity. We therefore opted for seven emulsified cycles, to allow partitioned PCR in the first critical priming steps. This was followed by emulsion breaking (by centrifugation) and the completion of the remaining 43 PCR cycles.

We investigated the enrichment of mixtures of three different plasmids: one plasmid sequence had no mismatches with the fusion primers template-specific regions, while the other two plasmids had deliberate mismatches with the primers. In each mixture, one of the three plasmid species constituted 90%, while the other two, either 9% or 1% of the total. Experiments were performed in triplicate for each of three different plasmid mixtures using the same set of nine fusion primers for the open and em-PCRs. After PCR enrichment the amplicons were sequenced on the 454 platform to investigate any random error or biased enrichment.

There was extremely uneven representation in NGS reads of the different variants depending on the number of fusion primer-template mismatches: our findings showed that a majority poorly matched variant represented less than 1% after enrichment, even when present at 90% in the original sample. The plasmid template with the most mismatches appeared as a minor variant relative to the perfectly-matched variant, even when it was at a 10-fold higher concentration in the sampled plasmid mixture. Seven cycles of emulsified PCR could not overcome the biased enrichment but resulted in a significantly increased representation of the plasmid with fewer mismatches, and a marginal effect on the one with the most mismatches. These findings emphasise the strong biased enrichment when using fusion primers, as seen by others (Berry et al. 2011).

Data from quantitative PCR (qPCR) is informative in understanding this phenomenon. Primer template mismatches often result in a cycle threshold (Ct) delay with qPCR or, too many mismatches could result in no detectable amplification (Pang et al. 2011). The Ct delay is thought to be due to the poor efficiency of a mismatched PCR during the first rounds of amplification when the imperfectly matched primer and templates lack stability essential for polymerase extension. However, once the primers are incorporated, the proportion of perfectly matched template (primer derived), compared to the original mismatched template, would exponentially increase with each cycle. Thus, the efficiency of amplification would improve.

Partitioning templates may enhance PCR efficiency by preventing competition of imperfectly matched templates with the primer-derived, perfect match. Nevertheless, the poor initial efficiency would still result in an initial delay and may explain the limited benefit of using only seven emulsified cycles. However, with too many emulsified cycles we found a poor PCR yield, apparently as the minute droplets reach early PCR plateau. Although, with an optimal cycle number, PCR plateau may be used to an advantage as when in an emulsion, a perfectly efficient droplet PCR would reach plateau earlier than a droplet with a lot of primer-template mismatches. This may handicap the perfectly matched reactions, compared to the mismatched ones and reduce enrichment bias. Finding the optimal number of cycles may require extensive validation and many more NGS experiments which fall outside the scope of this thesis. Our approach was nevertheless the first attempt to overcome fusion primer resampling error using an em-PCR approach, which allows droplet-partitioned PCR without expensive microfluidics and which did not require magnetic beads to facilitate amplicon recovery.

Furthermore, we have learned the importance of appropriate controls for NGS experiments. Reaction contaminants that can be amplified using the reaction primers or fusion primers could be perceived as minor variant populations. Bearing this in mind, we recommend similar control measures as are used for single-copy detection assays. As larger fragment PCRs are relatively inefficient, the no-template controls may not yield visible bands on gel electrophoresis, even when low-level contamination is present. Therefore nested PCR or qPCR could be used to screen no-template controls and reaction preparation areas despite being labour intensive.

Our investigations had the following limitations: In the first study, investigating minor variant resistance in patients failing a second-line PI regimen, we did not have samples available at the first-line failure episode and therefore could perform deep sequencing at both failure time points. It was therefore not known if there were any pre-existing minor drug-resistant variants before PI therapy. In this experiment we also did not quantify the cDNA before enrichment. Our assumptions with reference to minor variant thresholds were therefore based on the patient viral load and expected extraction and reverse transcription efficiencies.

In the second study investigating minor variant resistance in PMTCT-exposed infants we addressed this limitation and quantified the pre-enrichment cDNA. Thus we were able to report the average times a template was enriched based on the read coverage and original template quantity. We detected additional mutations with deep sequencing in these patients but were not able to assess the effect minor variant, NNRTI resistance mutations as these patients were initiated on a high genetic barrier, PI-based regimen. Whether these minor variant mutations would affect therapy response should NNRTIs be re-introduced in subsequent regimens, is unknown.

Our last experiment attempted to detect and improve fusion primer associated resampling error. The key limitation was that we did not know the optimal number of emulsified cycles (before breaking the emulsion by centrifugation), as too many emulsified cycles resulted in very low amplicon yield. This poor processivity was probably due to early PCR plateau in small droplets. Seven emulsified cycles, although able to reduced resampling error, were too few to overcome it, in the case of multiple template primer mismatches.

We have used NGS to characterise minor variants with HIV drug resistance in unique patient populations: infants exposed to a dual PMTCT regimen of zidovudine and nevirapine and adults who were failing a second-line protease inhibitor regimen without major PI resistance mutations detectable by bulk sequencing. Despite the potential of NGS for the detection of minor, drug-resistant variants, practical and technological challenges would need to be overcome before it could achieve wider implementation.

These challenges include pre-analytical, analytical and post-analytical aspects. Pre-analytically, the detection of very rare events requires adequate sample collection, especially in patients with low viral loads. Failing an adequate sample, rare variants would not be detected despite extensive enrichment and very high sequencing coverage. Analytically, PCR enrichment can introduce a particular selection bias, described as resampling error (Liu et al. 1996), especially when fusion primers are used (Berry et al. 2011). In our study of infant drug resistance after PMTCT exposure, we used parallel PCR reactions to reduce sampling error. This approach did not make use of fusion primers and is therefore less biased but, due to the high cost of ligating adaptors, this is less cost-efficient.

An innovative approach to overcome resampling error associated with fusion primers is the tagging of cDNA with random tags during reverse transcription. This strategy allows post-analytic correction of relative variant copy number by collapsing all reads with the same random tag into one (Jabara et al. 2011; Boltz et al. 2015; Zhou et al. 2015; Keys et al. 2015). Despite the value of this approach, it has been associated with a low recovery of sample diversity due to the insensitivity of the approach (Boltz et al. 2015; Zhou et al. 2015; Keys et al. 2015; Brodin et al. 2015). Our approach to reducing bias with a simple em-PCR has shown potential but requires further development to optimise its ability to overcome PCR resampling error. As the costs of NGS are expected to decrease in future, further improvement of sample enrichment is required to make optimal use of this technology. The improved and accurate use of targeted enrichment with fusion primers, which obviates expensive adaptor ligation, could make this much more affordable than Sanger sequencing, especially on the Ion Torrent (Life Technologies, California, USA) and MiSeq (Illumina, California, USA) platforms.

Post-analytically, the availability of open bioinformatic pipelines such as the UCSD's (University of California, San Diego) Datamonkey (Pond & Frost 2005), would facilitate quality assurance and interpretation of sequencing data. Our SANBI (South African National Bioinformatics Institute) collaborators have developed a bioinformatic alignment tool (RAMICS) which allows improved alignment of multiple sequences and hence the correction of spurious insertions or deletions generated through homopolymer read error on the 454 or Ion Torrent sequencing platforms (Wright & Travers 2014).

With recent advances in NGS technology and accompanied development in bioinformatics, there is a need to develop and customise workflows for particular applications such as HIV minor variant sequencing, making optimal use of the newer high coverage platforms such as MiSeq (Illumina, California, USA) or Ion Torrent (Life Technologies, California, USA). We have performed some of the first experiments on these platforms but further development would be required to fully utilise their potential, specifically in the clinical sector (Li & Kuritzkes 2013). This may expand the use of NGS to the characterization of low frequency variants in patients on long-term therapy, to characterise HIV persistence, which is has depended on the use of LD-PCR and single genome Sanger sequencing (Kearney et al. 2014). Developing improved workflows for HIV may also have application when sequencing other highly diverse biological populations with similar challenges.

5.2 References

Berry, D. et al., 2011. Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. *Applied and Environmental Microbiology* , 77 (21), pp.7846–7849.

Boltz, V.F. et al., 2015. Analysis of Resistance Haplotypes Using Primer IDs and Next Gen Sequencing of HIV RNA Methods. *CROI*, p.1. Available at: http://www.croiconference.org/sites/default/files/posters-2015/593.pdf [Accessed July 19, 2015].

Brodin, J. et al., 2015. Challenges with using primer IDs to improve accuracy of next generation sequencing. *PloS one*, 10(3), p.e0119123.

El-Khatib, Z. et al., 2010. Viremia and drug resistance among HIV-1 patients on antiretroviral treatment: a cross-sectional study in Soweto, South Africa. *AIDS (London, England)*, 24(11), pp.1679–1687.

Fisher, R. et al., 2012. Deep sequencing reveals minor protease resistance mutations in patients failing a protease inhibitor regimen. *Journal of virology*, 86(11), pp.6231–6237.

Jabara, C.B. et al., 2011. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50), pp.20166–20171.

Jourdain, G. et al., 2010. Association between detection of HIV-1 DNA resistance mutations by a sensitive assay at initiation of antiretroviral therapy and virologic failure. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 50(10), pp.1397–1404.

Kearney, M.F. et al., 2014. Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. *PLoS pathogens*, 10(3), p.e1004010.

Keys, J.R. et al., 2015. Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS research and human retroviruses*, 31(6), pp.658–668.

Li, J.Z., 2014. HIV-1 Drug-Resistant Minority Variants: Sweating the Small Stuff. *The Journal of infectious diseases*, 209(5), pp.639-641.

Li, J.Z. & Kurizkes, D.R., 2013. Clinical Implications of HIV-1 Minority Variants. *Clinical infectious diseases,* 56(11), pp.1667-1674.

Li, J.Z. et al., 2013. Impact of minority nonnucleoside reverse transcriptase inhibitor resistance mutations on resistance genotype after virologic failure. *The Journal of infectious diseases*, 207(6), pp.893–897.

Li, J.Z. et al., 2011. Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA*, 305(13), pp.1327–1335.

Li, J.Z. et al., 2012. Relationship between minority nonnucleoside reverse transcriptase inhibitor resistance mutations, adherence, and the risk of virologic failure. *AIDS (London, England)*, 26(2), pp.185–192.

Liu, S.L. et al., 1996. HIV quasispecies and resampling. *Science (New York, N.Y.)*, 273(5274), pp.415–416.

Metzner, K.J. et al., 2009. Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 48(2), pp.239–247.

Palmer, S. et al., 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *Journal of clinical microbiology*, 43(1), pp.406–413.

Pang, X. et al., 2011. Increased sensitivity for various rotavirus genotypes in stool specimens by amending three mismatched nucleotides in the forward primer of a real-time RT-PCR assay. *Journal of Virological Methods*, 172(1–2), pp.85–87.

Paredes, R. et al., 2010. Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *The Journal of infectious diseases*, 201(5), pp.662–671.

Pond, S.L.K. & Frost, S.D.W., 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)*, 21(10), pp.2531–2533.

Rabi, S.A. et al., 2013. Multi-step inhibition explains HIV-1 protease inhibitor pharmacodynamics and resistance. *The Journal of Clinical Investigation*, 123(9), pp.3848–3860.

Rosenbloom, D.I.S. et al., 2012. Antiretroviral dynamics determines HIV evolution and predicts therapy outcome. *Nature Medicine*, 18(9), pp.1378–1385.

Rowley, C.F. et al., 2010. Ultrasensitive detection of minor drug-resistant variants for HIV after nevirapine exposure using allele-specific PCR: clinical significance. *AIDS research and human retroviruses*, 26(3), pp.293–300.

Sutherland, K.A. et al., 2015. Gag-Protease Sequence Evolution Following Protease Inhibitor Monotherapy Treatment Failure in HIV-1 Viruses Circulating in East Africa. *AIDS research and human retroviruses*. [Cited ahead of print]

Wallis, C.L. et al., 2010. Protease inhibitor resistance is uncommon in HIV-1 subtype C infected patients on failing second-line lopinavir/r-containing antiretroviral therapy in South Africa. *AIDS research and treatment*, 2011.

Wright, I.A. & Travers, S.A., 2014. RAMICS: trainable, high-speed and biologically relevant alignment of high-throughput sequencing reads to coding DNA. *Nucleic acids research*, 42(13), p.e106.

Zhou, S. et al., 2015. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next Generation Sequencing of HIV-1 Genomic RNA Populations. *Journal of virology*. [Cited ahead of print]

Van Zyl, G.U. et al., 2011. Low lopinavir plasma or hair concentrations explain second line protease inhibitor failures in a resource-limited setting. *Journal of acquired immune deficiency syndromes (1999)*, 56(4), pp.333–339.