

https://dspace.org/

Evaluation of HIV- 1 Drug resistance Next Generation Sequencing Data Analysis Pipelines

Kana, Vibha

2025
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/49893
Kana, Vibha . (2025). Evaluation of HIV- 1 Drug resistance Next Generation Sequencing Data Analysis Pipelines [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/49893

Downloaded from a DSpace repository. https://dspace.org/

Evaluation of HIV- 1 Drug resistance Next Generation Sequencing Data Analysis Pipelines

Vibha Kana

Student Number: 0304450F

A dissertation submitted to the Faculty of Health Science, University of the Witwatersrand, in fulfilment of the requirements for the degree of Master of Science (MSc (Med)). Johannesburg, 2025.

DECLARATION

I, Vibha Kana, declare that this Dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University



14 day of November 2025 in Johannesburg

ABSTRACT (250 words)

Next-generation sequencing (NGS) has transformed HIV drug resistance (HIVDR) testing by enhancing sensitivity for detecting drug resistance mutations (DRM). However, variability in the analytical outcomes generated by different bioinformatics pipelines for HIVDR interpretation raises concerns about the reliability and reproducibility of reported results. This study presents a secondary analysis of sequencing data from a 2019 South African HIVDR surveillance study, evaluating six widely used NGS analysis pipelines, PASeq, HyDRA, MiCall, DeepChek HIV, Stanford HIVdb-NGS (beta), and Exatype. Focusing on eight key mutations (K65R, K70R, K103N, V106M, M184V, G190A, K219E, and P225H), pipeline performance was compared under default and customized 'Winnipeg' parameter settings, assessing sensitivity, specificity, accuracy, and linearity. Across pipelines, sensitivity ranged from 92.3% to 100.0% and specificity from 96.4% to 100.0%, depending on the applied detection thresholds, while linearity analysis (R^2 values) demonstrated strong agreement (≥0.98) across all platforms. Results demonstrated notable variability in DRM detection, especially for low-frequency variants. Implementation of Winnipeg settings marginally improved concordance across pipelines, reducing discrepancies in DRM identification. This study highlights the significant impact of bioinformatics parameters on HIVDR reporting and advocates for standardized quality control measures to enhance consistency in NGS-based resistance testing. By identifying pipeline-specific biases and proposing quality control recommendations, the findings contribute to improving the reliability of genotypic HIVDR testing, with important implications for clinical management, surveillance programs, and optimization of HIV treatment strategies in resource-limited settings.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Dr. Gillian Hunt and Dr. Adriaan Basson, for their immense patience, and mentorship, and for generously sharing their expertise in HIV drug resistance. Their guidance was fundamental to the success of this research.

I am especially grateful to Dr. Catherine Scheepers, Rudolph Serage, and Dr. Jonathan Featherston for their technical expertise, guidance in bioinformatics, and support with script writing. Their contributions were essential in refining my analyses and ensuring the accuracy of my findings.

I would like to acknowledge Exatype (Hyrax Biosciences) and DeepChek (Advanced Biological Laboratories) for providing access to their platforms through complimentary credits and licenses, which greatly facilitated the analyses conducted in this study.

A heartfelt thank you to the National Institute for Communicable Diseases (NICD) for providing the opportunity and resources to complete this qualification.

I also wish to sincerely thank my colleagues at the HIV Drug Resistance Laboratory — Ewalde Cutler, Cheri van der Walt, and Monalisa Kalimashe — for their continuous support, encouragement, and for providing the time and space I needed to complete this work. Special thanks to Cheri for always bringing laughter at moments when it was most needed.

Lastly, I am deeply grateful to my Mum and to David Lunuburg for their constant support, patience, and belief in me throughout this journey. I also wish to remember my late Dad, whose memory has been a source of strength and inspiration.

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION....3

1.1 Global HIV burden and treatment progress....3

1.2 Aetiology and Genetic Diversity of HIV....3

1.3 Genetic Composition and Functions of the HIV Genome....4

1.4 HIV Structure....5

1.5 HIV Life Cycle....6

1.6 HIV natural history and transmission....6

1.7 Antiretroviral Therapy (ART)....7

1.7.1 Nucleos(t)ide Reverse Transcriptase Inhibitors (NRTI)....7

1.7.2 Non-nucleoside reverse transcriptase inhibitors (NNRTI)....8

1.7.3 Protease Inhibitors (PI)....8

1.7.4 Integrase Inhibitors (INI) / Integrase strand transfer inhibitors (INSTI)....8

1.8 Evolution of Treatment Guidelines....8

1.9 Emerging HIV Drug Resistance....9

1.9.1 Prevalence and Trends of ADR in HIV....10

1.9.2 Surveillance of ADR in South Africa....11

1.10 HIV Drug Resistance Testing Methods....14

1.11 Alternative Methods for Genotypic HIVDR Testing....14

1.12 Next-Generation Sequencing (NGS)....15

1.13 Bioinformatics Pipelines HIVDR NGS data interpretation....17

1.13.1 Primary Analysis 18
1.13.2 Secondary Analysis 20
1.13.3 Tertiary Analysis 21
1.13.4 NGS file types 21
1.14 NGS Bioinformatics Pipelines for HIVDR Data Interpretation 23
1.15 Differences in NGS pipelines 23
1.16 Aims and Objectives 25
CHAPTER 2 - MATERIALS AND METHODS 27
2.1 National HIVDR surveillance study, 2019 27
2.1.1 Sample collection 27
2.1.2 Total Nucleic acid extraction 27
2.1.3 Reverse Transcription-Polymerase Chain Reaction (RT-PCR) Amplification and amplicon clean-up 27
2.1.4 Library preparation and sequencing 28
2.1.5 Data management and quality control 28
2.2 Comparative analysis of study sequencing data using six different pipelines 30
2.3 Pipeline data management procedures 31
2.3.1 DeepChek® 31
2.3.2 Exatype 33
2.3.3 HyDRA 36
2.3.4 MiCall 38

2.3.5 PASeq 40
2.3.6 Stanford HIVdb-NGS (beta) 42
2.3.7 Data processing for comparative analyses 47
2.3.8 Threshold Filtering 48
2.4 Statistical analysis 48
2.4.1 Generation of a reference standard as proxy for “true value” 48
2.4.2 Data Processing Classification and Threshold 48
2.4.3 Determination of Classification Outcomes 48
2.4.4 Determination of Sensitivity, Specificity, and Accuracy 48
2.4.5 Linearity 49
2.5 Ethics clearance 49
CHAPTER 3 - RESULTS 50
3.1 Introduction 50
3.2 Difference in minimum value reporting 50
3.3 Description of dataset 50
3.4 Comparison of frequency values of mutation detection obtained using default and recommended (Winnipeg) QC settings 52
3.4.1 Analytical sensitivity 52
3.4.2 Analytical specificity 55
3.4.3 Accuracy 58
3.4.4 Linearity 60
3.5 Investigating discrepant samples 63

CHAPTER 4 - DISCUSSION 68

4.1 Detection of mutations below 1.00% 69

4.2 Assessment of pipeline sensitivity, specificity, accuracy and linearity 70

4.3 Application of Winnipeg settings to improve agreement between pipelines . 71

4.4 Pipeline processes that may be contributing to the marginal differences noted 72

4.5 Examining Minimum Prevalence Thresholds: Variations in Mutation Detection Across NGS Pipelines 77

4.5.1 **Study Limitations:** Although this study provides valuable insights into the performance of multiple HIVDR bioinformatics pipelines, several limitations should be acknowledged. Firstly, the analysis was based on secondary data from a single national surveillance dataset, which may not fully capture the variability of sample quality or viral subtypes observed in routine diagnostic settings. Secondly, the evaluation was limited to eight key drug resistance mutations, and therefore does not represent the full spectrum of resistance-associated mutations across all drug classes. Thirdly, the reference standard used, a median-derived comparative measure, served as a practical proxy but not an absolute biological truth. Finally, differences in pipeline accessibility and version updates at the time of analysis may have influenced specific performance outcomes. 78

Despite these constraints, the study design and comparative framework provide a robust foundation for evaluating inter-pipeline variability and identifying critical factors influencing HIVDR detection accuracy. 78

4.5.2 Conclusion: 79

CHAPTER 5 REFERENCES 81

CHAPTER 6 APPENDIX 1 92

LIST OF FIGURES

Figure 1.1. **Global overview of HIV-1 subtypes** Map illustrating the geographical distribution of major HIV-1 subtypes (A–K) and circulating recombinant forms (CRFs) (Ndashimye, Reyes and Arts, 2023).....4

Figure 1.2.**Schematic representation of the HIV-1 genome and virion structure.** (A) Genome organisation showing the major structural, enzymatic, and regulatory genes. (B) Virion structure depicting the envelope glycoproteins, capsid, and associated enzymes) (Musumeci, Riccardi and Montesarchio, 2015).....5

Figure 1.3. **Prevalence of HIV-1 drug-resistance mutations among adults with viraemia, South Africa (2019).** Data derived from the national HIVDR surveillance programme) (Steegen et al., 2024). 12

Figure 1.4 **Workflow for comparative analysis of HIVDR pipelines.** Overview of the analytical process comparing six NGS pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, Stanford HIVdb-NGS) for detection of eight HIVDR mutations across 762 samples. 18

Figure 2.1: **Workflow for comparative pipeline analysis of HIV-1 drug-resistance data.** Overview of the analytical process comparing six NGS pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS) for detection of eight HIVDR mutations across 762 samples under default and Winnipeg settings. 31

Figure 3.1. Comparison of mutation frequency distributions across six NGS HIVDR pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS) for eight key drug-resistance mutations (K65R, K70R, K103N, V106M, M184V, G190A, K219E, P225H). Results are based on 753 successfully genotyped samples analysed under default settings. Outputs from six pipelines categorized into <1.00% (not detected), ≥1.00%–<5.00% (very low frequency variant), ≥5.00%–<20.00% (low frequency variant), ≥20.00%–<80.00% (intermediate frequency variant), and ≥80.00% (dominant variant). 51

Figure 3.2. **Linear regression analysis of mutation frequencies obtained from DeepChek, Exatype, and HyDRA pipelines.** Each point represents the observed mutation frequency versus the median (expected) frequency for values > 1 %.

Regression lines indicate concordance between reported and expected frequencies for these pipelines.>1.00% 61

Figure 3.3. **Linear regression analysis of mutation frequencies obtained from PASeq, Stanford HIVdb-NGS, and MiCall pipelines.** Each point represents the observed mutation frequency versus the median (expected) frequency for values > 1 %. Regression lines indicate concordance between reported and expected frequencies for these pipelines.>1.00% 62

LIST OF TABLES

Table 1.1.Clinical significance of most frequently detected mutations (K103N, M184V, V106M, K65R, K70R, P225H, K219E and G190A) within this dataset Extracted from https://hivdb.stanford.edu/. All listed mutations are located within the reverse transcriptase (RT) region of the pol gene and associated with resistance to nucleoside (NRTI) and/or non-nucleoside (NNRTI) reverse transcriptase inhibitors. (Clutter et al., 2016; Tang, Liu and Shafer, 2012).....12

Table 1.2. Overview of NGS Platforms: Key Features Including Sequencing Chemistries, Read Lengths, Throughput, Accuracy, and Limitations (Gu, Miller and Chiu, 2019; Manyana et al., 2021). 16

Table 1.3. Overview of file types used in the primary, secondary and tertiary analysis stages of NGS data. 22

Table 1.4. Quality control settings proposed by the Winnipeg consensus (Ji et al., 2018) 24

Table 2.1 Summary of Steps in the MiCall Pipeline with Input and Output Files. Available from https://cfe-lab/github.io/MiCall/steps 39

Table 2.2 Overview of Steps and Tools Used in the PASeq Algorithm for viral genotyping. Available from https://paseq.org/about.php.....41

Table 2.3 Overview of pipeline features, including coverage, quality thresholds. Pipelines used are DeepChek®, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS (beta).....45

Table 3.1. Sensitivity of mutation detection using pipeline-specific default and recommended QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%. Abbreviations: ND = Not Detected. 52

Table 3.2. Specificity of mutation detection using pipeline-specific default and recommended QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection

values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%. Abbreviations: ND = Not Detected. 55

Table 3.3. Accuracy of mutation detection using pipeline-specific default QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%. 58

- Table 3.4. List of samples with large differences in mutations frequency reporting across the different pipelines.....65

ABBREVIATIONS

AAVF	Amino Acid Variant Format
ADR	Acquired Drug Resistance
AIDS	Acquired Immune Deficiency Syndrome
APOBEC	Apolipoprotein B MRNA editing cyclic polypeptide-like
ART	Antiretroviral therapy
BAM	Binary Alignment Map
Bp	Base pair
BWA	Burrows–Wheeler Aligner
CSV	Comma Separated Value
DNA	Deoxyribonucleic acid
DTG	Dolutegravir
FASTA	Text-based file format for representing nucleotide sequences
FASTQ	Format for storing biological sequences and their quality scores
GATK	Genome analysis toolkit
HIV	Human Immunodeficiency Virus
HIVDR	HIV Drug Resistance
HIVVL	HIV Viral Load
IN	Integrase
INSTI	Integrase Strand Transfer Inhibitor
INDELS	Insertions and deletions
Kb	Kilobase
MALDI-TOF	Matrix-assisted Laser Desorption Ionization Time of Flight
µL	microlitre
mL	millilitre
NGS	Next-generation Sequencing
NICD	National Institute for Communicable Diseases
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	Nucleoside reverse transcriptase inhibitor
OLA	Oligonucleotide ligation assay
PCR	Polymerase Chain Reaction
PF	Passing Filter
PI	Protease inhibitor
PR	Protease
QC	Quality Control
QS	Quality Score
RNA	Ribonucleic acid
RT	Reverse Transcriptase
SA	South Africa
SAM	Sequence Alignment Map
SFF	Standard Flowgram Format
SNP	Single Nucleotide Polymorphism
TDR	Transmitted Drug Resistance
VCF	Variant Calling File
VL	Viral load
WHO	World Health Organization

CHAPTER 1 - INTRODUCTION

1.1 Global HIV burden and treatment progress

As of 2023, it was estimated that 39.9 million individuals were living with human immunodeficiency virus (HIV) globally, with approximately 30.7 million receiving antiretroviral therapy (ART) (UNAIDS, 2024a). Eastern and Southern Africa bears the highest burden, accounting for 65% of infections globally, or 20.8 million people living with HIV. Within this region, South Africa has the largest epidemic, with an estimated 7.8 million people living with HIV by the end of 2023; about 75% of these individuals, including both adults and children, were receiving ART (UNAIDS, 2024b). Although lifelong ART has transformed HIV management into a chronic condition, no approved vaccine for prevention or cure currently exists.

1.2 Aetiology and Genetic Diversity of HIV

HIV is classified under the Lentivirus genus, which is a subgroup of the Retroviridae family and consists of two main subgroups: HIV-1 and HIV-2 (Sharp and Hahn, 2011). HIV-1, responsible for the global pandemic, originated from zoonotic transfer of SIV_{cpz} in central African chimpanzees, while HIV-2 evolved from SIV_{sm} in West African sooty mangabeys. HIV-1 is further categorized into four groups: M, N, O, and P. Group M, which accounts for around 90% of HIV cases globally, can be subdivided into nine subtypes A, B, C, D, F, G, H, J, and K, along with over 157 circulating recombinant forms (CRF) and various unique recombinant forms (URF) (Sharp and Hahn, 2011; Hemelaar *et al.*, 2019; Foley *et al.*, 2024). Subtype distribution varies across regions, with subtypes A and D prevalent in East Africa, subtype B in Europe and the Americas, and subtype C in Southern Africa, India, and other Asian countries (Figure 1.1) (Hemelaar *et al.*, 2019; Ndashimye, Reyes and Arts, 2023).

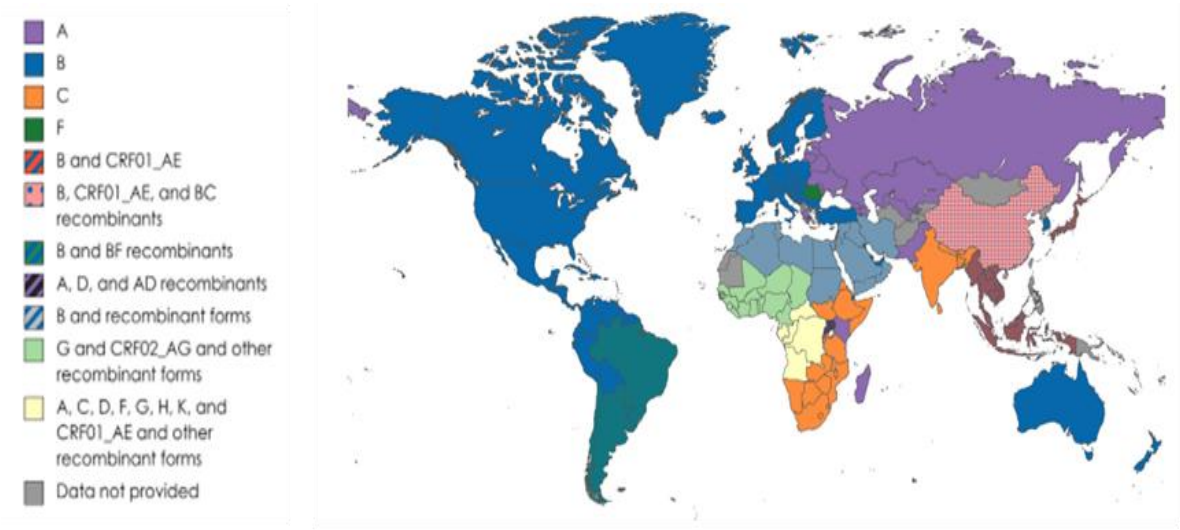


Figure 1.1. ***Global overview of HIV-1 subtypes*** *Map illustrating the geographical distribution of major HIV-1 subtypes (A–K) and circulating recombinant forms (CRFs) (Ndashimye, Reyes and Arts, 2023).*

1.3 Genetic Composition and Functions of the HIV Genome

The HIV genome is a positive-sense, single-stranded RNA approximately 9-10 kilobases (kb) in length (Li and Clercq, 2016).It contains nine major genes responsible for encoding viral proteins and enzymes essential for replication and pathogenicity. These genes include *gag* (p17 matrix, p24 capsid and p7 nucleoproteins), *pol* (p66 reverse transcriptase, p32 integrase and p11 protease enzymes), *env* (envelope glycoproteins gp120 and gp41), *tat*, *rev*, *vif*, *vpr*, *vpu*, and *nef* (Figure 1.2 (A)) (Musumeci, Riccardi and Montesarchio, 2015). Of these genes, *pol* is the most relevant for antiretroviral therapy and drug resistance testing, as it encodes the key viral enzymes targeted by antiretroviral drugs (Clutter et al., 2016; Shafer, 2002).

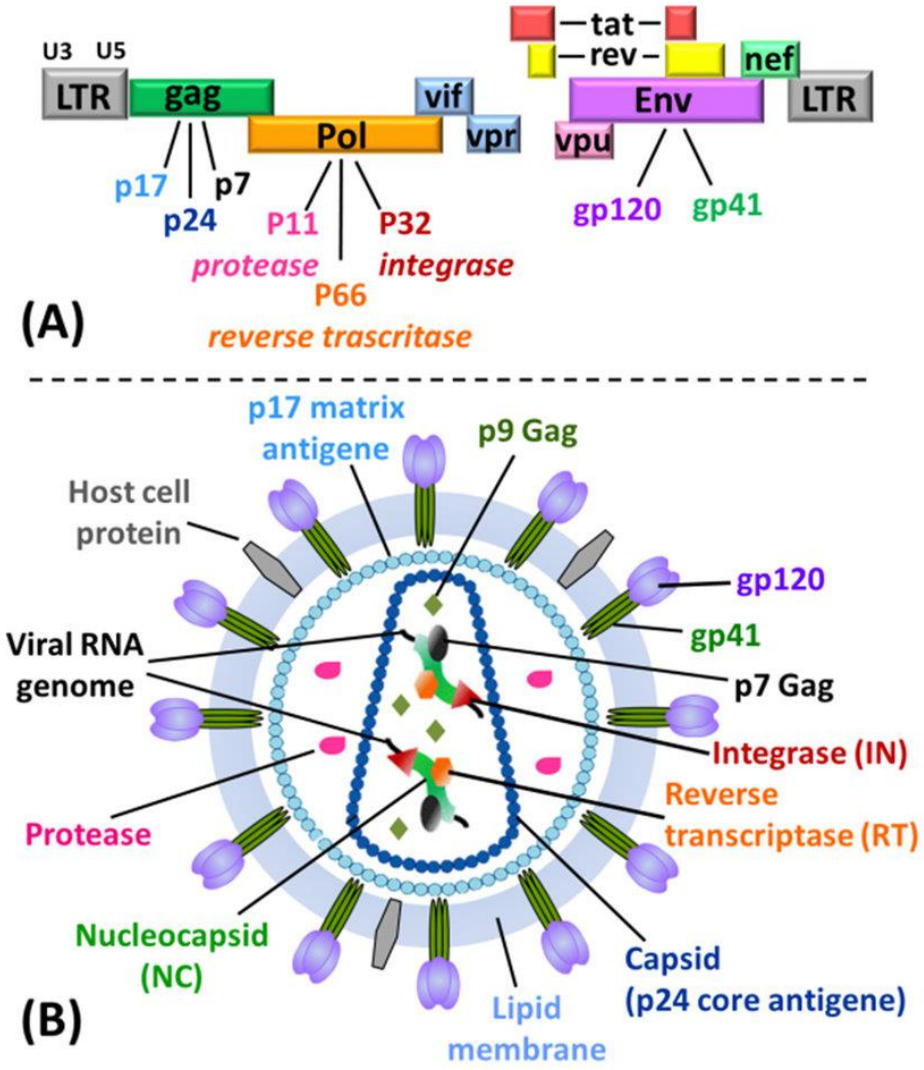


*Figure 1.2.**Schematic representation of the HIV-1 genome and virion structure.** (A) Genome organisation showing the major structural, enzymatic, and regulatory genes. (B) Virion structure depicting the envelope glycoproteins, capsid, and associated enzymes) (Musumeci, Riccardi and Montesarchio, 2015)*

1.4 HIV Structure

The HIV-1 virion is encased in a host-derived lipid bilayer membrane with trimeric envelope glycoproteins (gp120 and gp41), which mediate host cell entry (Figure 1.2B; Musumeci, Riccardi and Montesarchio, 2015). Within the virion are two copies of linear, single-stranded RNA (approximately 10 kb), along with essential enzymes including protease (PR), reverse transcriptase (RT) and integrase (IN). These components are packaged within a cone-shaped capsid composed of p24 and further supported by an underlying matrix layer (p17), which stabilizes the virion and facilitates assembly. Additional accessory proteins like *vpr, vpu,* and *vif* regulate nuclear import, enhance

replication, mediate CD4 degradation, support virion release, and contribute to viral pathogenesis.

1.5 HIV Life Cycle

HIV undergoes a multistep replication cycle that begins with attachment to CD4+ receptors on host T lymphocytes, followed by binding to either CCR5 or CXCR4 co-receptors (Woodham *et al.*, 2016). This interaction facilitates fusion of the viral envelope with the host cell membrane, allowing entry of the viral core. Inside the host cell, reverse transcriptase (RT) synthesizes complementary DNA (cDNA) from the viral RNA genome. The resulting cDNA is integrated into the host genome by the IN enzyme, establishing a proviral state (Masenga *et al.*, 2023).

Following integration, the virus uses the host cell's transcriptional and translational machinery to produce viral RNA and proteins during infection (Maertens, Engelman and Cherepanov, 2022). During assembly, the protease (PR) enzyme cleaves polyprotein precursors into functional units required for forming new virions. These components are directed to the cell membrane, where they assemble and bud from the host cell, acquiring a lipid envelope in the process. The mature virions are then released into circulation and can infect other CD4+ T cells or immune system targets (Woodham *et al.*, 2016).

RT, IN, and PR are essential for HIV replication and are the primary targets of antiretroviral therapy (Maertens, Engelman and Cherepanov, 2022). Inhibiting these enzymes disrupts key stages of the viral life cycle, forming the basis of effective HIV treatment strategies.

1.6 HIV natural history and transmission

Following initial infection, the virus rapidly disseminates during the acute phase, sometimes causing flu-like symptoms (Gougeon *et al.*, 1998; Woodham *et al.*, 2016). HIV then enters a clinical latent phase with low-level replication without notable symptoms, but it continues to damage the immune system and deplete CD4+ T cells. Over time, HIV progressively impairs the immune system by directly infecting and killing CD4+ T cells, as well as triggering chronic immune activation and inflammation. This immune dysfunction weakens the immune response, leaving individuals vulnerable to opportunistic infections, certain cancers, and other complications. In the

absence of treatment, HIV progressively reduces CD4+ T cell levels, ultimately causing severe immunosuppression and the onset of acquired immunodeficiency syndrome (AIDS). AIDS is characterized by life-threatening opportunistic infections and specific cancers (Gougeon *et al.*, 1998).

HIV transmission occurs through unprotected sexual contact, contact with infected blood or contaminated needles, and from mother to child during pregnancy, delivery, or breastfeeding (Woodham *et al.*, 2016).

1.7 Antiretroviral Therapy (ART)

ART aims to suppress the replication of the HIV virus by targeting key viral enzymes (PR, RT and IN), as well as other proteins involved in the HIV replication cycle (Clutter *et al.*, 2016). Effective ART halts the progression to AIDS, significantly lowering rates of illness and death, while enhancing the overall quality of life for individuals living with HIV (Clutter *et al.*, 2016; Peng *et al.*, 2023). Despite its effectiveness, ART is unable to eliminate the latent reservoir formed early during infection (Chou, Maggirwar and Marsden, 2024).

Currently, there are over 30 antiretroviral drugs available, categorized into six classes based on their mode of action: nucleoside or nucleotide reverse transcriptase inhibitors (NRTI), non-nucleoside reverse transcriptase inhibitors (NNRTI), protease inhibitors (PI), integrase strand transfer inhibitors (INSTI), CCR5 antagonists, and fusion inhibitors (Temereanca and Ruta, 2023). Recent developments include new drug types such as attachment inhibitors and capsid inhibitors, as well as longer-acting agents to address challenges like drug resistance and improve adherence.

1.7.1 Nucleos(t)ide Reverse Transcriptase Inhibitors (NRTI)

Nucleoside or nucleotide reverse transcriptase inhibitors (NRTIs) are synthetic nucleoside and nucleotide mimetics that compete with natural nucleotides during reverse transcription, and become incorporated into the growing DNA chain, thereby inhibiting viral replication. (Sever et al., 2024). This class of drugs includes agents such as zidovudine, tenofovir, emtricitabine, lamivudine, and abacavir. Initially introduced as standalone treatments in the early 1990s, NRTIs are now primarily used as part of combination antiretroviral therapy, typically serving as the foundational component (‘backbone’) alongside agents from other antiretroviral classes.

1.7.2 Non-nucleoside reverse transcriptase inhibitors (NNRTI)

NNRTIs interact with a hydrophobic pocket adjacent to the reverse transcriptase active site, inducing structural changes that impair the enzyme's function (Sever *et al.*, 2024). Examples of NNRTIs include efavirenz and nevirapine, while second-generation agents consist of etravirine, rilpivirine, and doravirine. Although NNRTIs are effective in suppressing viral replication, they are prone to the rapid development of resistance due to their low genetic barrier.

1.7.3 Protease Inhibitors (PI)

Protease inhibitors inhibit the enzymatic activity of HIV protease, thereby blocking the cleavage of viral polyproteins essential for producing mature, infectious virions (Sever *et al.*, 2024). Commonly used PIs include darunavir, atazanavir, lopinavir, and ritonavir.

1.7.4 Integrase Inhibitors (INI) / Integrase strand transfer inhibitors (INSTI)

Integrase strand transfer inhibitors act by targeting the active site of the integrase enzyme, thereby blocking the integration of viral DNA into the host genome (Sever *et al.*, 2024). Current integrase inhibitors in clinical use include raltegravir, elvitegravir, dolutegravir (DTG), cabotegravir, and bictegravir. These agents are characterized by strong antiviral potency, a high genetic barrier to resistance and favourable tolerability profiles.

Non-nucleoside reverse transcriptase inhibitors are particularly prone to resistance due to their low genetic barrier, meaning that a single point mutation can significantly reduce drug susceptibility. In contrast, NRTIs and INSTIs generally require multiple mutations to confer high-level resistance, providing a higher genetic barrier. Protease inhibitors, while potent, also exhibit a moderate to high barrier to resistance, but their use is often limited by tolerability and drug–drug interactions (Clutter et al., 2016; Parikh et al., 2017).

1.8 Evolution of Treatment Guidelines

The World Health Organization (WHO) has issued treatment guidelines for individuals diagnosed with HIV, offering standardized, evidence-based recommendations for

initial and salvage ART regimens (World Health Organization, 2021a). From the early 2000s until 2018, first-line regimens included a combination of two NRTI plus an NNRTI, and substituting the NNRTI with a PI for second-line or salvage regimens. More recently, the formulation of fixed-dose combinations allowed for improved adherence, reduced pill burden, and simplified supply chain and procurement strategies. In 2018, WHO recommended the implementation of DTG-based regimens, specifically the fixed-dose combination of tenofovir, lamivudine, and dolutegravir, as the preferred first-line treatment due to DTG's higher efficacy, reduced side effects, and lower likelihood of resistance development (World Health Organization, 2018). This transition was reinforced in 2019, when updated guidelines urged the widespread adoption of DTG-based regimens, including for pregnant women and those of childbearing potential, following the resolution of initial safety concerns (World Health Organization, 2019). In cases where the first-line regimen fails, as defined by detectable HIV viral load (HIVVL), a second-line regimen is introduced which incorporates a ritonavir-boosted PI to enhance treatment response.

The 2019 revision of South Africa's National HIV guidelines integrated recommendations from the World Health Organization, reflecting important advances in HIV management and therapeutic approaches (South African National Department of Health, 2019). These guidelines endorse the use of DTG-based regimens as the preferred first-line treatment for individuals initiating ART.

1.9 Emerging HIV Drug Resistance

The increased global access to ART has inadvertently driven the evolution of resistance to ART, both at the individual patient level and across public health systems worldwide, posing a significant challenge to achieving optimal treatment outcomes (Shafer, 2002). Suboptimal drug levels facilitate the rapid selection of resistance-associated viral mutations, which can compromise treatment effectiveness and lead to virologic failure. Poor or inconsistent adherence to ART is a major contributor to acquired drug resistance, as intermittent drug exposure allows viral replication in the presence of subtherapeutic drug concentrations, promoting the emergence of resistant variants (Bennett et al., 2009; Clutter et al., 2016). Additionally, the high genetic diversity of HIV and rapid replication rate introduces additional complexities in formulating treatment strategies that can effectively combat this virus (Shafer, 2002; Clutter *et al.*, 2016).

Drug resistance in HIV can be categorized into two main types: acquired drug resistance (ADR) and transmitted drug resistance (TDR) (Clutter *et al.*, 2016):

- Acquired Drug Resistance (ADR): ADR arises during ART as a consequence of viral adaptation to selective drug pressure, resulting in the emergence of resistant HIV variants within treated individuals. This commonly results from inadequate treatment adherence, pharmacological interactions, or the use of suboptimal therapeutic regimens. As the virus replicates in the presence of these drugs, mutations can emerge that confer resistance, leading to treatment failure and disease progression.
- Transmitted Drug Resistance (TDR): TDR refers to drug-resistant strains of HIV that are passed from one individual to another, typically through unprotected sexual contact or sharing of needles. TDR complicates treatment for newly diagnosed individuals, who may begin therapy with a pre-existing resistance profile. The prevalence of TDR varies by region and population, influenced by local ART practices and the epidemiology of HIV transmission.

The growing presence of drug-resistant HIV strains in resource-limited environments poses a substantial risk to the sustainability of ART programs and epidemic control (World Health Organization, 2021b) . In many resource-limited settings, systemic challenges can undermine the long-term success of antiretroviral therapy programmes. Limited access to high-resistance-barrier regimens, inconsistent drug supply chains, gaps in clinical oversight, and reduced frequency of viral load testing contribute to environments where resistant strains of HIV are more likely to emerge and circulate. These vulnerabilities are further compounded by patient-level factors such as irregular follow-up and challenges with continuity of care (World Health Organization, 2021b). To mitigate the continued rise and spread of HIV drug resistance (HIVDR) in resource-limited settings, the World Health Organization recommends adopting structured surveillance systems designed to detect and respond to health system-level contributors that facilitate the development and transmission of resistance.

1.9.1 Prevalence and Trends of ADR in HIV

Acquired drug resistance trends have been documented through various approaches, including clinical trials, cohort monitoring, and national surveillance efforts (Boender *et al.*, 2015). An aggregated analysis of global studies conducted between January 2006

and May 2013, examining long-term outcomes of NNRTI-based antiretroviral therapy in low- and middle-income countries, found that approximately 85% of patients maintained virological suppression for up to five years. However, among those who did not achieve viral suppression, resistance to NRTIs, NNRTIs, or both drug classes were observed in 70% to 90% of cases.

The World Health Organization's 2021 HIV Drug Resistance Report highlighted concerning trends in acquired drug resistance across 29 resource-limited countries. Viral load suppression rates among individuals on treatment ranged widely from 63% to 99%. Among those experiencing virological failure (defined as viral load ≥1,000 copies/mL), the proportion with ADR to NNRTIs was found in 50% to 97% of cases, while ADR to NRTI was observed in 16.8% to 82.8% of patients (World Health Organization, 2021b). Population surveys conducted within South Africa have revealed ADR rates reaching 75.6% among affected individuals (Moyo *et al.*, 2020). Levels of ADR to dolutegravir-based regimens are reported as 3.9 – 19.6 %% in countries implementing routine surveillance activities between 2021 and 2022 (World Health Organization, 2024).

1.9.2 Surveillance of ADR in South Africa

In 2019, the National Institute for Communicable Diseases (NICD) implemented a national HIVDR surveillance protocol aimed at assessing the prevalence of resistance among adults receiving ART who experienced virological failure (defined as a viral load ≥1,000 copies/mL), using remnant plasma samples collected during routine viral load monitoring. This initiative aligns with similar regional surveillance efforts across sub-Saharan Africa, where next-generation sequencing (NGS) technologies are increasingly being adopted to inform treatment strategies and support public health decision-making in high-burden settings (Steegen et al., 2024).

As part of a national surveillance initiative, remnant plasma samples from routine HIV viral load tests were systematically selected across 16 National Health Laboratory Service (NHLS) sites. Out of 779 specimens earmarked for HIVDR analysis, genotyping was successfully conducted on 762 samples using next-generation sequencing (NGS) of the *pol* gene region. After applying inclusion criteria, 753 of these were retained for final analysis. Overall, 72.1% (95% CI: 66.8%–76.9%) of the samples exhibited resistance to at least one antiretroviral drug class. Non-nucleoside reverse

transcriptase inhibitor-associated resistance mutations were identified in 70.5% (95% CI: 64.7%–75.7%) of cases, while NRTI resistance was present in 49.0% (95% CI: 44.7%–53.3%). In contrast, resistance to protease inhibitors (PIs) was uncommon, detected in only 2.2 % of the specimens (95% CI: 1.3%–3.5%). The most prevalent mutations, defined as those detected at a consensus threshold of ≥20.00%, were K103N, M184V, V106M, K65R, K70R, P225H, K219E and G190A (Figure 1.3) all located within the reverse transcriptase (RT) region of the pol gene.

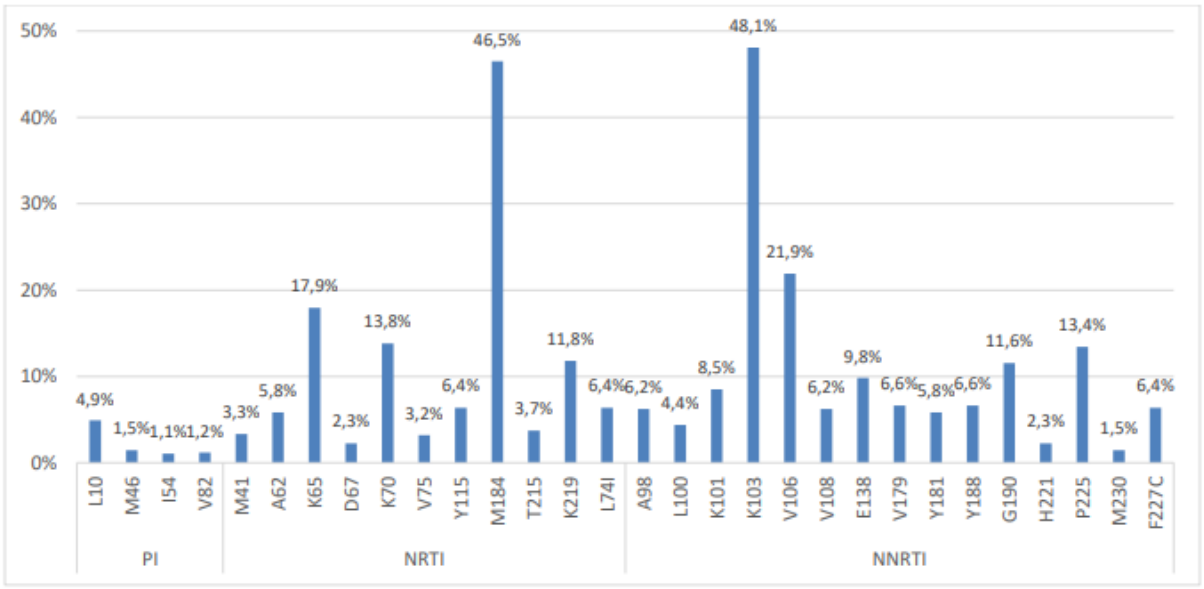


Figure 1.3. ***Prevalence of HIV-1 drug-resistance mutations among adults with viraemia, South Africa (2019).*** *Data derived from the national HIVDR surveillance programme) (Steegen et al., 2024).*

The clinical significance of these eight mutations and associated nucleic acid changes are summarised in Table 1.1 below. These mutations occur within the reverse transcriptase (RT) region of the HIV-1 pol gene and are primarily associated with resistance to nucleoside (NRTI) and non-nucleoside (NNRTI) reverse transcriptase inhibitors (Clutter et al., 2016; Tang, Liu and Shafer, 2012).

Table 1.1.Clinical significance of most frequently detected mutations (K103N, M184V, V106M, K65R, K70R, P225H, K219E and G190A) within this dataset Extracted from https://hivdb.stanford.edu/. All listed mutations are located within the reverse transcriptase (RT) region of the pol gene and associated with resistance to nucleoside (NRTI) and/or non-nucleoside (NNRTI) reverse transcriptase inhibitors. (Clutter et al., 2016; Tang, Liu and Shafer, 2012)

Mutation	Clinical relevance
K103N	K103N is a nonpolymorphic mutation conferring resistance to NNRTIs, commonly selected in individuals treated with nevirapine and efavirenz. This mutation decreases susceptibility to nevirapine by approximately 50-fold and to efavirenz by about 20-fold. K103N is also recognized as the most

	frequently transmitted drug resistance mutation.
M184V	M184V/I mutations are primarily selected by lamivudine and emtricitabine, causing a greater than 200-fold decrease in susceptibility to these drugs. They also reduce susceptibility to abacavir by approximately threefold. Conversely, M184V/I mutations increase susceptibility to zidovudine and tenofovir by about twofold and slow the development of resistance to these medications. Additionally, M184V/I are linked to diminished viral replication both in vitro and in vivo.
V106M	V106M is a nonpolymorphic mutation mainly induced by exposure to efavirenz and nevirapine. It is observed more frequently in subtype C viruses due to requiring only a single nucleotide substitution (GTG [Valine] → ATG [Methionine]), whereas other subtypes need two nucleotide changes (GTA [Valine] → ATG [Methionine]). This mutation leads to more than a 30-fold decrease in susceptibility to both nevirapine and efavirenz.
K65R	K65R is a mutation selected primarily by tenofovir and abacavir, and less frequently by lamivudine. It reduces susceptibility to tenofovir and abacavir by approximately two-fold according to the PhenoSense assay, with slightly higher reductions reported by other testing methods. Resistance to lamivudine and emtricitabine is more pronounced, with a 5 to 10-fold decrease in susceptibility. Interestingly, K65R increases susceptibility to zidovudine, except when found alongside the Q151M mutation. Due to a bidirectional antagonistic effect, K65R rarely co-occurs with thymidine analogue mutations. K65R is more commonly observed in HIV subtype C viruses. Biochemical analyses indicate that a unique sequence context in subtype C—specifically, a series of five consecutive adenosines preceding the second position of the K65 codon—makes this region more prone to mutation during reverse transcription. Experimental studies have shown that subtype C viruses develop K65R more rapidly than subtype B viruses under tenofovir pressure. Retrospective clinical data also suggest that patients infected with subtype C who experience virological failure on tenofovir-containing regimens have a higher likelihood of developing the K65R mutation. Despite these subtype differences, K65R remains one of the most frequent NRTI-associated resistance mutations, second only to M184V, in individuals receiving tenofovir-based therapy. Importantly, no evidence has shown that subtype C infections confer a greater risk of virological failure on tenofovir regimens compared to other subtypes
K70R	K70R is classified as a type 2 thymidine analogue mutation that develops primarily in response to zidovudine and stavudine treatment. This mutation is linked to an approximate five-fold reduction in susceptibility to zidovudine, while it has little to no significant impact on the effectiveness of abacavir and tenofovir.
P225H	**P225H** is a nonpolymorphic accessory mutation commonly selected during efavirenz therapy. It typically co-occurs with the K103N mutation. When present together, K103N and P225H can lead to a more than 50-fold reduction in susceptibility to nevirapine and efavirenz, and a 5- to 10-fold reduction in susceptibility to doravirine.
K219E	Accessory thymidine analogue mutations (TAMs) selected by zidovudine and stavudine. While K219Q/E alone confer minimal resistance, their presence alongside other TAMs contributes to a cumulative reduction in zidovudine susceptibility and may enhance overall NRTI resistance.

G190A	G190A is a nonpolymorphic mutation selected in persons receiving nevirapine and efavirenz. It confers >50-fold reduced nevirapine susceptibility and 5-10-fold reduced efavirenz susceptibility

1.10 HIV Drug Resistance Testing Methods

HIVDR testing approaches provide information on the resistance profile of the virus circulating in a patient and, in combination with other clinical information such as treatment history and HIVVL, allows healthcare providers to tailor antiretroviral therapy to the individual patient. HIVDR testing methods include genotypic and phenotypic resistance testing (Inzaule *et al.*, 2017; Parkin *et al.*, 2023).

Genotypic testing involves analysing the viral genome to detect mutations that confer resistance to antiretroviral drugs (Clutter *et al.*, 2016; Parikh *et al.*, 2017). The most commonly used technique involves amplifying targeted regions of the HIV genome, such as IN, RT, and PR genes, followed by Sanger sequencing for mutation detection. These sequences are then compared against established databases and algorithms to predict the likely efficacy of different drug regimens based on presence of specific genetic mutations (Hirsch *et al.*, 1998; Shafer, 2002). The most used HIVDR interpretation algorithms are the Stanford University HIV Drug Resistance Database HIVdb Program (Liu and Shafer, 2006) Rega 8 (Vercauteren *et al.*, 2013), and Agence Nationale de Recherches sur le SIDA (ANRS) (Singh, 2017). Genotypic HIVDR testing is highly accurate, cost-effective, can be performed relatively quickly and can be performed on the viral RNA when the plasma HIVVL ranges from at least 500 copies/ml (Carr *et al.*, 2023).

Phenotypic testing assesses the functional response of the virus to antiretroviral drugs by measuring the virus's capacity to replicate when exposed to varying drug concentrations (Hirsch *et al.*, 2005). As it is a culture-based method and provides valuable insights into drug resistance, however, it is more resource-intensive and less frequently used in clinical practice compared to genotypic testing.

1.11 Alternative Methods for Genotypic HIVDR Testing

In addition to Sanger sequencing, alternative methods available for genotypic HIVDR testing are available. Some of these methods include:

- Next-generation sequencing (NGS): NGS technologies allow for parallel

sequencing of multiple viral genes or the entire HIV genome. This approach provides a more comprehensive view of the viral population and can detect low-frequency drug resistance mutations (Ávila-Ríos *et al.*, 2020). In addition to its sensitivity for low-frequency mutations, NGS enables the characterization of intra-host viral diversity, allowing for detailed analysis of quasispecies dynamics and the evolution of drug resistance within individuals over time (Ávila-Ríos et al., 2020; Gibson et al., 2014).

- Oligonucleotide ligation assay (OLA): OLA is a hybridization-based technique that uses sequence-specific probes to detect drug resistance mutations (Panpradist *et al.*, 2019). It is a sensitive and specific method for genotyping drug resistance mutations in HIV.
- Mass spectrometry-based methods: Mass spectrometry approaches, including matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF), have been explored as alternative methods for genotypic HIV drug resistance testing (Lee *et al.*, 2013). These methods detect specific mutations by analysing the mass differences between wild-type and mutant sequences.
- DNA microarray technology: DNA microarrays consist of immobilized probes that can detect specific drug resistance mutations (Martín *et al.*, 2016). This high-throughput method allows for the parallel assessment of multiple resistance mutations.

1.12 Next-Generation Sequencing (NGS)

Next-generation Sequencing is increasingly adopted as a promising alternative to Sanger sequencing for clinical HIVDR testing, owing to its greater scalability, higher throughput, and enhanced sensitivity for detecting low-abundance mutations (Inzaule *et al.*, 2017). By simultaneously sequencing millions of DNA fragments, NGS can quantify variants within a HIV quasispecies, including those that are not detected when using Sanger sequencing technologies (Gibson *et al.*, 2014; Ávila-Ríos *et al.*, 2020). Additionally, incorporating index sequences enables the multiplexing of samples within a single sequencing run (MacConaill *et al.*, 2018).

Different technologies and instruments are commercially available for NGS sequencing. These include Illumina platforms (Illumina, San Diego, CA, USA), Ion Torrent platforms (Thermo Fisher Scientific, Waltham, MA, USA), BGISEQ (BGI, Cambridge, MA), Nanopore Sequencing (Oxford Nanopore Technologies, United

Kingdom) and PacBio (Pacific Biosciences, CA, USA) (Gu, Miller and Chiu, 2019; Manyana *et al.*, 2021). The distinct chemistries and features of each platform are outlined in Table 1.2.

Table 1.2. Overview of NGS Platforms: Key Features Including Sequencing Chemistries, Read Lengths, Throughput, Accuracy, and Limitations (Gu, Miller and Chiu, 2019; Manyana et al.*, 2021).*

Manufacturer	Platforms	Chemistry	Read Length	Throughput	Accuracy	Key Limitations
Illumina	iSeq100, MiniSeq, MiSeq, NextSeq, NovaSeq	Sequencing by Synthesis, bridge PCR	150-300 bp	High	>99.9%	Short reads
Thermo Fisher	Ion S5, Ion PGM, Ion Protov	Ion semiconductor, emulsion PCR	50-400 bp	Moderate	High	Short reads
Pacific Biosciences	Sequel II, Sequel IIe, Sequel	Single Molecule Real-Time	60,000 bp	Low-Moderate	87-99% (with error correction)	Higher error rate, expensive
BGI	BGISEQ	DNA Nanoball Sequencing	100 bp	High	>99%	Short reads
Oxford Nanopore Technologies	MinIon, GridION, PromethION	Nanopore	>100,000 bp	Low-High (scalable)	92-98%	Lower accuracy, high error rate

The Illumina platform offers integrated systems for genetic variation and biological function analysis across multiple biological domains, including agriculture and medicine (Pereira, Oliveira and Sousa, 2020). Renowned for its scalability and precision, this platform supports various sequencing approaches, including whole-genome, targeted, and RNA sequencing. Illumina sequencing works by fragmenting the sample into smaller pieces, attaching specific adapters to the fragments, and immobilizing them on the surface of a flow cell. Through bridge amplification, clusters of double-stranded DNA are formed, enabling simultaneous amplification of DNA fragments. The sequencing process itself involves a cyclic addition of labelled nucleotides (sequencing by synthesis), which are detected through laser excitation and imaging. Base calling is then performed based on the fluorescent signals captured during each cycle. Despite its high accuracy, the technology has limitations related to

dye activity and spectral overlap.

1.13 Bioinformatics Pipelines HIVDR NGS data interpretation

NGS bioinformatics is the application of computational and analytical techniques to process and analyse data generated by NGS technologies (Pereira, Oliveira and Sousa, 2020). The raw sequencing data obtained from sequencing instruments undergo a series of bioinformatics processes, commonly referred to as a pipeline. Through the bioinformatics pipeline, the raw sequencing data is processed and interpreted, enabling the identification and characterization of genetic variations present in the sample. Next-generation sequencing data processing is typically divided into three stages: base calling and quality control (QC) in primary analysis, read alignment and variant calling in secondary analysis, and functional or clinical interpretation of results in tertiary analysis (Figure 1.4) (Yohe and Thyagarajan, 2017; Pereira, Oliveira and Sousa, 2020).

Although most HIVDR bioinformatics pipelines follow the same three analytical stages; primary, secondary, and tertiary analysis (Ji et al., 2018; Barbitoff et al., 2022; Olson et al., 2022). Each pipeline implements these steps differently based on its underlying algorithms and computational logic. The differences extend beyond quality control processes and reflect the unique assumptions made during data processing. For example, pipelines may use distinct alignment strategies, variant-calling criteria, or statistical models for filtering sequencing noise, which can influence mutation detection outcomes. These variations mean that even when analysing the same dataset, pipelines can yield slightly different results because of differences in algorithm design, parameter settings, and interpretation frameworks.

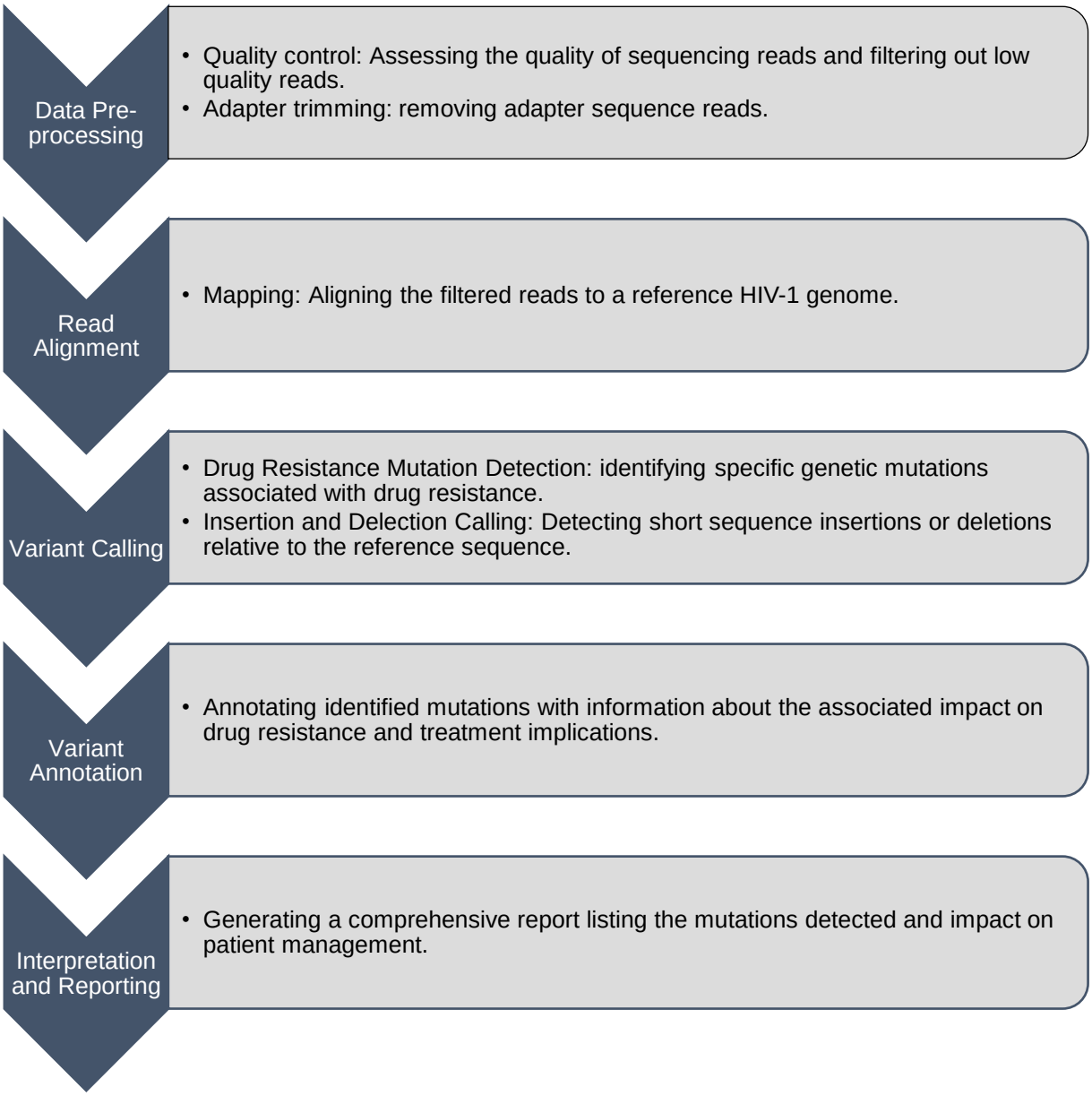


Figure 1.4 ***Workflow for comparative analysis of HIVDR pipelines.*** *Overview of the analytical process comparing six NGS pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, Stanford HIVdb-NGS) for detection of eight HIVDR mutations across 762 samples.*

1.13.1 Primary Analysis

Primary analysis is the initial stage of data processing that focuses on the recognition and evaluation of raw sequencing data generated. (Quail *et al.*, 2012; Pereira, Oliveira and Sousa, 2020). This stage includes tasks such as base calling, quality control (Q-scores), and data format conversion, transforming raw sequencing data into high-quality data suitable for further analysis.

- Base Calling: Base calling is the process of converting raw signal data from NGS into DNA sequence information. For Illumina platforms, this involves analysing

fluorescent signals generated as nucleotides are incorporated into the growing DNA chain using the sequencing-by-synthesis approach (Yohe and Thyagarajan, 2017; Pereira, Oliveira and Sousa, 2020). Challenges include the physical proximity of densely packed DNA clusters, which can cause crosstalk and interference between neighbouring clusters. This requires complex image processing to align images to cluster positions, extract intensity values, and correct for crosstalk. Algorithms such as the Bustard algorithm were initially used to convert these fluorescence signals into DNA sequence data. Recent advancements have introduced more sophisticated algorithms like Real-Time Analysis and Dragen, which significantly reduce base-calling errors, enhancing accuracy and performance in real-time sequencing (Koboldt, 2020).

- Phred Scores: Base-calling accuracy is assessed using Phred scores, which indicate the likelihood of incorrect base calls (Pfeifer, 2017). For example, a Q10 score represents a 10% chance of an error, while a Q30 score indicates a 0.1% chance, correlating to 99.9% accuracy.
- Trimming: The trimming process removes adapter fragments and poor-quality bases from sequencing data to increase the precision of downstream analysis (Chen *et al.*, 2017; Pereira, Oliveira and Sousa, 2020). Tools such as BTrim, AdapterRemoval, and Trimmomatic are used for this process, with the choice of tool depending on dataset characteristics and analysis requirements (Pereira, Oliveira and Sousa, 2020; Kayani *et al.*, 2021; Hakimzadeh *et al.*, 2023).
- Fast-All (FASTA) + Quality (FASTQ) files: The DNA sequence data and quality scores obtained from base calling are combined into FASTQ records. Each record includes a sequence identifier, the DNA sequence, a separator line, and quality scores, reflecting the confidence in each base call (Lo and Chain, 2014). FASTQ files, containing both DNA sequences and quality scores, are essential for downstream applications such as variant discovery and genome assembly. In contrast, FASTA files store only the nucleotide or amino acid sequences in a simplified text-based format without quality information. FASTQ files, containing both sequence and quality data, are essential for downstream applications such as variant discovery and genome assembly.
- Demultiplexing, using barcodes to label DNA fragments, allows for pooling and sequencing multiple samples, which are then sorted and stored as FASTQ files (Hu *et al.*, 2021).
- Quality assessment of raw data is performed using tools like FastQC, which

generates comprehensive reports with graphical representations of various quality aspects. Sequence reads meeting quality thresholds can proceed to alignment or mapping against a reference genome (Chen *et al.*, 2017).

1.13.2 Secondary Analysis

Following primary analysis, secondary analysis involves deeper processing of the data (Pereira, Oliveira and Sousa, 2020). Key tasks include read alignment, variant calling, and quality filtering, with the goal of identifying genetic variations and characterizing genomic features.

- Sequence Alignment: In the secondary analysis of NGS data, sequencing reads are either aligned to a reference genome to facilitate variant detection or assembled de novo to reconstruct genomic sequences without relying on a reference (Oliver, Hart and Klee, 2015). Common alignment tools include Burrows-Wheeler Aligner (BWA), Bowtie-2, HISAT2, STAR, and GSNAP, each catering to different sequencing data types and analysis needs (Bani Baker *et al.*, 2020; Kim, Ji and Yi, 2020; Donato *et al.*, 2021). BWA and Bowtie-2 are preferred for short DNA sequences, while HISAT2 and STAR are specialized for RNA-Seq. Aligning short sequencing reads can be challenging, especially when dealing with repetitive regions and inconsistencies between the sequencing reads and the reference genome.
- Post-Alignment Processing: To enhance the reliability of variant calling, post-alignment steps include the removal of duplicate reads, local realignment around Insertion and Deletion (INDEL) regions, and recalibration of base quality scores (Sezerman *et al.*, 2019). Tools such as SAMtools, Picard, and Genome Analysis Toolkit (GATK) are commonly used for these tasks.
- Variant Calling: The variant calling process enables the detection of sequence-level changes, including single nucleotide polymorphisms (SNP), insertions, deletions, and complex structural variants (Koboldt, 2020). This process involves comparing sequenced reads to a reference genome and determining differences. Tools for variant calling include SAMtools, GATK, Freebayes, and Ion Torrent's Torrent Variant Caller (Lahens *et al.*, 2017; Kanzi *et al.*, 2020). The output, in Variant Call Format (VCF), is a standardized format for storing genetic variant information, enabling further analysis and interpretation (Pabinger *et al.*, 2014).
- Structural Variants: Structural variants include larger-scale genetic alterations such

as deletions, duplications, inversions, insertions, and translocations. Long-read sequencing technologies are increasingly used to identify and characterize these structural variations, which can enhance diagnostic yield for unresolved genetic diseases (Kumar, Cowley and Davis, 2019).

1.13.3 Tertiary Analysis

Tertiary analysis represents the advanced stage of data interpretation following primary and secondary analyses (Pereira, Oliveira and Sousa, 2020). It focuses on integrating and contextualizing genetic variants to extract meaningful biological and clinical insights.

Key tasks in tertiary analysis include (Magi *et al.*, 2010; Oliver, Hart and Klee, 2015):

- Variant Annotation: Annotating variants with information on genomic location, functional impact, disease associations, and population frequency to aid in prioritization and interpretation.
- Variant Prioritization and Interpretation: Filtering and prioritizing variants based on functional impact, conservation, allele frequency, and disease associations to identify potentially pathogenic variants.
- Pathway and Functional Analysis: Assessing the impact of variants on biological pathways and functions, including gene-gene interactions and pathway enrichment.
- Clinical Relevance and Reporting: Evaluating the clinical significance of variants in relation to the patient's clinical presentation and generating reports to guide clinical interpretation.
- Integration with Clinical and Phenotypic Data: Combining genetic variant data with clinical and phenotypic information to understand the genetic basis of diseases or conditions.

1.13.4 NGS file types

Table 1.3 outlines the file types, corresponding extensions, and descriptions associated with each stage of the NGS data analysis pipeline, in particular, primary, secondary, and tertiary analysis.

Table 1.3. Overview of file types used in the primary, secondary and tertiary analysis stages of NGS data.

Stage	File Type	File Extension	Description
Primary Analysis	Raw Data	.bcl	Raw base call files from Illumina sequencers.
	Base Calling Output	.fastq	FASTQ files containing sequence reads and their associated quality scores.
	Quality Control Reports	.html, .png	Reports generated by tools like FastQC to assess the quality of raw sequencing data.
Secondary Analysis	Trimmed FASTQ Files	.fastq	FASTQ files after trimming adapters and low-quality bases. Processed data, where adapters and low-quality regions have been removed for cleaner analysis.
	Aligned Reads	.sam, .bam	SAM (plain text) and BAM (binary) formats for storing aligned sequence data.
	Binary alignment map (BAM) files	.bam	BAM files after post-processing steps like base recalibration and duplicate removal.
	Variant Calls	.vcf	VCF (Variant Call Format) files containing detected variants.
Tertiary Analysis	Annotated Variants	.vcf, .tsv	VCF or TSV files with additional annotations for variants, providing functional information.
	Coverage Data	.bed	BED files summarizing the coverage of specific genomic regions.
Additional Files	Reference Genome	.fasta	FASTA files containing reference genome sequences used for alignment.
	Index Files	.bai, .tbi	Index files for BAM (.bai) and VCF (.tbi), enabling efficient querying of large datasets.
	Demultiplexed FASTQ	.fastq	FASTQ files generated after demultiplexing reads assigned to individual samples based on barcodes. Raw data, separated by sample but unprocessed for quality or adapter removal.
	Summary Reports	.html, .pdf, .csv	Reports summarizing sequencing results or variant findings, often generated at the end of pipelines.

1.14 NGS Bioinformatics Pipelines for HIVDR Data Interpretation

Next-generation sequencing bioinformatics pipelines have been developed specifically for HIVDR data interpretation, including V-Phaser, ShoRah, VirVarSeq, MinVar, V-pipe, Hivmmer, Geno2Pheno, MiCall, HyDRA, PASeq, DeepChek HIV, Smart GeneHIV, Vela Sentosa HIV, Hyrax Exatype (Ji *et al.*, 2018) and, very recently, Quasiflow (Ssekagiri *et al.*, 2022) and HIV-DRIVES (Kanyerezi *et al.*, 2024). V-Phaser, ShoRah, VirVarSeq, MinVar, V-pipe and Hivmmer are freely available but require bioinformatic skills, and DeepChek HIV, Smart GeneHIV, Vela Sentosa HIV and Hyrax Exatype are commercial platforms with potential licensing fees. Geno2Pheno, MiCall, HyDRA and PASeq are freely available and suitable for users with limited bioinformatics experience. Developed by separate research teams, these pipelines employ distinct analytical workflows and produce varied output formats.

The pipelines process raw sequence data typically in FASTQ format, which contain sequence reads along with associated quality scores (Ji *et al.*, 2018). The data management process begins with quality control, which involves removing adapter sequences, discarding low-quality reads, and mapping the remaining reads to a reference genome. Once the sequencing data is processed, the pipelines focus on identifying mutations, especially those linked to drug resistance. The pipelines often rely on curated databases, such as the Stanford HIV Drug Resistance Database (Stanford HIVdb) (Tang, Liu and Shafer, 2012) which compiles known resistance mutations and provides interpretation frameworks.

The outputs of these pipelines vary but generally include (Blassel *et al.*, 2021):

- Consensus sequences: Representing the overall viral sequence present in the sample.
- Variant calls: Listing detected mutations, including low-frequency variants.
- Drug resistance reports: Detailed reports of drug resistance mutations, along with interpretations regarding the susceptibility of the virus to various antiretroviral drugs.

1.15 Differences in NGS pipelines

Differences in HIVDR NGS pipeline reporting have previously been published (Lee *et al.*, 2020). These pipelines vary in ease of use, accessibility, and underlying

assumptions. A lack of standardization complicates cross-study comparisons and may affect treatment decisions. A study by (Lee *et al.*, 2020) reported differences in HIVDR amino acid variant calling when comparing NGS data generated by six clinical laboratories as part of a proficiency testing scheme. Raw NGS data in FASTQ format were provided by each participating lab and analysed using five HIVDR pipelines (HyDRA, MiCall, PASeq, Himmer and DEEPGEN). Assessments of sensitivity, specificity, accuracy and linearity showed reduced sensitivity of detection amongst low abundance variants and proposed a minimum variant threshold of 2% for reliable reporting. A similar study comparing three pipelines and using data generated from 11 virology laboratories in Italy at ≥1.00% threshold showed fair concordance among variants present at frequencies on ≥10% only (Armenia *et al.*, 2024). A study comparing data generated across ten laboratories similarly showed rates of pairwise sequence identity at 99.1%, 98.7%, 98.3% and 97.3% at threshold levels of 20%, 15%, 10% and 5%, respectively (Parkin et al., 2020).

To address this, an international symposium in 2018 led to the Winnipeg consensus, which outlined optimal bioinformatics strategies for the analysis of NGS-based HIVDR data (Ji *et al.*, 2018). Key functionalities were organized into five analytical components: (1) read quality control, (2) alignment and reference mapping, (3) variant calling, (4) HIV drug resistance reporting, and (5) extended data management. A summary of the recommendations from the Winnipeg consensus on bioinformatic strategies for NGS HIVDR data analysis pipelines is provided in Table 1.4 below.

Table 1.4. Quality control settings proposed by the Winnipeg consensus (Ji et al.*, 2018)*

Functional Blocks	Objectives	Consensus on strategies
NGS read quality control/quality assurance	Ensure only high-quality reads are retained for downstream analysis	• Use an average read quality score ≥ 25 • Exclude reads shorter than 75 bases • Perform contamination checks
NGS read alignment and reference mapping	Enable accurate alignment of reads to a reference genome	• Use the HIV-1 HXB-2 reference sequence • Apply a short-read aligner • Analyse the full *pol* gene • Handle insertions and deletions appropriately

HIV variant calling and variant quality control	Ensure reliable detection of nucleotide and amino acid variants	• Conduct QC for nucleotide variant calling • Detect amino acid changes directly from NGS reads (not consensus) • Optional: perform secondary QC for minority variants
HIVDR interpretation and reporting	Accurately interpret and report drug resistance mutations	• Use validated drug resistance interpretation tools and curated databases (e.g., Stanford HIVdb)

Despite these recommendations, inconsistent reporting of DRMs across pipelines remains a concern, particularly in resource-limited settings (Ji *et al.*, 2018). Variations in pipeline parameters, such as consensus thresholds, read depth, quality scores, and variant detection criteria, may significantly affect reproducibility and detection sensitivity.

Given South Africa's high HIV burden and the widespread adoption of NGS technologies in national reference laboratories, there is a pressing need to evaluate the extent of pipeline-specific variability using real-world HIV surveillance data. HIV drug resistance surveillance provides essential data to guide national treatment policies and optimize patient management **(**World Health Organization, 2021; Inzaule et al., 2021). The detection of low-frequency drug resistance variants, made possible through NGS-based testing, has direct clinical relevance as these minority mutations can rapidly expand under selective drug pressure, leading to treatment failure if not identified early (Ávila-Ríos et al., 2020)**.** Therefore, accurate detection and interpretation of both major and minority variants are critical for informing clinical decision-making, improving regimen selection, and supporting effective HIVDR surveillance strategies at both the population and programmatic levels.

1.16 Aims and Objectives

The aim of this project was to investigate differences in HIVDR reporting across six NGS pipelines (PASeq, HyDRA, MiCall, DeepChek HIV, Stanford HIVdb-NGS (beta) and Exatype-NGS) and to establish if further data pre-processing QC recommendations would improve reproducibility across these pipelines.

The study objectives are:

- Perform a comparative analysis of six NGS HIVDR pipelines (PASeq, HyDRA,

MiCall, DeepChek HIV, Stanford HIVdb-NGS (beta), and Exatype) using secondary data from the 2019 South African HIVDR surveillance study, focusing on eight resistance-associated mutations: K65R, K70R, K103N, V106M, M184V, G190A, K219E, and P225H.

- Evaluate the performance of each pipeline under default and customized Winnipeg settings in terms of sensitivity, specificity, accuracy, and linearity.
- Assess the impact of different mutation frequency thresholds (≥1 %, ≥2 %, ≥5 %, and ≥20 %) on concordance and reporting accuracy.

CHAPTER 2 - MATERIALS AND METHODS

A secondary data analysis was performed on the 2019 HIVDR national surveillance study to evaluate six NGS HIVDR pipelines (DeepChek, Exatype, HyDRA, MiCall, PaSeq and Stanford HIVdb-NGS (beta)). While this study primarily involved secondary data analysis, for reference purposes, the following description outlines the wet lab and sequencing activities conducted for the HIVDR 2019 surveillance study.

2.1 National HIVDR surveillance study, 2019

2.1.1 Sample collection

A secondary data analysis was carried out using NGS data from a nationally representative survey data estimating levels of HIVDR amongst South African patients in care and with viraemia in 2019 (Steegen *et al.*, 2024). The survey involved gathering remnant HIVVL samples from 16 HIVVL testing laboratories operated by the National Health Laboratory Services. Sampling was carried out between May and June 2019. A total of 779 samples from adult patients aged ≥18 years and with an unsuppressed HIVVL (defined as ≥1,000 copies/ml) were selected for HIVDR NGS. Testing was performed at the NICD by members of the Centre for HIV and STI HIVDR lab team, with the researcher participating in the testing process.

2.1.2 Total Nucleic acid extraction

HIV-1 RNA was isolated from 500µL of plasma samples with the MagNA Pure 96 extraction system and the MagNA Pure 96 DNA and Viral NA Large Volume Kit adhering to manufacturer instructions (Roche Diagnostics, Mannheim, Germany). The resulting eluted total nucleic acid was collected within 96 well plates.

2.1.3 Reverse Transcription-Polymerase Chain Reaction (RT-PCR) Amplification and amplicon clean-up

The Applied Biosystems HIV-1 Genotyping Amplification Module Kit (Thermo Scientific, MA, USA) was used to amplify specific regions of the HIV genome. The amplification targeted codons 6-99 of the PR gene and codons 1-250 of the RT region within the HIV-pol gene, resulting in a 1.1 kb amplicon. To assess amplification, PCR products were loaded onto a 1% agarose gel containing ethidium bromide and viewed under UV light. This was to verify their size, integrity, and absence of contamination.

Successful PCR products with a clear band size of 1.1 kb underwent purification using AMPure XP beads (Beckman Coulter, CA, USA).

2.1.4 Library preparation and sequencing

Concentrations of PCR products were quantified for each sample using the dsDNA HS assay kit and Qubit® fluorometer (Thermo Scientific, MA, USA), and adjusted to a final concentration of 0.2 ng/µl. Library preparation was performed using the 96-sample Nextera® XT DNA Library Preparation Kit (Illumina, CA, USA), following the manufacturer's protocol for tagmentation, indexing, purification, and normalization. The libraries were then normalized to 2nM, pooled in equal volumes and diluted 30-fold using the supplied hybridization buffer. To enhance quality control and sequence diversity, a 20% spike-in of a 20pM PhiX control library (Illumina, CA, USA) was included in the final library pool. A PhiX control library was included to increase base diversity and assist with error calibration during Illumina sequencing, supporting accurate cluster identification and base calling, particularly for amplicon-based HIV sequencing where sequence diversity is low. The samples were then sequenced on the MiSeq platform using the v3 600-cycle MiSeq Reagent Kit v3 (600 cycles). A total of 11 96-well plates were sequenced for this study, with each plate containing between 80 and 90-pooled samples.

2.1.5 Data management and quality control

Raw sequence data from Illumina MiSeq machines were generated as paired-end FASTQ files for each sample.

To ensure the reliability and accuracy of sequencing results, quality control processes were implemented using the Sequencing Analysis Viewer (SAV) tool designed to visualize and assess critical metrics indicative of run performance and data quality on the MiSeq platform.

The key steps and measures in this process are as follows:

i. Data Extraction and Setup.
   After sequencing completion, raw data are transferred from the sequencing instrument to a designated analysis computer. The files processed include:
   - RunInfo.xml: Metadata about the sequencing run.

- RunParameters.xml: Detailed parameters used in the sequencing.
- InterOp folder: Contains various performance metrics.

ii. Visualization and Metric Analysis.

In the SAV environment, the following metrics are evaluated:

- % Base: Percentage of bases successfully called at each sequencing cycle, indicating overall accuracy.
- %Q≥30: Proportion of bases with a quality score exceeding 30, reflecting confidence in base calls.
- Error Rate: Calculated using the PhiX control, an acceptable error rate is typically <1.00%.
- Flow Cell Chart: Visual representation of performance, including:
  - Intensity: Signal strength of sequencing clusters; low signals may indicate issues with chemistry or sample quality.
  - Full Width at Half Maximum: Measures cluster size; variations may suggest problems in cluster generation.
  - Signal-to-Noise Ratio: High ratios correlate with enhanced data quality.

iii. Read Size and Yield Evaluation.

Read size is calculated using the formula:

Read Size (MB) = (% Reads Identified PF/100) x (Reads PF) x (Read Length/1,000,000)
PF: Passing Filter
MB: Megabytes

A variation of up to 10% in read size and yield is acceptable. Deviations necessitate re-sequencing with fresh DNA.

iv. Passing Criteria for QC.

A sequencing run is deemed acceptable if it meets the following criteria:

- %Q≥30: Ideally ≥80% for high-quality data.
- Error Rate: Should remain below 1%.
- Cluster Density: Must be within the optimal range for the specific instrument.
- % Reads Passing Filter (PF): High percentages indicate successful quality filtering.

v. Review and Final Approval.

Any metrics falling outside acceptable ranges required further action, such as

re-running samples or optimizing preparation protocols. Samples with poor quality scores or high error rates are re-sequenced starting from the primary sample extraction.

All sequencing runs performed passed QC criteria and met internal performance standards required by the NICD sequencing core.

2.2 Comparative analysis of study sequencing data using six different pipelines

FASTQ files generated from the 11 sequencing runs constituting 762 individual samples were uploaded into each of the six pipelines assessed as per pipeline requirements. The HIVDR analysis was performed using pipeline default settings and repeated using Winnipeg settings. For each analysis, threshold settings were applied as nil, ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%.

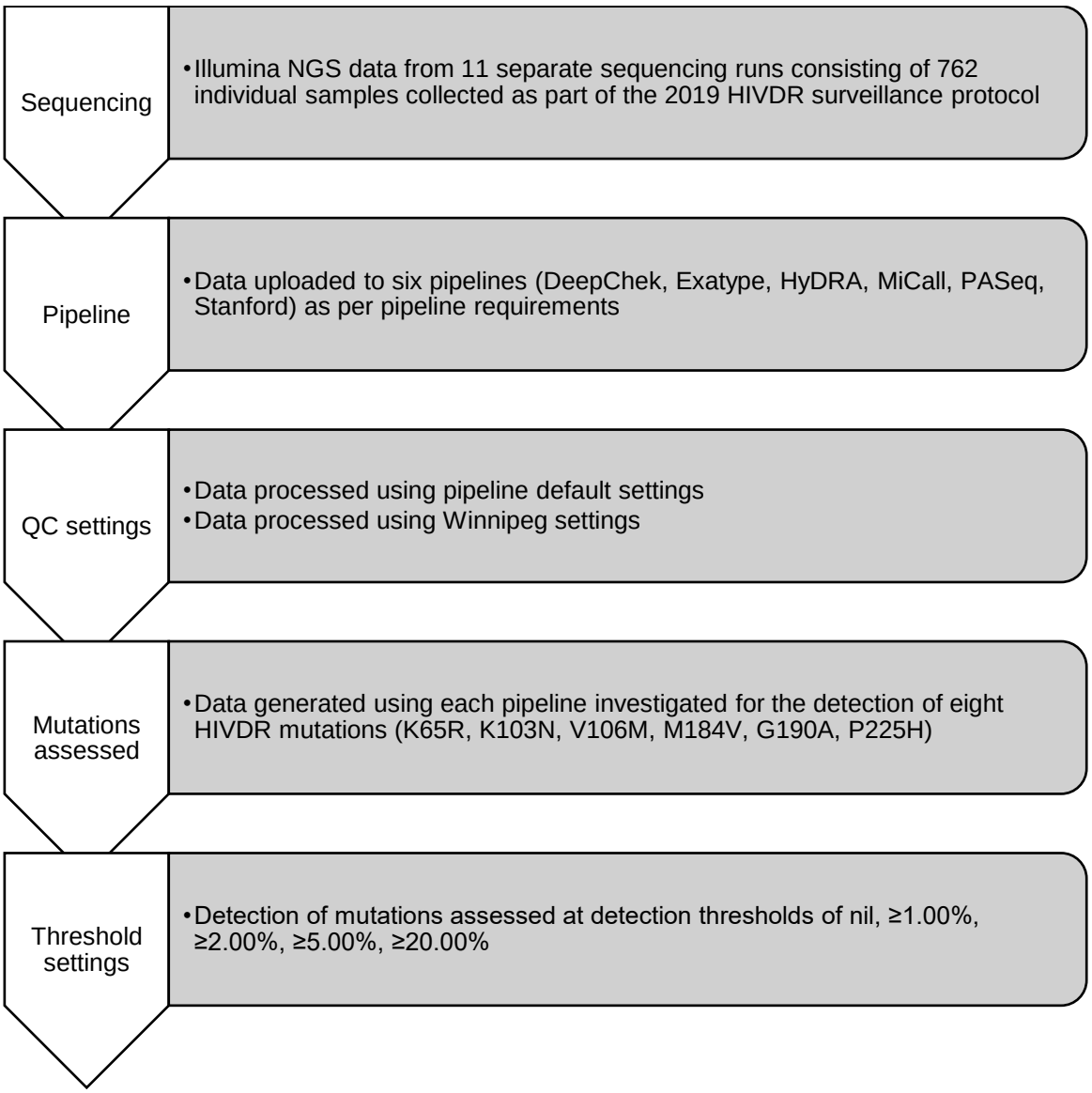


Figure 2.1: ***Workflow for comparative pipeline analysis of HIV-1 drug-resistance data.*** *Overview of the analytical process comparing six NGS pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS) for detection of eight HIVDR mutations across 762 samples under default and Winnipeg settings.*

2.3 Pipeline data management procedures

2.3.1 DeepChek®

DeepChek® is a commercial HIV NGS HIVDR pipeline developed by Advanced Biological Laboratories (ABL, Luxembourg) and is available at https://www.ablsa.com/laboratory-solutions/DeepChek-software/. The software is designed to process both NGS and Sanger sequencing data. DeepChek® primarily focuses on performing comprehensive genomic data analyses, including genotyping, subtyping, detection of amino acid mutations, and characterization of nucleotide

changes. DeepChek® is a web-based software system that is accessed locally via pre-configured servers and does not require additional bioinformatic information technology expertise. The system supports multiple formats for sequencing data, including FASTA and FASTQ formats, and can process NGS data covering targeted genes or whole genome data in binary alignment map (BAM) and sequence alignment map (SAM) formats. Data outputs can be obtained in both comma separated value (CSV) and FASTA formats. The system also incorporates a homopolymer correction algorithm, which effectively reduces sequencing errors across various data sets.

2.3.1.1 Uploading of data files

DeepChek requires for paired-end data files to be upload one sample at a time in fastq.gz format.

2.3.1.2 Configuration settings

The analysis configuration settings were set as default or adjusted to reflect the Winnipeg guidelines for the repeat analysis (Table 1.3). The samples were processed in by uploading paired end FASTQ files, which were stored in the library for any future analysis. The NGS Alignment Engine used the Burrows–Wheeler Aligner (BWA) for Illumina sequencing data

2.3.1.3 Output files

The pipeline generated the following reports:

i. The quality information report provides comprehensive data on the mutations, including prevalence, sequencing coverage, expert system validation, and associated quality scores in Microsoft Excel format, allowing for easy data analysis and manipulation. This report was generated from data in nucleotide format.
ii. The quality information report generated from amino acid format provides data on amino mutations, prevalence, significance according to HIVdb, sequencing coverage, expert system validation, quality scores, and confirmation using Sanger sequencing in Microsoft Excel format. This worksheet was used to extract relevant data for the study.
iii. DeepChek® Clinical summary report provided a concise overview of the

analysis results and relevant clinical information. It was available in both PDF and TIFF formats, ensuring accessibility and compatibility with various devices and software.

2.3.2 Exatype

Exatype is developed by Hyrax Biosciences, Cape Town, South Africa (https://exatype.com) and offers a genotyping solution for SARS-CoV-2 and HIV. It supports both Sanger and NGS sequencing platforms, specifically Illumina and Ion Torrent for NGS analysis. Exatype operates on the purchasing of credits and allows for direct upload of sequencing data from instruments and integration with LIMS/LIS systems. The software has capabilities to differentiate sequencing errors, including homopolymer errors, from actual mutations. This enables seamless analysis of large NGS plates and accurate mutation calling even at a low prevalence of 1% with a CSV or FASTA output.

2.3.2.1 Uploading of data

Sequence data with the extension fastq.gz files were uploaded to the Exatype platform. As paired-end sequencing had been performed, two files were required per sample.

2.3.2.2 Reference sequence

In 2021, Exatype transitioned from utilizing the HXB2 reference sequence to the Los Alamos National Library (LANL) Subtype B pol reference sequence (2004) as a reference sequence for alignment. The use of subtype B references, including HXB2, remains consistent with standard practices for HIVDR analysis and reporting (Ji et al., 2018). The impact on alignment accuracy for subtype C sequences, which predominate in South Africa, is minimal, as HXB2-based coordinate systems are widely adopted for cross-study comparability and standardized mutation position reporting. Exatype incorporates the Stanford HIV Drug Resistance Algorithm for detecting HIVDR mutations.

2.3.2.3 Prevalence

The amino acid and nucleotide prevalence files contained the following information: the reference name, wild-type, position, observed variant, percentage prevalence of

the observed variant, coverage at that position and total read count. Exatype uses the total read count to calculate the prevalence of deletions, which means that coverage and total read count may sometimes differ. The default minimum prevalence cut-off for HIV (Targeted), which is utilized for drug resistance calls and displayed in the nucleotide and amino acid prevalence files, was set at 15%.

2.3.2.4 Determining the Minimum Coverage Required at a Variant Position

The prevalence cut-off and minimum coverage are related: lower prevalence cut-offs require higher coverage for accurate variant calling. Minimum coverage for variant calling at each position is determined based on a sample size beta function, taking into account the worst-case sequencing error rate observed in high-throughput sequencing data. The required coverage is determined by the variant prevalence cut-off. For example, if the prevalence cut-off was set at 10%, a minimum coverage of 30 sequences was needed. Similarly, if the prevalence cut-off was set at 15%, a minimum coverage of 20 sequences was required.

2.3.2.5 Generation of a consensus sequence

Exatype creates a consensus sequence for each sample analysed. The consensus sequence is generated by selecting the most common nucleotide observed at each position from the mapped sequence reads of the sample. For HIV (Targeted) variants, the default minimum prevalence score for inclusion in the consensus sequence is 60%. For INDELS, the minimum prevalence score was set at 80%.

2.3.2.6 Average quality score (QS) of the read

The algorithms in place conduct quality checks and trimming on the sequence reads. By default, specific minimum quality score thresholds are set for each nucleotide. For Illumina, the threshold is 30.

2.3.2.7 NGS read quality control/quality assurance

Exatype carried out various quality control procedures on sequence reads to eliminate low-quality data and ensure that only high-quality data is utilized for subsequent analysis. These quality control steps include quality trimming and read filtering.

Exatype uses Qtrim for quality trimming. The default settings are as follows:

i. Minimum read length: 50 nucleotides.
ii. Minimum quality threshold: The threshold for both the entire read and individual nucleotides at the 3’ end was 30 for Illumina and for Ion Torrent

After quality trimming, Exatype employs several quality filters during reference mapping and variant calling stages that may lead to the discarding of reads.

i. Read Length: exclusion of reads exceeding 6,500 nucleotides
ii. Mismatch to Reference Sequence: exclusion of reads without a match. Exatype conducts an initial alignment process to determine the reference sequence that corresponds to each read. If a sequence read does not match any of the reference sequences, it is discarded.
iii. Minimum Coverage: excluding reads that fail alignment criteria. In the alignment process, a minimum portion of a read must align to the reference sequence for it to be considered valid. For Illumina and Ion Torrent, this minimum is set at 20%. If a read does not meet this criterion, it is discarded and not included in further analysis.
iv. Low Quality Bases: exclusion and X Calling. Exatype takes into account quality scores on a per-read basis when identifying mutations and other variations. If the quality of an individual nucleotide falls below a certain threshold (30 for Illumina and 25 for Ion Torrent), Exatype discards that nucleotide and labels it as an X. In coding data, if the discarded nucleotide is part of a codon, the resulting amino acid is also called as an X, unless the nucleotide represents a four-fold degenerate position.
v. Alignment score for HIV (Targeted) on Illumina and Ion Torrent platforms is set to 50.

2.3.2.8 Output files

After the analysis with Exatype, the following outputs were generated:

i. Coverage: The coverage results are stored in a file named ".coverage.zip". This file contains information about the sequencing depth and coverage across the genome.
ii. Consensus sequences: The consensus sequences of the analysed samples are stored in a file called "All_reference_relative.consensus.fasta". This file provides

the consensus sequence for each sample, representing the most prevalent sequence at each position in the genome.

iii. Consensus FASTA: The consensus sequences are also available in a .zip folder in FASTA format. This folder contains the consensus sequences for all the samples analysed.

iv. Prevalence: This report provides information about the locus, wild-type sequence, locus position, variant, prevalence, coverage, and read count for each observed variant. This data helps assess the presence and characteristics of variants at specific positions within the analysed samples. Data was extracted from this file for subsequent analysis.

v. Mutation details: The "Mutations.csv" file provides details about the identified mutations, including the positions, the reference base, and the observed variant base.

2.3.3 HyDRA

HyDRA is a Perl-based bioinformatics pipeline, designed specifically for analysing NGS data to detect drug resistance mutations in HIV-1. The pipeline maps sequencing reads to the HIV-1 HXB2 reference genome (GenBank accession no. K03455) using Bowtie2 as the alignment tool. It incorporates quality control measures and variant calling thresholds to identify HIV drug resistance–associated mutations, referencing both the Stanford HIV Drug Resistance Database and the 2009 WHO list for surveillance of transmitted resistance (Bennett *et al.*, 2009). Raw NGS data is uploaded onto HyDRA Web with the option of customizing the analysis parameters. The default analysis parameters are stringent and can be customized to meet different user requirements. HyDRA Web enables users to perform HIV-1 drug resistance analysis on NGS data without the need for advanced computational skills or infrastructure. Upon registration, users can upload their datasets, which remain accessible on the platform’s secure server for 90 days. The results include a comprehensive drug resistance report, consensus sequences, variant information, alignment files (BAM), and mutation summaries—all available for download. The platform’s intuitive interface supports scalable use for global research and surveillance initiatives

2.3.3.1 Uploading of data

The platform accepts input files in FASTQ, compressed FASTQ (GZ), and SFF formats. Users have the flexibility to upload or remove files as needed. Although there is no restriction on the number of files that can be submitted, it is important to consider that adding more files may result in longer processing times.

2.3.3.2 HyDRA Configuration and Filtering Parameters

User-defined parameters could be adjusted via the Advanced Options menu to refine the analysis. One such parameter, Consensus Percent, specifies the minimum allele frequency required for a nucleotide to be included in the final consensus sequence. The default threshold was 20%, mimicking the detection sensitivity of traditional Sanger sequencing; thus, low-frequency variants below this cut-off were excluded from the consensus.

The mutation reporting was primarily based on the Stanford HIV Drug Resistance Database, supplemented with annotations from the WHO's 2009 list of Surveillance Drug Resistance Mutations.

Target Coverage determines the average read depth per base used in the final alignment. By default, this was set to 10,000 reads to balance computational speed with sensitivity to minority variants. To achieve this, reads were randomly down-sampled after quality and length filtering but before alignment. Setting this value to -1 disables down-sampling, allowing all reads to be retained in the analysis.

Under the "Filtering" section, options were available to filter out MiSeq reads that did not meet specified criteria:

i. For reads, a length cut-off was set to filter out reads shorter than the specified length. The default was 100 nucleotides. Additionally, a score cut-off was set to filter out reads with an average quality score lower than the specified score. The default cut-off was 30, which represented a Phred-like quality score.
ii. For variants, an error rate was set based on the specific sequencing platform. The default was set to the Illumina MiSeq platform error rate of 0.0021 (0.21%). Minimum variant quality (QS 30), minimum read depth (100bp), and minimum allele count (5%) were also specified to control the quality of variant calls.

iii. For mutations, a minimum amino acid frequency was set to determine the threshold for considering a mutation in the drug resistance report. The default was 0.01 (1%).
iv. The provided HXB2 pol gene sequence (GenBank #: K03455) served as the reference for mapping.

2.3.3.3 Output files

Exportable results can be accessed for all samples encompassed in the analysis. These results entail a downloadable CSV file that contained a concise report summarizing the HIVDR mutations identified within each sample. The report specifically lists mutations that surpass the designated frequency threshold as defined in the advanced options. For deeper analysis and examination, the CSV file could be opened and viewed in Excel. It covers gene locations (PR, RT and IN), classification (major or minor mutation), surveillance status, wild-type and mutant identities, variant frequency, and coverage depth at those sites. Data was extracted from this file for subsequent analysis.

A FASTA file containing consensus sequences for each analysed sample can also be downloaded. The consensus sequences replicate those obtained through Sanger sequencing, incorporating mixed bases based on the predefined "Consensus Percent" cut off of ≥20.00%.

2.3.4 MiCall

The MiCall software is provided as open source; however, the option to make modifications is available after purchasing a permissive license. This allows the running of the modified version of the program on a server without the requirement to publish the source code.

MiCall utilizes various third-party components, each governed by its own license. The following is a compilation of these components alongside their respective licenses:

- Requests: Apache 2.0 license
- Python 3: Python 3 license
- Bowtie2, IVA, and Python-Levenshtein: GNU General Public License (GPL)
- Matplotlib: Matplotlib license

- Reportlab: BSD license
- Pyyaml and Cutadapt: MIT license

These tools were deployed through the installation and set-up of Docker, by pulling and running the following command: sudo docker run cfelab/micall:vX.Y.

2.3.4.1 Uploading data and output files

Table 2.1 highlights certain steps in in the MiCall pipeline, indicating the input files required and the corresponding output files generated at each step. The pipeline encompasses data pre-processing, alignment, mapping, variant calling, and resistance prediction.

Table 2.1 Summary of Steps in the MiCall Pipeline with Input and Output Files. Available from https://cfe-lab/github.io/MiCall/steps

Step Name	Purpose	Input Files	Output Files
summarize_run	Extracts PhiX control error rates and read length information from sequencing run data	InterOp folder, RunInfo.xml or BaseSpace parameters	quality.csv (error rates grouped by tile)
filter_quality	Evaluates PhiX quality metrics to flag low-quality tiles and cycles	quality.csv	bad_cycles.csv (rejected cycles and tiles)
trim_fastqs	Applies trimming based on quality filter and removes adapter sequences	Paired FASTQ files, bad_cycles.csv	trimmed1.fastq, trimmed2.fastq
prelim_map	Performs an initial mapping of reads to multiple HIV reference sequences	Trimmed FASTQ reads	prelim.csv (SAM-format alignments)
sam2aln	Extracts aligned portions, removes duplicate reads, and detects insertions	remap.csv	aligned.csv, conseq_ins.csv, failed_read.csv, clipping.csv
aln2counts	Counts nucleotides and amino acids at each position and constructs consensus sequences	aligned.csv, g2p_aligned.csv, clipping.csv, conseq_ins.csv	nuc.csv, amino.csv, concordance.csv, conseq.csv, insertions.csv, etc.
coverage_plots	Converts amino acid count data into coverage plots for gene regions	amino.csv	coverage_maps.tar, coverage_scores.csv
cascade_report	Summarizes read retention at each pipeline stage	Internal data from previous modules	Internal QC metrics
resistance	Applies Stanford HIVdb algorithm to	amino.csv, optionally amino_midi.csv	resistance.csv, mutations.csv,

	determine resistance profiles		resistance_fail.csv
genreport	Compiles and formats a printable drug resistance report	resistance.csv, mutations.csv	report.pdf

Docker is a self-contained environment within a computer. A script was provided on the MiCall operating page to set up a Docker environment with all the necessary tools for the execution of MiCall through isolated containers. These containers can run on various operating systems and environments and can also be used in on-premises or local environments. Docker allowed for portability and flexibility in deploying applications across different computing environment.

2.3.4.2 Output files

Table 2.1 outlines the individual components of the MiCall pipeline, detailing the tasks performed at each stage along with the corresponding input and output files. The Mutations.csv file was used for further analysis.

2.3.5 PASeq

PASeq (Polymorphism Analysis by Sequencing; https://www.paseq.org/) is a web service designed to facilitate viral genotyping using NGS technologies and developed at Fundació Privada Institut de Recerca de La Sida-Caixa (“IrsiCaixa”). It specializes in HIV resistance genotyping while also offering the capability to detect amino acid polymorphisms in other viral genomes.

PASeq combines multiple open-source bioinformatics tools with custom and proprietary scripts, operating through a web-based interface connected to internal cloud computing resources. It also includes a built-in link to the Stanford HIV Drug Resistance Database, enabling automated interpretation of drug resistance mutations.

2.3.5.1 Uploading and filtering data

The PASeq algorithm uses multiple tools and steps to analyse NGS paired-end FASTQ data for viral genotyping. These steps include quality control, contamination control, subtyping, alignment, consensus calling, variant calling, interpretation, and report generation. Optional steps such as specific threshold selection, Apolipoprotein B mRNA Editing Catalytic Polypeptide-like (APOBEC) filtering, haplotype reconstruction,

and sequence duplicate removal can also be included. The analysis and report parameters can be configured according to specific requirements. PASeq allows for uploads in smaller batches (10 samples at a time, with a size limit of 1 GB) or larger batches through a Google Drive link.

Table 2.2 Overview of Steps and Tools Used in the PASeq Algorithm for viral genotyping. Available from https://paseq.org/about.php

Step	**Function**
Quality Control	Pipeline uses Trimmomatic software to discard those sequencing reads that are of low quality
Contamination Control	Filtering of non-viral data using BBmap
Reference-Based Read Alignment	To enable accurate variant calling, sequencing reads must be aligned to a reference genome. PASeq supports user selection of subtype-specific HIV-1 reference sequences for this purpose.
Consensus calling to obtain a Sanger-like sequence	The consensus sequence is generated by selecting the predominant nucleotide at each aligned position, producing a result comparable to traditional Sanger sequencing.
Variant Calling	After alignment, each amino acid position is assessed against the reference sequence to identify and call variations.
Variant Interpretation	Based on Stanford-HIVdb in terms of resistance or fitness
Report Generation	Through the web or pdf file

2.3.5.2 Output files

The output files generated by PASeq included the following:

i. Consensus with degenerate positions (FASTA): The consensus sequence with degenerate positions represented by ambiguity codes (e.g. N, R, Y) was provided in FASTA format.
ii. Variant File in Amino Acid Variant Format (AAVF) Format: A file in AAVF was generated, containing information about the detected variants, including their positions and amino acid changes.
iii. Amino acid Variants CSV format: A CSV file was produced, providing detailed information about the amino acid variants detected during the analysis, including their positions, wild-type amino acids, and mutant amino acids. The provided columns contain information about the file, reference sequence, protein, mutation, nucleotide variants, position, frequency, coverage in both

forward and reverse directions, mutation counts, presence of incomplete coverage, total coverage, and potentially a ratio or proportion related to the mutation. Data was extracted from this output file for subsequent analysis.

iv. Reference Guided Alignment - SAM format: A SAM (Sequence Alignment/Map) format file was generated, containing alignment information between the analysed sequences and a reference sequence.

v. Coverage Plot: A PDF file was created, presenting a graphical representation of the sequencing coverage depth across the analysed regions or positions.

vi. Consensus (FASTA): The consensus sequence was provided in FASTA format, representing the most predominant base or amino acid at each position.

vii. Resistance Interpretation (Consensus sequence): Resistance interpretation was performed using the consensus sequencing providing insights into the detection of mutations linked to drug resistance about the presence of mutations associated with drug resistance.

viii. Additionally, the outputs included FASTA files and resistance interpretation at different predefined or user-defined threshold values.

2.3.6 Stanford HIVdb-NGS (beta)

The HIVdb-NGS (beta) platform allows users to submit PR, RT, and IN codon frequency tables (CodFreq Files) or AAVF from the HYDRA pipeline. Results are displayed across eight mutation detection thresholds, allowing users to evaluate the presence of unusual mutations as well as signature APOBEC-related mutations. It is important to note that thresholds with a high number of these mutations may indicate mutations caused by machine errors, PCR errors, or G-to-A hypermutations.

By default, genotypic resistance interpretations are provided at the 20% threshold, but users can choose to view the results at lower thresholds using a drop-down menu. A mutation summary table at multiple thresholds assists users in identifying thresholds that minimize the risk of false-positive mutations. Genotypic resistance interpretations are not provided for thresholds where more than 1% of positions exhibit unusual mutations.

2.3.6.1 Uploading of files

HIVdb-NGS (beta) has a requirement for codon frequency files as the input for

processing. These files were generated within the platform by uploading and thereafter converting FASTQ files to the necessary codon frequency files. CodFreq Files are text files that consist of five columns. These columns include the gene (PR, RT, or IN), amino acid position, coverage, sequenced codon, and the number of reads containing that specific codon. HIVdb-NGS (beta) offers flexibility by accepting AAVF files from third-party NGS pipelines like PASeq or HyDRA. This alternative input format accommodates users who have already processed their data through these pipelines.

2.3.6.2 Output files

i. Sequence read summary: The program output provided a summary of sequence reads for regions of HIV-1 PR, RT, and/or IN based on the uploaded CodFreq file. It included the following information:
   - Read Depth: The median number of sequencing reads covering each nucleotide position, indicating the coverage level across the targeted regions.
   - Subtype Assessment: The program determines the subtype by analyzing the consensus of all reads at a 20% mutation detection threshold. This method of subtyping is similar to that used by the main HIVdb program.
   - Minimal Read Depth: This value could be modified according to a selected preference, which excluded positions with reads below the specified depth. In case positions with reads below the specified depth are identified, the program will issue a warning specifying the number of affected codons. The default minimal read depth used by the program is 1,000.
   - Mutation Detection Threshold: The minimum proportion of reads necessary for a specific amino acid to be included in the report can be specified according to specific requirements or preferences. By default, this threshold is set to 20%.

ii. Multi-Threshold Mutation Summary Table:
   Displays mutation detection thresholds (20%, 10%, 5%, 2%, 1%, 0.5%, 0.2%, and 0.1%). Columns include:
      - Mutation detection threshold
      - Usual mutations (amino acid differences from consensus B sequence, excluding unusual mutations)
      - Unusual mutations (amino acids with prevalence <0.01% in published HIVDB group M sequences)
      - Drug resistance mutations with mutation penalty score

    - Mutations suggestive of G-to-A hypermutation
    - APOBEC-context drug resistance mutations resulting from APOBEC-mediated G-to-A hypermutations)
- Low abundance mutations:
    - An option to view a histogram showing the prevalence of variants found in less than 20% of sequence reads was available.
    - A spreadsheet with low-abundance codons listed in descending order of prevalence in the NGS reads. The spreadsheet contained details such as the gene name, amino acid position, read count at each position, sequenced codon, the count of reads containing that codon, the codon's amino acid translation, the percentage of reads featuring the codon, the prevalence of the amino acid and codon in HIVDB group M Sanger sequences, and annotations indicating whether the mutation was associated with drug resistance, classified as unusual, identified as a signature APOBEC mutation, or an APOBEC-context DRM. Data was extracted from this file for subsequence analysis

2.3.6.3 Gene-specific drug-resistance interpretations

Gene-specific drug-resistance interpretations were provided for mutations that were detected above the mutation detection threshold; subject to the condition that the selected threshold was equal to or greater than 0.5% and that fewer than 1% of positions exhibited an unusual mutation.

Table 2.3 below summarizes the key features and differences between the six pipelines.

Table 2.3 Overview of pipeline features, including coverage, quality thresholds. Pipelines used are DeepChek®, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS (beta)

	DeepChek®	**Exatype**	**HyDRA**	**MiCall**	**PASeq**	**Stanford HIVdb NGS (beta)**
URL	https://www.ablsa.com/overview/DeepChek® /	https://exatype.com/	https://HyDRA.canada.ca	https://github.com/cfe-labMiCall	https://www.paseq.org	https://hivdb.stanford.edu/hivdb/by-reads/
Version (default, reanalysis)	12.3; 13.4	3.0.3; 4.2.5	v1.7.0	14.2; -	Not indicated	Beta
Cost	License Fees	Cost per sample	Freeware	Freeware	Freeware	Freeware
Bioinformatic IT needs	No	No	No	Some	No	No
Cloud-based / data management and storage	Local Installation or secured hosting / remote access	Online data storage	Online data storage	Data stored externally	Online data storage	Data stored externally
NGS sequencing platform	Illumina, Ion Torrent, PacBio	Illumina, Ion Torrent	Illumina, Ion Torrent	Illumina	Illumina	Not indicated
Consensus Sequence threshold percentage	20%	15%	20%	20%	20%	20%
Read aligner	BWA	Bowtie2	Bowtie2	Bowtie	Bowtie2	MINIMAP2
Minimum Length Cut-off	Not indicated	50	100	100	70	15
Score cut off	No QS filtering	30	30	15	25	15

Minimum variant quality	Not indicated	20	30	Not indicated	Not indicated	Not indicated
Minimum read depth	100	100	100	100	Not indicated	50
Minimum Amino acid Frequency	1% (0,01)	1% (0,01)	1% (0,01)	5% (0,05)	1% (0,01)	1% (0.01)
Reference sequence		LANL Consensus subtype B pol (2004)	HXB-2	HIV1-C-BR-JX140663-seed	HXB-2	LANL Consensus subtype B pol (2004)

2.3.7 Data processing for comparative analyses

Individual CSV outputs from each platform were consolidated and the eight mutations of interest were filtered out using a custom-designed script in R. The variant frequencies for each mutation were then compiled from all platforms into a CSV file.

The script performed the following steps:

- The working directory was set to the location of the input files containing the pipeline results. Data files for each pipeline were imported using the read.csv() function, with appropriate delimiters specified (semicolon for DeepChek, HyDRA, and PASeq; comma for Exatype). Only necessary columns for Exatype were retained to streamline processing.
- A predefined vector of mutation positions (K65R, K70R, K103N, V106M, M184V, G190A, K219E, P225H) was defined. For each pipeline, rows containing these mutation positions were identified and extracted using the which() function combined with filtering criteria.
- Data from each pipeline were processed to ensure consistency across datasets. Specific steps for each pipeline included:
- DeepChek: Relevant columns (Sample ID, Mutation, Prevalence) were selected, and a column for the tool name was added.
    - Exatype: The mutation column was parsed to extract the specific mutation using strsplit() and sapply().
    - HyDRA: Sample names were standardized, and mutation data was formatted as a concatenation of the wildtype, position, and mutation columns.
    - MiCall: Sample names were extracted and standardized, while relevant columns (Mutation, Prevalence) were retained.
    - PASeq: Mutation positions and prevalence values were parsed, and sample identifiers were extracted from filenames.
    - Processed data from all pipelines were consolidated into a single data frame using the rbind() function. Each row in the consolidated dataset included the following fields:
    - Sample ID: The unique identifier for each sample.
    - Tool: The pipeline from which the data originated.
    - Mutation: The detected mutation (e.g., K65R).

- Prevalence: The frequency of the detected mutation.
- Output Generation: The final compiled dataset was written to a CSV file (combined_file.csv), providing a unified dataset for subsequent statistical and comparative analyses.

2.3.8 Threshold Filtering

Pipelines were evaluated using different detection thresholds (nil, ≥1.00%, ≥2.00%, ≥5.00%. ≥20.00%. All analyses were repeated using each threshold setting and new spreadsheet generated.

2.4 Statistical analysis

2.4.1 Generation of a reference standard as proxy for “true value”

In the absence of a true value, a hypothetical “ground truth” was generated for further assessments of pipeline performance. This hypothetical value or reference standard was calculated as the median of all reported frequencies from all six pipelines (Lee *et al.*, 2020).

2.4.2 Data Processing Classification and Threshold

Pipeline-specific outputs were classified into qualitative categories (detected or not-detected) based on frequency detection value above or below the predefined thresholds.

2.4.3 Determination of Classification Outcomes

Pipeline outputs (positive/negative) were compared to the reference standard to determine true positive, false positive, false negative, and true negative classifications.

2.4.4 Determination of Sensitivity, Specificity, and Accuracy

Pipeline sensitivity, specificity and accuracy was determined using the formulas below:

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

Where TP = True Positive, FP = False Positive, FN = False Negative, TN = True Negative

2.4.5 Linearity

The linear range of each assay was assessed by comparing variant frequencies reported by each pipeline to reference standard values. Linear regression analysis was performed in Microsoft Excel (version 2309; Microsoft Corporation, Redmond, WA, USA) to evaluate the correlation and accuracy between the observed and expected frequencies.

2.5 Ethics clearance

The study was carried out in compliance with the approval granted by the University of the Witwatersrand Medical Human Research and Ethics Committee, under the ethics clearance certificate number M210525 (see Appendix 1).

CHAPTER 3 - RESULTS

3.1 Introduction

This chapter presents the comparative results of HIV drug resistance (HIVDR) analysis across six next-generation sequencing (NGS) pipelines. The objective was to evaluate the consistency of mutation detection and to determine how differences in pipeline design and parameter settings affect the reliability of HIVDR reporting. The analyses were conducted using sequencing data generated from the 2019 national HIVDR surveillance study.

Data processing abilities of six NGS pipelines were compared to assess their performance in analysing a dataset, measured as analytical sensitivity, specificity, accuracy and linearity. The pipelines used were DeepChek, Exatype, HyDRA, MiCall, PASeq and Stanford HIVdb-NGS (beta). The outputs generated from these pipelines were in the form of .csv files, which listed frequency values of mutations per sample. Data analysis was restricted to the detection of key mutations K65R, K70R, K103N, V106M, M184V, G190A, K219E, P225H only. The mutation frequency values reported by each pipeline were extracted and merged into a single combined dataset to facilitate direct comparative analysis across all platforms. Pipeline-generated data were then compared to the reference standard, as discussed in section 2.4.1.

3.2 Difference in minimum value reporting

Mutation frequency data from the six pipelines showed that DeepChek, Exatype and PASeq reported data with a range of 0.00% to 100.00%, whereas HyDRA and Stanford HIVdb-NGS (beta) were reporting data with a range of 1.00% to 100.00%. Frequency data from MiCall were reported as 5.00% to 100.00%. To accurately reflect a consensus outcome, it was preferable to then exclude all prevalence frequency values of <1.00% to ensure the reference standard reflects the majority of platforms.

3.3 Description of dataset

Pipeline-generated frequency data were categorized into <1.00% (not detected), ≥1.00% to <5.00%, ≥5.00 to <20.00%, ≥20.00 to <80% and ≥80% as shown in Figure 3.1. The most prevalent mutations detected were K103N and M184V, as evidenced by more than half of samples exhibiting frequency reads at ≥1.00% threshold, and most

being the dominant variant. The mutations V106M, K65R and P225H were prevalent in ~25%, 20% and 15% of samples, respectively. K70R, G190A, K219E were least prevalent with ~10% of samples being positive for these mutations.

The frequency at which the mutations were detected was generally consistent across all pipelines. When comparing data generated from each pipeline and categorized into <1.00% (not detected), ≥1.00% to <5.00% (very low frequency variant), ≥5.00% to <20.00% (low frequency variant), ≥20.00 to <80.00%, or ≥80.00% (dominant variant), minimal variation was shown in distribution of frequency categories. Small differences were noted in detection of the P225H mutation using the DeepChek at below 5.00%.

Overall, the distribution of mutation frequencies across all samples showed that most of the eight target drug resistance mutations (K65R, K70R, K103N, V106M, M184V, G190A, K219E, and P225H) were detected at frequencies ≥80%, indicating that these were dominant variants within the circulating viral quasispecies.

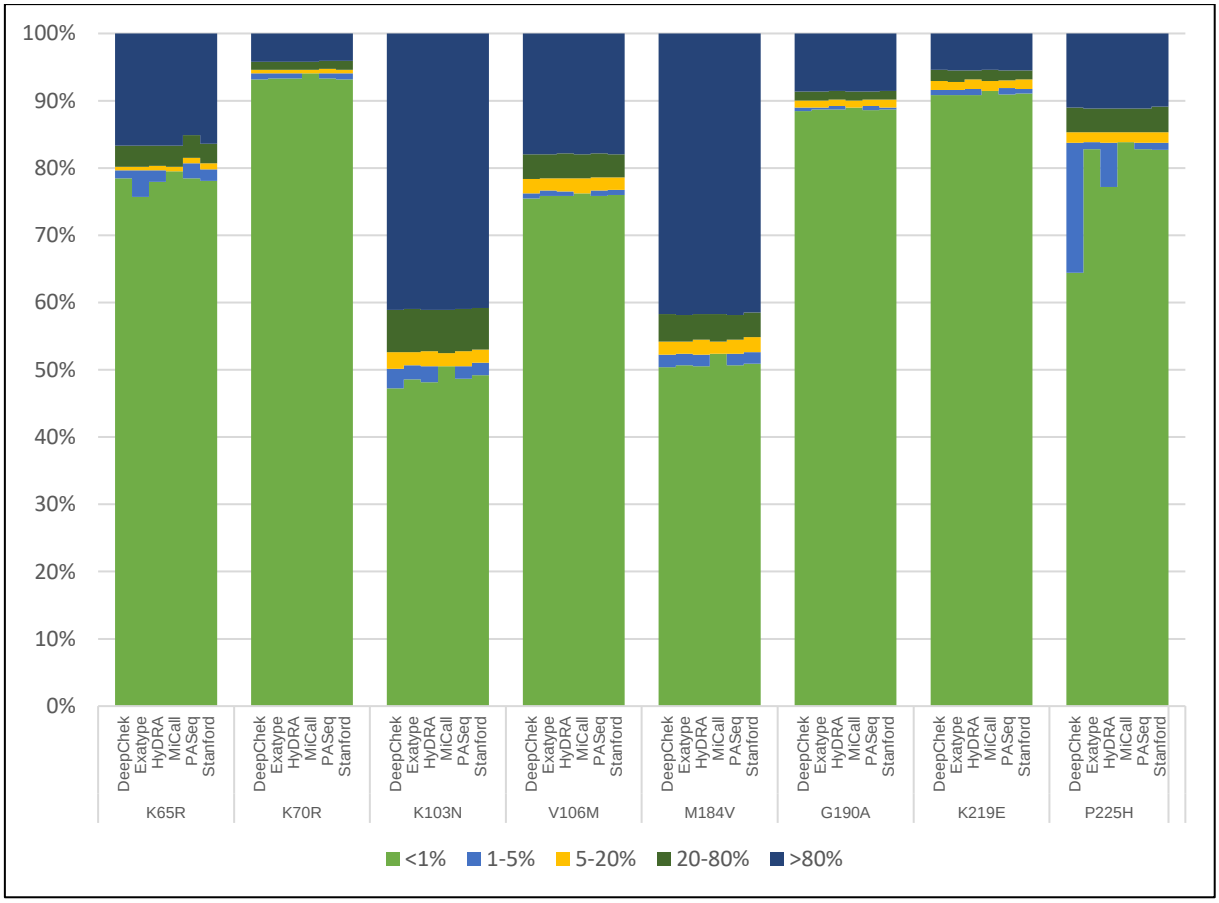


Figure 3.1. Comparison of mutation frequency distributions across six NGS HIVDR pipelines (DeepChek, Exatype, HyDRA, MiCall, PASeq, and Stanford HIVdb-NGS) for eight key drug-resistance mutations (K65R, K70R, K103N, V106M, M184V, G190A, K219E, P225H). Results are based on 753 successfully genotyped samples analysed under default settings. Outputs from six pipelines categorized

into <1.00% (not detected), ≥1.00%–<5.00% (very low frequency variant), ≥5.00%–<20.00% (low frequency variant), ≥20.00%–<80.00% (intermediate frequency variant), and ≥80.00% (dominant variant).

3.4 Comparison of frequency values of mutation detection obtained using default and recommended (Winnipeg) QC settings

3.4.1 Analytical sensitivity

Analytical sensitivity was determined as described in section 2.4.4 to assess the feasibility of the mutation being detected when it exists, as indicated by the presence of the mutation on the reference standard. Sensitivity was assessed using data generated from each pipeline assessed (DeepChek, Exatype, HyDRA, MiCall, PASeq and Stanford HIVdb-NGS (beta) (Table 3.1) relative to that of the median value obtained from all pipelines. Sensitivity for detection of existing variants was determined using threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%.

Table 3.1. Sensitivity of mutation detection using pipeline-specific default and recommended QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%. Abbreviations: ND = Not Detected.

Pipeline	Mutation	Default QC settings				Winnipeg QC settings			
		≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold	≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold
DeepChek	K65R	**97.62%**	**99.38%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K70R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	V106M	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	M184V	100.00%	100.00%	100.00%	**99.71%**	100.00%	100.00%	100.00%	**99.71%**
	G190A	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K219E	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	P225H	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	ALL	**99.73%**	**99.93%**	100.00%	**99.92%**	100.00%	100.00%	100.00%	**99.92%**
Exatype	K65R	**99.40%**	**99.38%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K70R	100.00%	**97.96%**	100.00%	100.00%	100.00%	100.00%	100.00%	**97.56%**

	K103N	**99.49%**	**99.48%**	**99.73%**	**99.72%**	**99.24%**	100.00%	**99.73%**	**99.72%**
	V106M	100.00%	**99.46%**	**99.44%**	**99.39%**	100.00%	100.00%	100.00%	100.00%
	M184V	100.00%	**99.73%**	**99.73%**	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K219E	100.00%	**98.53%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	P225H	**98.50%**	**99.22%**	**99.19%**	100.00%	**99.24%**	**99.21%**	**98.37%**	100.00%
	ALL	**99.66%**	**99.44%**	**99.71%**	**99.85%**	**99.73%**	**99.93%**	**99.78%**	**99.85%**
HyDRA	K65R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K70R	100.00%	100.00%	100.00%	100.00%	**98.08%**	100.00%	100.00%	**97.56%**
	K103N	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	V106M	100.00%	100.00%	100.00%	**99.39%**	100.00%	100.00%	100.00%	**99.39%**
	M184V	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	100.00%	100.00%	**97.62%**	100.00%	100.00%	100.00%	**97.62%**	100.00%
	K219E	100.00%	**98.53%**	**98.44%**	**98.11%**	100.00%	100.00%	**98.44%**	**98.11%**
	P225H	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	ALL	100.00%	**99.93%**	**99.78%**	**99.85%**	**99.93%**	100.00%	**99.78%**	**99.77%**
MiCall	K65R	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	K70R	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	K103N	ND	ND	**99.73%**	100.00%	ND	ND	**99.73%**	100.00%
	V106M	ND	ND	100.00%	**99.39%**	ND	ND	100.00%	100.00%
	M184V	ND	ND	**99.73%**	**99.71%**	ND	ND	**99.73%**	**99.71%**
	G190A	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	K219E	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	P225H	ND	ND	**99.19%**	100.00%	ND	ND	**99.19%**	100.00%
	ALL	ND	ND	**99.78%**	**99.85%**	ND	ND	**99.78%**	**99.92%**
PASeq	K65R	**97.02%**	**96.25%**	**94.84%**	**94.00%**	**97.02%**	**96.25%**	**94.84%**	**94.00%**
	K70R	100.00%	**97.96%**	100.00%	**97.56%**	**98.08%**	**97.96%**	100.00%	**97.56%**
	K103N	**99.24%**	**99.74%**	100.00%	100.00%	**99.49%**	100.00%	100.00%	100.00%

	V106M	100.00%	**99.46%**	**99.44%**	**98.79%**	100.00%	**99.46%**	**99.44%**	**99.39%**
	M184V	100.00%	**99.73%**	**99.73%**	100.00%	**99.73%**	**99.19%**	**99.73%**	100.00%
	G190A	100.00%	100.00%	**97.62%**	100.00%	100.00%	100.00%	**97.62%**	100.00%
	K219E	**98.57%**	**98.53%**	**96.88%**	100.00%	100.00%	100.00%	**98.44%**	100.00%
	P225H	**98.50%**	**98.44%**	100.00%	100.00%	100.00%	**99.21%**	100.00%	100.00%
	ALL	**99.25%**	**99.09%**	**98.99%**	**99.08%**	**99.38%**	**99.16%**	**99.07%**	**99.15%**
Stanford HIVdb-NGS (beta)	K65R	**98.81%**	**99.38%**	**99.35%**	**98.00%**	**98.81%**	**99.38%**	**99.35%**	**99.33%**
	K70R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	**98.22%**	**98.70%**	**98.94%**	**99.44%**	**98.48%**	**98.70%**	**98.94%**	**99.44%**
	V106M	**98.91%**	**98.91%**	**98.88%**	**98.79%**	**98.37%**	**98.37%**	**98.32%**	**98.78%**
	M184V	**99.47%**	**99.19%**	**99.18%**	**99.14%**	**99.20%**	**98.66%**	**99.18%**	**99.14%**
	G190A	100.00%	100.00%	**97.62%**	100.00%	100.00%	100.00%	**97.62%**	100.00%
	K219E	**97.14%**	**98.53%**	**98.44%**	**98.11%**	**97.14%**	**98.51%**	**98.44%**	**98.11%**
	P225H	**99.25%**	**99.22%**	100.00%	100.00%	**99.24%**	**99.21%**	100.00%	100.00%
	ALL	**98.91%**	**99.09%**	**99.21%**	**99.16%**	**98.77%**	**98.88%**	**99.14%**	**99.31%**

The sensitivity for data generated by each pipeline, at the four different thresholds applied, and by mutation analysed was high, with ranges of 94.00% – 100.00% determined. Overall, assessment of sensitivity did not significantly increase with increasing threshold values; however, sensitivity was marginally lower when analysing data at the ≥1.00% threshold. There was no evidence of any mutation showing overall reduced or increase sensitivity relative to others. There was a tendency for lower sensitivity from the data analysed using the PASeq and Stanford HIVdb-NGS (beta) pipelines, and to a lesser degree, Exatype pipeline, indicating that these pipelines were not detecting mutations relative to the other pipelines. Application of the Winnipeg setting resulted in marginal improvement of assessments of sensitivity using all pipeline data.

Analytical sensitivity was comparable across most pipelines, with DeepChek, HyDRA, and MiCall demonstrating consistently high detection rates for dominant variants. Minor variability was observed at lower frequency thresholds (1–5%), particularly in PASeq and Exatype, which occasionally under-reported variants near the detection

limit. These findings suggest that differences in default variant-calling thresholds and quality filtering parameters may influence the detection of low-abundance mutations.

3.4.2 Analytical specificity

Analytical specificity was determined as described in section 2.4.4 to assess the feasibility of the mutation not being detected when it does not exist, or a measure of false positivity. Specificity was calculated from each pipeline assessed (Table 3.2). Specificity was determined using threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%.

Table 3.2. Specificity of mutation detection using pipeline-specific default and recommended QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%. Abbreviations: ND = Not Detected.

Pipeline	Mutation	Default QC settings				Winnipeg QC settings			
		≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold	≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold
DeepChek	K65R	100.00%	100.00%	100.00%	**99.84%**	100.00%	100.00%	100.00%	**99.84%**
	K70R	**99.86%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	**97.83%**	**98.67%**	**99.22%**	**99.75%**	**97.83%**	**98.41%**	**99.22%**	**99.75%**
	V106M	**99.48%**	**100.00%**	**99.66%**	100.00%	**99.48%**	100.00%	**99.66%**	**99.83%**
	M184V	**99.48%**	**99.49%**	100.00%	**99.28%**	**99.74%**	100.00%	100.00%	**99.28%**
	G190A	**99.70%**	100.00%	100.00%	**99.85%**	**99.85%**	100.00%	100.00%	**99.85%**
	K219E	100.00%	100.00%	100.00%	**99.86%**	100.00%	**99.86%**	100.00%	**99.86%**
	P225H	**78.06%**	**99.68%**	100.00%	100.00%	**77.97%**	**99.69%**	100.00%	100.00%
	ALL	**96.68%**	**99.81%**	**99.89%**	**99.85%**	**96.72%**	**99.81%**	**99.89%**	**99.83%**
Exatype	K65R	**96.97%**	**99.17%**	100.00%	**99.84%**	**98.15%**	**99.34%**	100.00%	**99.84%**
	K70R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	100.00%	100.00%	100.00%	**99.50%**	100.00%	100.00%	**99.74%**	**99.50%**
	V106M	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	**99.83%**	100.00%
	M184V	100.00%	100.00%	100.00%	**99.52%**	**99.74%**	100.00%	100.00%	**99.52%**
	G190A	100.00%	100.00%	100.00%	**99.85%**	100.00%	100.00%	100.00%	**99.85%**

	K219E	100.00%	100.00%	100.00%	**99.72%**	100.00%	100.00%	100.00%	**99.86%**
	P225H	100.00%	100.00%	100.00%	100.00%	**99.84%**	**99.84%**	**99.84%**	100.00%
	ALL	**99.61%**	**99.89%**	100.00%	**99.83%**	**99.72%**	**99.89%**	**99.94%**	**99.85%**
HyDRA	K65R	100.00%	**99.83%**	100.00%	100.00%	100.00%	**99.83%**	100.00%	100.00%
	K70R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	**99.73%**	100.00%	100.00%	100.00%	**99.73%**	**99.73%**	**99.74%**	**99.75%**
	V106M	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	**99.83%**
	M184V	**99.74%**	**99.49%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K219E	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	P225H	**93.48%**	**99.68%**	100.00%	**93.48%**	**80.03%**	**99.69%**	100.00%	**80.03%**
	ALL	**99.07%**	**99.89%**	100.00%	100.00%	**97.26%**	**99.91%**	**99.98%**	**99.96%**
MiCall	K65R	ND	ND	**99.84%**	**99.84%**	ND	ND	**99.84%**	**99.84%**
	K70R	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	K103N	ND	ND	**99.74%**	**99.50%**	ND	ND	**99.74%**	**99.50%**
	V106M	ND	ND	**99.66%**	100.00%	ND	ND	**99.66%**	100.00%
	M184V	ND	ND	100.00%	**99.28%**	ND	ND	100.00%	**99.28%**
	G190A	ND	ND	100.00%	**99.85%**	ND	ND	100.00%	**99.85%**
	K219E	ND	ND	**99.86%**	**99.86%**	ND	ND	**99.86%**	**99.86%**
	P225H	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	ALL	ND	ND	**99.89%**	**99.83%**	ND	ND	**99.89%**	**99.83%**
PASeq	K65R	**99.83%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K70R	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	V106M	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	M184V	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	**99.85%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K219E	100.00%	100.00%	100.00%	100.00%	100.00%	**99.86%**	100.00%	100.00%
	P225H	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

	ALL	**99.96%**	100.00%	100.00%	100.00%	100.00%	**99.98%**	100.00%	100.00%
Stanford HIVdb-NGS (beta)	K65R	**99.83%**	**99.83%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K70R	**99.86%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	V106M	**99.83%**	100.00%	100.00%	100.00%	**99.83%**	100.00%	100.00%	100.00%
	M184V	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	**99.85%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K219E	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	P225H	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	ALL	**99.94%**	**99.98%**	100.00%	100.00%	**99.98%**	100.00%	100.00%	100.00%

The specificity for all pipelines at each threshold was high, with a range 96.68% – 100.00%. Overall, assessment of sensitivity did not significantly improve with increasing threshold values; however, specificity was slightly reduced when analysing data at the >1.00% threshold and slightly improved when analysing data at the ≥5.00% threshold. There was a tendency for lower specificity from the data analysed using DeepChek, Exatype and HyDRA, indicating that these pipelines were detecting more mutations than the other pipelines. Application of the Winnipeg setting resulted in marginal improvement of assessments of specificity using all pipeline data.

The DeepChek and HyDRA pipelines detected the P225H mutation amongst a higher proportion of samples, resulting in a specificity determination of 77.97% and 93.48%, respectively. The DeepChek pipeline results showed 138 individual specimens with P225H frequency values with IQR 1.06 – 1.23%, whereas the HyDRA pipeline generated 41 samples with P225H results with IQR 1.05 – 1.19%.

Analytical specificity remained consistently high across all pipelines, indicating a low occurrence of false-positive variant calls. Slight reductions in specificity were observed in pipelines employing less stringent filtering or variant quality control, particularly in genomic regions with lower read coverage or homopolymeric stretches. These differences likely reflect variations in algorithmic assumptions—such as the choice of aligner, variant-calling thresholds, and quality-score filters—that influence whether borderline variants are classified as true or false positives. Pipelines applying stricter consensus and read-depth criteria, such as HyDRA and Stanford HIVdb-NGS,

minimize false positives but may omit true low-frequency variants, whereas more permissive tools like DeepChek can overcall minority variants. This highlights the balance between maximizing sensitivity and maintaining specificity in HIVDR variant detection.

3.4.3 Accuracy

Analytical accuracy was determined as described in section 2.4.4 and calculated using data from each pipeline assessed (DeepChek, Exatype, HyDRA, MiCall, PASeq and Stanford HIVdb-NGS (beta) (Table 3.3). Accuracy for detection of mutation was determined using threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%.

Table 3.3. Accuracy of mutation detection using pipeline-specific default QC settings. Mutations analysed were K65R, K70R, K103N, V106M, M184V, G190A, K219E and P225H. Analysis was performed at threshold detection values of ≥1.00%, ≥2.00%, ≥5.00% and ≥20.00%.

Pipeline	Mutation	Default QC settings				Winnipeg QC settings			
		≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold	≥1.00% Threshold	≥2.00% Threshold	≥5.00% Threshold	≥20.00% Threshold
DeepChek	K65R	**99.48%**	**99.87%**	100.00%	**99.87%**	100.00%	100.00%	100.00%	**99.87%**
	K70R	**99.87%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	**98.95%**	**99.34%**	**99.61%**	**99.87%**	**98.95%**	**99.21%**	**99.61%**	**99.87%**
	V106M	**99.61%**	100.00%	**99.74%**	100.00%	**99.61%**	100.00%	**99.74%**	**99.87%**
	M184V	**99.74%**	**99.74%**	100.00%	**99.48%**	**99.87%**	100.00%	100.00%	**99.48%**
	G190A	**99.74%**	100.00%	100.00%	**99.87%**	**99.87%**	100.00%	100.00%	**99.87%**
	K219E	100.00%	100.00%	100.00%	**99.87%**	100.00%	**99.87%**	100.00%	**99.87%**
	P225H	**81.89%**	**99.74%**	100.00%	100.00%	**81.76%**	**99.74%**	100.00%	100.00%
	ALL	**97.41%**	**99.84%**	**99.92%**	**99.87%**	**97.51%**	**99.85%**	**99.92%**	**99.85%**
Exatype	K65R	**97.51%**	**99.21%**	100.00%	**99.87%**	**98.56%**	**99.48%**	100.00%	**99.87%**
	K70R	**100.00%**	**99.87%**	100.00%	100.00%	100.00%	100.00%	100.00%	**99.87%**
	K103N	**99.74%**	**99.74%**	**99.87%**	**99.61%**	**99.61%**	100.00%	**99.74%**	**99.61%**
	V106M	100.00%	**99.87%**	**99.87%**	**99.87%**	100.00%	100.00%	**99.87%**	100.00%

	M184V	100.00%	**99.87%**	**99.87%**	**99.74%**	**99.87%**	100.00%	100.00%	**99.74%**
	G190A	100.00%	100.00%	100.00%	**99.87%**	100.00%	100.00%	100.00%	**99.87%**
	K219E	100.00%	**99.87%**	100.00%	**99.74%**	100.00%	100.00%	100.00%	**99.87%**
	P225H	**99.74%**	**99.87%**	**99.87%**	100.00%	**99.74%**	**99.74%**	**99.61%**	100.00%
	ALL	**99.62%**	**99.79%**	**99.93%**	**99.84%**	**99.72%**	**99.90%**	**99.90%**	**99.85%**
HyDRA	K65R	100.00%	**99.87%**	100.00%	100.00%	100.00%	**99.87%**	100.00%	100.00%
	K70R	100.00%	100.00%	100.00%	100.00%	**99.87%**	100.00%	100.00%	**99.87%**
	K103N	**99.87%**	100.00%	100.00%	100.00%	**99.87%**	**99.87%**	**99.87%**	**99.87%**
	V106M	100.00%	100.00%	100.00%	**99.87%**	100.00%	100.00%	100.00%	**99.74%**
	M184V	**99.87%**	**99.74%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	G190A	100.00%	100.00%	**99.74%**	100.00%	100.00%	100.00%	**99.74%**	100.00%
	K219E	100.00%	**99.87%**	**99.87%**	**99.87%**	100.00%	100.00%	**99.87%**	**99.87%**
	P225H	**94.62%**	**99.74%**	100.00%	**94.62%**	**83.46%**	**99.74%**	100.00%	**83.46%**
	ALL	**99.29%**	**99.90%**	**99.95%**	**99.97%**	**97.90%**	**99.93%**	**99.93%**	**99.92%**
MiCall	K65R	ND	ND	**99.87%**	**99.87%**	ND	ND	**99.87%**	**99.87%**
	K70R	ND	ND	100.00%	100.00%	ND	ND	100.00%	100.00%
	K103N	ND	ND	**99.74%**	**99.74%**	ND	ND	**99.74%**	**99.74%**
	V106M	ND	ND	**99.74%**	**99.87%**	ND	ND	**99.74%**	100.00%
	M184V	ND	ND	**99.87%**	**99.48%**	ND	ND	**99.87%**	**99.48%**
	G190A	ND	ND	100.00%	**99.87%**	ND	ND	100.00%	**99.87%**
	K219E	ND	ND	**99.87%**	**99.87%**	ND	ND	**99.87%**	**99.87%**
	P225H	ND	ND	**99.87%**	100.00%	ND	ND	**99.87%**	100.00%
	ALL	ND	ND	**99.87%**	**99.84%**	ND	ND	**99.87%**	**99.85%**
PASeq	K65R	**99.21%**	**99.21%**	**98.95%**	**98.82%**	**99.34%**	**99.21%**	**98.95%**	**98.82%**
	K70R	100.00%	**99.87%**	100.00%	**99.87%**	**99.87%**	**99.87%**	100.00%	**99.87%**
	K103N	**99.61%**	**99.87%**	100.00%	100.00%	**99.74%**	100.00%	100.00%	100.00%
	V106M	100.00%	**99.87%**	**99.87%**	**99.74%**	100.00%	**99.87%**	**99.87%**	**99.87%**
	M184V	100.00%	**99.87%**	**99.87%**	100.00%	**99.87%**	**99.61%**	**99.87%**	100.00%

	G190A	**99.87%**	100.00%	**99.74%**	100.00%	100.00%	100.00%	**99.74%**	100.00%
	K219E	**99.87%**	**99.87%**	**99.74%**	100.00%	100.00%	**99.87%**	**99.87%**	100.00%
	P225H	**99.74%**	**99.74%**	100.00%	100.00%	100.00%	**99.87%**	100.00%	100.00%
	ALL	**99.79%**	**99.79%**	**99.77%**	**99.80%**	**99.85%**	**99.79%**	**99.79%**	**99.82%**
Stanford HIVdb-NGS (beta)	K65R	**99.61%**	**99.74%**	**99.87%**	**99.61%**	**99.74%**	**99.87%**	**99.87%**	**99.87%**
	K70R	**99.87%**	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	K103N	**99.08%**	**99.34%**	**99.48%**	**99.74%**	**99.21%**	**99.34%**	**99.48%**	**99.74%**
	V106M	**99.61%**	**99.74%**	**99.74%**	**99.74%**	**99.48%**	**99.61%**	**99.61%**	**99.74%**
	M184V	**99.74%**	**99.61%**	**99.61%**	**99.61%**	**99.61%**	**99.34%**	**99.61%**	**99.61%**
	G190A	**99.87%**	100.00%	**99.74%**	100.00%	100.00%	100.00%	**99.74%**	100.00%
	K219E	**99.74%**	**99.87%**	**99.87%**	**99.87%**	**99.74%**	**99.87%**	**99.87%**	**99.87%**
	P225H	**99.87%**	**99.87%**	100.00%	100.00%	**99.87%**	**99.87%**	100.00%	100.00%
	ALL	**99.69%**	**99.77%**	**99.82%**	**99.82%**	**99.69%**	**99.74%**	**99.80%**	**99.85%**

Assessment of accuracy showed results of ≥99.48% for most of the mutations analysed and using the majority of pipelines, showing the ability to analyse the data accurately is mostly consistent across all pipelines. Accuracy of K65R mutation was slightly lower at 97.51% when analysed using the Exatype pipeline, and accuracy of P225H detection was lower at 81.76% using DeepChek and 83.46% using HyDRA.

Accuracy across pipelines was generally high, particularly for mutations detected at frequencies above 20%. However, discrepancies were noted at lower variant frequencies, where DeepChek and MiCall occasionally reported false negatives for minority variants, while PASeq tended to slightly overestimate frequencies near the 5% threshold. These inconsistencies likely reflect differences in variant-calling stringency and handling of ambiguous base positions. Despite these minor variations, all pipelines achieved acceptable accuracy within expected analytical limits.

3.4.4 Linearity

The linear range was determined using linear regression analysis where mutation frequencies obtained from each pipeline were analysed against the median frequency values (Figure 3.2). This included all frequency data present at ≥1.00% and plotted

relative to reference standard data present at ≥1.00%. Each pipeline data was represented relative to the reference standard (expected frequency) data.

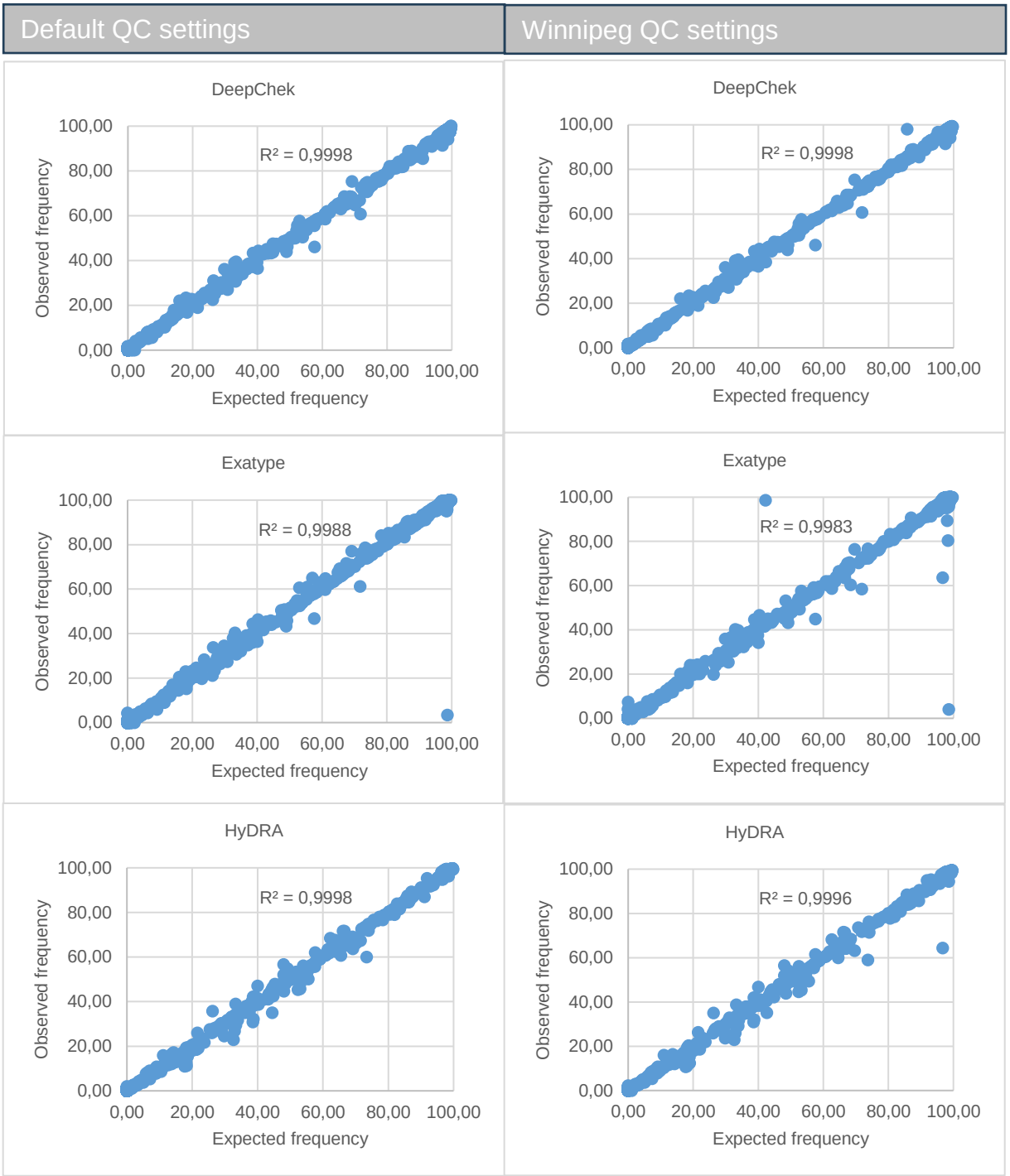


Figure 3.2. ***Linear regression analysis of mutation frequencies obtained from DeepChek, Exatype, and HyDRA pipelines.*** *Each point represents the observed mutation frequency versus the median (expected) frequency for values > 1 %. Regression lines indicate concordance between reported and expected frequencies for these pipelines.>1.00%*

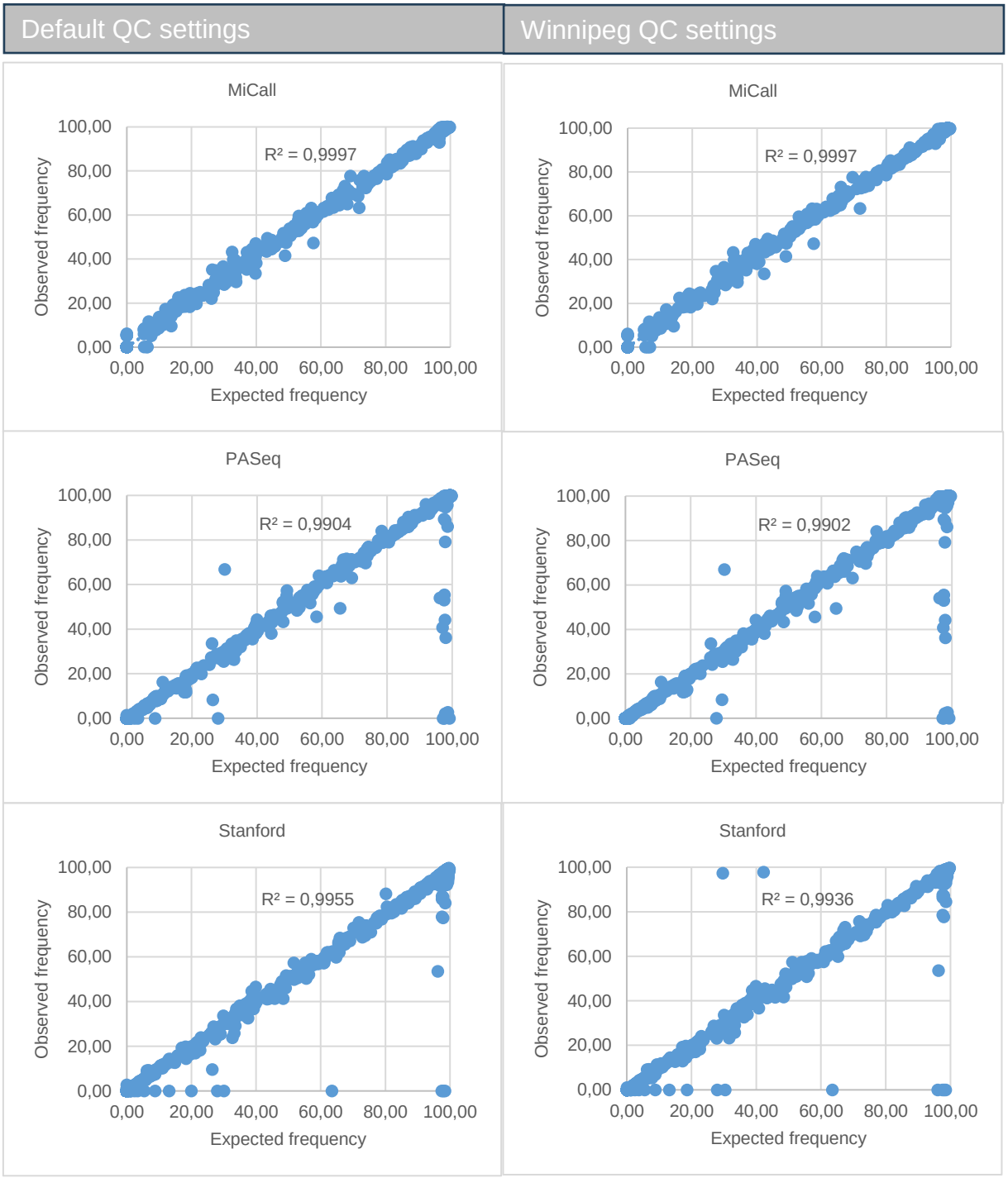


Figure 3.3. ***Linear regression analysis of mutation frequencies obtained from PASeq, Stanford HIVdb-NGS, and MiCall pipelines.*** *Each point represents the observed mutation frequency versus the median (expected) frequency for values > 1 %. Regression lines indicate concordance between reported and expected frequencies for these pipelines.>1.00%*

Assessment of linearity for frequency values relative to the median (expected) frequency from the DeepChek, HyDRA and MiCall pipelines were high, with correlation coefficients of 0.9998 and 0.9997 calculated. The correlation coefficients of data analysed using the Exatype, PASeq and Stanford HIVdb-NGS (beta) pipelines were marginally lower at 0.9983, 0.9902 and 0.9936, respectively.

Linearity was assessed by comparing observed mutation frequencies against reference values using linear regression analysis performed in Microsoft Excel (version 365). All pipelines demonstrated strong correlation ($R^2 > 0.95$), indicating reliable quantitative performance across frequency ranges. Minor deviations from perfect linearity were observed for low-frequency variants, particularly in Exatype and DeepChek, suggesting potential rounding or normalization effects during variant quantification. These findings reinforce the importance of consistent frequency scaling when comparing outputs across platforms.

3.5 Investigating discrepant samples

Twenty-six samples exhibited major discrepancies in mutation frequency detection (Table 3.4). These included:

- Five samples where K103N detection showed variation between the pipelines:
  - One sample (MK0192) where K103N was detected at ~30% by most pipelines and 66.86% by PASeq and not detected (0.00%) by Stanford
  - One sample (MK0222) where K103N was dominant variant (99%) by most pipelines and ~40% by PASeq
  - One sample (MK0409) where K103N was dominant variant (~99%) by most pipelines and ~64% by Exatype and HyDRA using Winnipeg settings only
  - One sample where K103N was detected as dominant variant (~99%) but minor variant (3.38, 4,10%) by Exatype
  - One sample (PE0187) where K103N reported as dominant variant and not detected by Stanford, under default and Winnipeg settings
- Fourteen samples where K65R showed variation between the pipelines:
  - Three samples (AD0223, AD0280, DG0236) where K65R was dominant variant (~99%) by most pipelines and not detected (0.00%) by PASeq
  - Five samples (DG0239, ED0129, IA0159, NG0151, NG0261) where K65R reported as dominant variant (~99%) and minor variant by PASeq

    - Five samples with lower frequency reported by PASeq relative to other pipelines (FR006, IA0314, MD0192, NE0412, PE0134)
    - One sample with K65R reported at ~30% by other pipelines and not detected (0.00% by PASeq and Stanford
- One sample (ED0478) where K70R was reported at a lower frequency by PASeq (~54%)
- Three samples where M184V showed variation between the pipelines:
    - Two samples (IA0152, NE0476) where M184V was the dominant variant (~99%) by other pipelines and not detected (0.00%) by Stanford
    - One sample (MK0027) where M184V was reported at ~40% by other pipelines and ~98% by DeepChek and Stanford using Winnipeg settings
- Three samples where V106M showed variation between pipelines:
    - One sample where V106M was dominant variant (~99%) using other pipelines and not detected (0.00%) using Stanford under Winnipeg settings
    - One sample (CM0900) where V106M was dominant variant (~63%) using other pipelines and not detected (0.00%) using Stanford

- *Table 3.4. List of samples with large differences in mutations frequency reporting across the different pipelines*

		Default Settings						Winnipeg settings					
SID	Mutation	DeepChek	Exatype	HyDRA	MiCall	PASeq	Stanford	DeepChek	Exatype	HyDRA	MiCall	PASeq	Stanford
MK0192	K103N	30.09	28.65	30.42	28.36	66.86	**0.00**	30.09	30.50	30.33	28.36	66.86	**0.00**
MK0222	K103N	91.41	99.65	98.91	99.66	**40.72**	97.10	91.41	99.70	98.03	99.66	**40.72**	97.43
MK0409	K103N	96.67	99.56	98.82	99.57	99.71	97.68	96.67	**63.60**	**64.40**	99.57	99.71	98.31
PE0140	K103N	98.04	**3.38**	99.16	99.78	99.84	98.75	98.04	**4.10**	98.54	99.78	99.84	98.92
PE0187	K103N	97.47	96.55	98.92	99.63	99.69	**0.00**	97.47	97.10	98.21	99.63	99.69	**0.00**
AD0223	K65R	97.60	99.75	99.38	99.70	**0.00**	96.13	97.60	99.90	98.72	99.70	**0.00**	96.89
AD0280	K65R	97.25	99.56	98.99	99.77	**0.00**	95.60	97.25	99.20	98.01	99.77	**0.00**	96.43
DG0236	K65R	99.20	99.88	99.63	99.85	**0.00**	98.10	99.20	99.90	99.40	99.85	**0.00**	98.37
DG0239	K65R	98.73	99.63	99.22	99.66	**2.62**	97.87	98.73	99.70	98.89	99.66	**2.62**	98.04
ED0129	K65R	30.97	33.84	26.45	35.12	**8.25**	**9.57**	30.97	29.60	26.08	35.12	**8.25**	**97.29**
FR0006	K65R	97.60	99.61	99.14	99.66	**55.40**	**77.84**	97.60	99.50	98.36	99.66	**55.40**	**78.59**
IA0159	K65R	97.87	99.47	99.09	99.54	**2.10**	97.46	97.87	99.40	98.79	99.54	**2.10**	97.92
IA0314	K65R	98.02	99.75	99.13	99.73	**36.16**	96.73	98.02	99.70	98.79	99.73	**36.16**	97.21
MD0192	K65R	97.81	99.72	99.48	99.69	**44.12**	97.36	97.81	99.80	98.92	99.69	**44.12**	97.97
NE0412	K65R	97.89	99.46	99.00	99.66	**79.12**	**77.38**	97.89	99.50	98.41	99.66	**79.12**	**77.66**

NG0151	K65R	97.85	99.63	99.23	99.63	**1.56**	95.12	97.85	99.50	98.61	99.63	**1.56**	95.71
NG0261	K65R	97.93	99.61	99.09	99.72	**1.20**	96.52	97.93	99.60	98.35	99.72	**1.20**	97.10
NG0493	K65R	28.04	30.68	28.93	29.26	**0.00**	**0.00**	28.04	27.90	28.93	29.26	**0.00**	**0.00**
PE0134	K65R	97.56	99.46	99.13	99.50	**53.00**	85.80	97.56	99.50	98.53	99.50	**53.00**	86.44
ED0478	K70R	96.25	98.57	98.22	98.68	**54.02**	**53.48**	96.25	98.50	97.41	98.68	**54.02**	**53.52**
IA0152	M184V	98.50	95.05	99.22	99.78	99.83	**0.00**	98.50	96.60	98.68	99.78	99.83	**0.00**
MK0027	M184V	38.49	37.74	40.64	33.46	42.25	39.75	38.49	**98.60**	41.09	33.46	42.25	**97.80**
NE0467	M184V	97.97	95.60	98.76	99.33	99.49	**0.00**	97.97	95.20	98.44	99.33	99.49	**0.00**
AD0151	V106M	95.48	98.58	97.71	99.55	99.71	96.79	95.48	96.50	95.99	99.55	99.71	**0.00**
CM0897	V106M	20.01	20.26	20.64	19.34	18.29	**0.00**	22.79	21.50	21.72	23.04	20.09	20.23
CM0900	V106M	63.46	63.41	65.93	67.74	63.84	**0.00**	63.46	61.80	64.94	67.74	63.85	**0.00**

Of the five samples with discrepant K103N detection, three would have resulted in the mutation not being reported at all thresholds and possibly resulting in an incorrect HIVDR diagnosis. Of these, two were not reported by the Stanford pipeline, and one by the Exatype pipeline.

Of the fourteen samples with discrepant K65R detection, one would have resulted in the mutation not being detected at the >20% threshold, a further four at the >5% threshold, and a further four at the >1.00% threshold. All fourteen discrepancies were noted in data generated by the PASeq pipeline, whereas one sample was not detected using Stanford

Using data generated by the Stanford pipeline, 2 samples with the M184V mutation were not detected, and 3 samples with the V106M mutation were not detected.

CHAPTER 4 - DISCUSSION

This study aimed to address the analytical variability observed across bioinformatics pipelines used for HIVDR interpretation. By comparing six widely used NGS analysis platforms (PASeq, HyDRA, MiCall, DeepChek HIV, Stanford HIVdb-NGS (beta), and Exatype) the research sought to evaluate how differences in pipeline design, parameter settings, and variant-calling strategies influence the detection and reporting of key drug resistance mutations. The goal was to identify factors contributing to inconsistencies in reporting and to propose quality control measures that could improve reproducibility and reliability in HIVDR surveillance and clinical testing.

The onset of HIVDR is a major obstacle to successful and sustainable ART, at a patient and programmatic level (Ji, Sandstrom, *et al.*, 2020). Genotypic HIVDR testing identifies mutations associated with resistance to ART agents and drug classes and is used to identify appropriate drug regimens for optimal patient management, research and surveillance purposes. Next-generation sequencing approaches to HIVDR mutation detection are increasingly being adopted, due to improved scalability, throughput, and enhanced sensitivity for detection of low frequency variants. Specifically, studies comparing data generated using conventional Sanger sequencing and NGS data show superior ability to detect mutations associated with HIVDR present at frequencies below 20%, which is considered as Sanger-equivalent (Alidjinou *et al.*, 2017). This enhanced ability to detect lower frequency mutations supports accurate diagnosis and treatment decision-making.

Despite the described advantages of NGS for HIVDR testing, challenges still exist for wide-scale adoption of these solutions, including high cost of infrastructure requirements and reagents, training of personnel, validated of laboratory testing protocols, standardized data analysis outputs, and external quality assessment programs (Ávila-Ríos *et al.*, 2020).

Next-generation sequencing workflows include wet-lab procedures (nucleic acid extraction, sample amplification by PCR or enrichment and library preparation), sample sequencing and data analysis (Ji, Sandstrom, *et al.*, 2020). Each step can introduce variations and potential biases that can affect the NGS output. This study investigated differences in NGS data analysis pipelines only, by using a single set of data generated from a surveillance project performed in 2019. It is important to note that the results

generated here do not consider possible errors, biases or variations introduced during wet lab and sequencing steps.

Studies evaluating different variant calling tools in large-scale clinical sequencing projects emphasize the critical importance of assessing sensitivity, specificity, and accuracy in mutation detection (Yao *et al.*, 2020). Similarly, Sorelle et al. (2013) assessed bioinformatic pipelines for somatic mutation detection in cancer sequencing data, highlighting the need for comprehensive evaluation of these metrics to ensure /reliable results (SoRelle, Wachsmann and Cantarel, 2020). Furthermore, Liu et al. compared variant detection tools for genetic analysis using NGS data and underscored the significance of sensitivity, specificity, and overall accuracy in optimizing mutation detection pipelines (Liu *et al.*, 2013).

Data comparisons were performed by generating a 'reference standard' from which each pipeline data was compared individually. The reference standard value was determined by generating a median value of all six pipeline data points for each sample and for each of the eight mutations analysed. In order the ensure the representativeness of the reference standard, it was considered when a minimum of four of six values were available. This comparative approach was originally described by Ji *et al.* (2018) and Ji, Parkin, *et al.* (2020) and was applied in the current study to enable consistent and objective comparison between pipeline outputs (Ji *et al.*, 2018; Ji, Parkin, *et al.*, 2020).

It is important to note that this analysis does not reflect pipeline superiority or ability to accurately quantify mutations, given that the 'true value' is not known in real-world datasets. The analysis compares data generated using six different pipelines and provides a platform to identify discrepancies for further investigation and potential processing improvement.

4.1 Detection of mutations below 1.00%

After completion of data processing, it was noted that mutation frequency data from the DeepChek, Exatype and PASeq pipelines reported data with a range of 0.00% to 100.00%, HyDRA and Stanford HIVdb-NGS (beta) were reporting data with a range of 1.00% to 100.00%, and MiCall reported with a range of 5.00% to 100.00%. This does not infer that the HyDRA, Stanford HIVdb-NGS (beta) pipelines were unable to

determine frequency values below their minimum reporting threshold but rather contained mechanisms to remove all frequency values below their minimum reporting values respectively. It was not possible to generate a meaningful consensus outcome below 1% as only three of six pipelines were contributing data, therefore it was necessary to then exclude all prevalence frequency values of <1.00% to ensure the reference standard reflects the majority of platforms.

While next-generation sequencing enables the detection of variants at frequencies below 1.00 %, these minor variants are generally considered research-only and are not yet routinely actionable in clinical decision-making, as their impact on treatment outcomes remains uncertain (Wensing et al., 2022; Zhou et al., 2023). Nonetheless, pipelines that exclude variants below 1.00 % may overlook emerging resistance signals that could inform early detection of resistance evolution in surveillance contexts (Wensing et al., 2022; Zhou et al., 2023).

4.2 Assessment of pipeline sensitivity, specificity, accuracy and linearity

Assessments of sensitivity, specificity and accuracy showed very high agreement between the data generated from each of the six pipelines, with sensitivity, specificity and accuracy values of between 99% and 100.00% generated in most instances, indicating that very small numbers of specimens showed disagreement between the different pipelines. A slight trend towards reduced sensitivity from the data analysed using the PASeq and Stanford HIVdb-NGS (beta) pipelines, and to a lesser degree, Exatype pipeline, indicating that these pipelines were not detecting mutations relative to the other pipelines. Similarly, there was a slight tendency towards lower specificity from the data analysed using DeepChek, Exatype and HyDRA, indicating that these pipelines were detecting more mutations than the other pipelines. Assessments for pipeline accuracy showed results of ≥99.48% for most of the mutations analysed and using the majority of pipelines, showing the ability to analyse the data accurately is mostly consistent across all pipelines. For this analysis, some exceptions were noted, with accuracy of K65R mutation being slightly lower at 97.51% when analysed using the Exatype pipeline, and accuracy of P225H detection was lower at 81.76% using DeepChek and 83.46% using HyDRA.

Assessment of linearity for frequency values relative to the median (expected) frequency from all pipelines were very high, with the DeepChek, HyDRA and MiCall pipelines showing correlation coefficients of 0.9998 and 0.9997, and Exatype, PASeq and Stanford HIVdb-NGS (beta) pipelines showing correlation coefficients that were marginally lower at 0.9983, 0.9902 and 0.9936, respectively. However, these values remain extremely high, indicating very high levels of agreement in data generated from each of the six pipelines. However, some exceptions are noted in Figures 3.2 and 3.3, with notable exceptions highlighted in Table 3.4. The findings of this analysis are notably higher that those reported in previous studies, where sensitivity values of >99% and specificity values of >91% were reported (Lee *et al.*, 2020).

It is important to note that, in the absence of a definitive biological "truth," the sensitivity and accuracy values reported in this study should be interpreted as relative metrics (Ji et al., 2018; Zook et al., 2019). They reflect agreement between computational outputs rather than absolute diagnostic certainty, underscoring the dependence of performance measures on pipeline-specific definitions and reference assumptions.

4.3 Application of Winnipeg settings to improve agreement between pipelines

Guidelines for standardization of NGS data processing of HIVDR data have been proposed in the "Winnipeg Consensus" (Ji *et al.*, 2018). These guidelines include standardization of NGS read quality control, real alignment and reference mapping, and variant calling and quality control. Most pipelines in existence were developed independently and in the absence of reference and reporting guidelines.

The Winnipeg settings recommend a lower consensus sequence percentage threshold (5-15%) compared to the default settings (20%) (Ji *et al.*, 2018). This difference in the consensus sequence percentage threshold could potentially make the Winnipeg settings more sensitive in detecting minor resistance variants present at lower frequencies. However, this increased sensitivity may come at the cost of a higher risk of false positive variant calls, as the lower threshold may be more susceptible to technical artefacts or sequencing errors.

The Winnipeg settings additionally recommend using the Bowtie2 read aligner algorithm, while some pipelines use the Burrows-Wheeler Aligner (Ji *et al.*, 2018). The choice of read aligner can impact the accuracy of variant detection, as different

algorithms have varying strengths and weaknesses in terms of alignment sensitivity, computational efficiency, and the ability to handle INDELS and complex genomic regions.

The Winnipeg settings implement a higher minimum length cut-off of 75 base pairs for variant calls, compared to no such cut-off in the default settings (Ji *et al.*, 2018). This more stringent length requirement in the Winnipeg pipeline could potentially filter out some true positive variant calls, especially for shorter resistance-associated mutations, which may negatively impact the overall sensitivity of the platform.

The Winnipeg settings apply a minimum variant quality score threshold of 30, while the default settings do not utilize any quality score filtering (Ji *et al.*, 2018). This quality score threshold in the Winnipeg pipeline could help to reduce the number of false positive variant calls by prioritizing high-confidence variant assignments. However, this quality score filtering may also result in the exclusion of some true positive variant calls, particularly those with lower quality scores.

While the effort to standardize elements of HIVDR NGS reporting are encouraged, application of these settings had minimal if any impact on assessments of sensitivity, specificity, accuracy and linearity in the dataset analysed.

4.4 Pipeline processes that may be contributing to the marginal differences noted

These disparities noted in this study, as described in Section 3.5, refer to the minor differences in mutation frequency detection observed among the six HIVDR analysis pipelines. These discrepancies, albeit rare, are likely attributed to variations in algorithm design and parameter settings across the compared pipelines. Each pipeline may employ distinct variant calling criteria and error correction methods, leading to differences in sensitivity for detecting mutations, especially those present at low frequencies or in regions of low coverage (Zook *et al.*, 2019; Barbitoff *et al.*, 2022).

The discrepancies described in Section 3.5 were not random but followed pipeline-specific patterns. For instance, PASeq and Stanford HIVdb-NGS frequently under-detected low-frequency variants such as K65R and M184V, likely because of more stringent consensus thresholds and variant-quality filters that prioritize specificity over sensitivity. Conversely, DeepChek and Exatype occasionally reported minority variants

at higher frequencies, indicating more permissive filtering or differences in their alignment scoring algorithms. HyDRA and MiCall demonstrated intermediate performance, maintaining a balance between sensitivity and specificity. These patterns suggest that algorithmic design choices, such as read-alignment approach, indel handling, and quality-score weighting, play a decisive role in each pipeline's ability to detect clinically relevant mutations. Clinically, missing or under-reporting key resistance mutations may delay regimen optimization and contribute to continued virologic failure in resource-limited settings where confirmatory testing is unavailable.

Moreover, differences in pipeline optimization efforts may contribute to variability in sensitivity, specificity and accuracy of variant detection. Pipelines undergo iterative refinement and updates aimed at enhancing their performance, and disparities in optimization strategies among pipelines can result in variations for mutation detection (Barbitoff *et al.*, 2022; Olson *et al.*, 2022).

Optimization in the context of HIVDR bioinformatics pipelines refers to the continuous process of refining analytical parameters and computational algorithms to enhance performance (Barbitoff et al., 2022; Olson et al., 2022). This may include adjustments to read-alignment scoring matrices, variant quality thresholds, consensus percentage cut-offs, and filtering strategies aimed at improving the balance between sensitivity and specificity. Some pipelines, such as HyDRA and MiCall, undergo iterative updates that integrate improved reference databases or adopt enhanced variant-calling modules, while others employ adaptive algorithms that automatically calibrate quality metrics based on sequencing depth or error profiles. These refinements can substantially influence the detection of low-frequency variants and the reproducibility of results. Pipelines that are more frequently optimized or incorporate machine learning–based error correction tend to maintain higher accuracy and stability across diverse datasets, while those lacking ongoing refinement may exhibit greater variability in mutation detection performance

Critical evaluation of bioinformatics algorithms, quality control measures, and optimization efforts is essential to understand the observed differences in sensitivity and ensure reliable mutation detection in HIV drug resistance testing.

Variabilities in sensitivity and specificity, especially at lower thresholds, highlight the importance of cautious interpretation of mutation data and the need for tailored pipeline

selection and optimization. Blassel *et al*. (2021) highlight the complexities and challenges in detecting drug resistance mutations in HIV, stressing that variations in bioinformatics approaches can significantly impact the reliability of mutation detection (Blassel *et al.*, 2021). They discuss how high error rates in sequencing data, differences in read alignment algorithms, and the choice of reference sequences can all lead to variability in mutation detection outcomes. Blassel *et. al* further emphasize the critical role of threshold settings and bioinformatics pipeline configurations, noting that lower thresholds can increase sensitivity but reduce specificity by allowing more false positives. Variant calling algorithms should be robust and comprehensive validation of bioinformatics tools are necessary to ensure consistent and reliable mutation detection, which is crucial for accurate clinical decision-making. This aligns with our findings that pipelines such as Exatype and Stanford maintain high specificity across different conditions, reinforcing the importance of careful optimization to minimize false positives and ensure accurate mutation detection.

DeepChek relies on the Burrows-Wheeler alignment algorithm, a widely used method for short read alignment. This algorithm constructs a data structure called the Burrows-Wheeler Transform, which allows for efficient alignment by indexing the reference genome. While efficient, the sensitivity of DeepChek's specificity to parameter adjustments, particularly in the absence of quality score filtering, highlights potential challenges in accurately detecting mutations under certain settings (Ávila-Ríos *et al.*, 2020; Ji, Sandstrom, *et al.*, 2020). Variations in specificity observed, such as the drastic decrease in specificity for mutations like K65R and K70R in the Winnipeg settings, underscore the algorithm's susceptibility to parameter choices and its impact on mutation detection reliability.

In contrast, Exatype and HyDRA utilize the Bowtie alignment algorithm, which prioritizes speed and sensitivity in aligning short sequencing reads to a reference genome. Bowtie employs an indexing technique to efficiently search for exact matches between reads and the reference, enabling rapid alignment even in large genomic datasets (Langmead and Salzberg, 2012). Bowtie's ability to handle diverse sequence data efficiently likely contributes to Hydra's better specificity in comparison across all mutations and thresholds in both settings. Bowtie achieves this through its efficient alignment algorithm, which is designed to manage a wide range of sequence variations, including those with gaps and errors. This robust handling of sequence data

ensures accurate alignment, thereby enhancing the specificity of mutation detection tasks. The consistent high specificity observed across mutations in both default and Winnipeg settings suggests the robustness and reliability of this alignment strategy, which may contribute to its consistent performance irrespective of parameter variations.

Bowtie combines the strengths of the full-text minute index with dynamic programming algorithms to achieve high speed, sensitivity, and accuracy. This combination is particularly beneficial for clinical applications where high levels of accuracy and reliability in mutation detection are crucial. The consistent performance of Bowtie in aligning sequencing reads, even with gaps caused by true insertions and deletions or sequencing errors, underscores its suitability for clinical use. This adaptability and efficiency in handling sequence data help ensure that tools like Hydra can achieve and maintain high specificity in mutation detection (Langmead and Salzberg, 2012).

PASeq introduces a unique approach to mutation detection with its reliance on an in-house variant caller that prioritizes coverage and read balance over quality scores. This approach diverges from traditional methods that heavily rely on quality scores for variant calling. The variable specificity observed across mutations, particularly in the Winnipeg settings, suggests that this strategy may benefit certain mutations more than others. Further exploration into the intricacies of this variant calling approach is warranted to fully understand its impact on mutation detection accuracy.

Stanford adopts a full codon-based variant calling approach, which considers variations at the codon level rather than individual nucleotides. This approach allows for a more comprehensive assessment of mutations within coding regions, enhancing the accuracy and reliability of mutation detection. Combined with a minimum coverage requirement, Stanford achieves consistent perfect specificity across all mutations and thresholds in both settings, reflecting the high degree of reliability and accuracy conferred by its algorithmic strategies (Zook *et al.*, 2019; Koboldt, 2020).

The use of Bowtie2 in Exatype, HyDRA, MiCall, and PaSeq provides a balance of speed and sensitivity in aligning short reads, contributing to their robust specificity performance. Conversely, DeepChek's reliance on the Burrows-Wheeler Aligner and lack of quality score filtering can lead to higher sensitivity but lower specificity, particularly noticeable in the significant drop in specificity for mutations like K65R and

K70R under different settings. Additionally, strict quality score cut-offs (e.g., 30 for Exatype and HyDRA) enhance specificity by filtering out low-quality reads, whereas no QS filtering in DeepChek allows for more data inclusion but at the risk of increased noise. Minimum length cut-offs also vary, with longer cut-offs improving specificity by ensuring high confidence reads but potentially reducing sensitivity. Similarly, the minimum read depth requirement and amino acid frequency cut-offs influence the accuracy of variant detection, with lower thresholds increasing sensitivity and higher thresholds improving specificity.

The introduction of the Winnipeg settings, which applied more stringent quality thresholds and lower consensus frequency cut-offs, further demonstrated how parameter optimization can modulate algorithmic behaviour across pipelines. Under these modified settings, pipelines such as HyDRA and PASeq showed improved concordance and sensitivity, indicating that their alignment and variant-calling modules benefit from enhanced depth and quality control. In contrast, DeepChek and MiCall displayed increased variability in specificity, suggesting that algorithmic designs lacking adaptive filtering mechanisms are more sensitive to changes in parameter inputs. These findings reinforce that discrepancies between pipelines are not solely a function of algorithm choice but arise from the interaction between pipeline design and user-defined analytical thresholds. Such differences underscore the importance of standardized parameter recommendations, like those outlined in the Winnipeg consensus, to promote reproducible HIVDR analyses across laboratories and surveillance programs (Ji et al., 2018; Barbitoff et al., 2022).

Understanding these nuances is crucial for selecting the appropriate pipeline and optimizing its settings for accurate HIV drug resistance mutation detection.

The observed discordance among pipelines can be attributed to a combination of algorithmic and parameter-driven factors (Zook et al., 2019; Olson et al., 2022). Pipelines using different read aligners, such as Bowtie2 or Burrows–Wheeler Aligner, apply varying scoring schemes and error-tolerance thresholds, which influence how mismatches and insertions are interpreted. Similarly, differences in variant-calling logic—whether codon-based, frequency-based, or coverage-weighted—affect the classification of low-frequency variants. These algorithmic distinctions collectively explain much of the observed variability in sensitivity and specificity across pipelines, underscoring the need for harmonized analytical standards to minimize discordant

HIVDR calls.

These findings correspond with the patterns observed in Section 3.4, where certain pipelines demonstrated higher sensitivity but slightly reduced specificity, indicating a tendency toward overcalling low-frequency variants. Such overcalling may inflate estimates of resistance prevalence, particularly at thresholds below 5%, potentially leading to premature regimen switching or unnecessary clinical concern. Conversely, pipelines with stricter variant filters may undercall genuine mutations, resulting in false negatives that obscure early indicators of emerging resistance. From a surveillance perspective, both outcomes carry important implications: overcalling can distort prevalence estimates of transmitted or acquired resistance, whereas undercalling may delay detection of critical resistance patterns. Balancing sensitivity and specificity through optimized variant thresholds is therefore essential for achieving reliable and actionable HIVDR surveillance outcomes (Zhou et al., 2023; Wensing et al., 2022; Ji et al., 2018).

4.5 Examining Minimum Prevalence Thresholds: Variations in Mutation Detection Across NGS Pipelines

The investigation into the minimum prevalence thresholds for detecting specific mutations in HIVDR testing through NGS pipelines provides valuable insights into the detection capabilities of these pipelines. Research in this area has highlighted challenges and opportunities associated with implementing NGS technology for HIVDR testing.

Analysis of this dataset showed very little variation in sensitivity, specificity, accuracy and linearity across the threshold settings of ≥2.00%, ≥5.00%, and ≥20.00%, with slightly lower values reported when using the ≥1.00% threshold. These findings suggest that all pipelines have the capability to accurately detect mutations below previously recommended thresholds of Sanger equivalent (≥20.00%), such as those proposed by the Winnipeg consensus (≥2.00%).

However, questions arise regarding the clinical relevance of mutations detected at these low frequencies. Clutter *et al.* (2016) and Inzaule *et al.* (2018) highlight the critical importance of defining clinically meaningful mutation detection thresholds to guide treatment decisions and support effective epidemiological monitoring in the context of

HIV management (Clutter *et al.*, 2016; Inzaule *et al.*, 2018). Whilst improving detection sensitivity is advantageous, it is necessary to exclude false-positive reads. Further research is also necessary to ascertain whether mutations identified at low frequencies necessitate changes in treatment regimens or have prognostic implications for patient outcomes. Leveraging ultrasensitive testing methods and consensus mutation data can provide valuable insights into mutation significance and guide personalized treatment strategies based on NGS data.

The study emphasizes the importance of investigating minimum prevalence thresholds in HIV drug resistance testing using NGS pipelines. Next-generation sequencing allows us to detect low-frequency mutations effectively, even in minority variants of the virus. These variants, present in less than 20% of the viral population, can impact patient care and treatment strategies. Ongoing research collaboration is essential for refining NGS pipelines, aligning detection thresholds, and ensuring clinical reliability of mutation detection. Standardizing thresholds further will optimize treatment strategies and patient outcomes.

Given the natural variability of HIV, researchers should consider that optimal threshold determination can be dependent on the type of mutation, its frequency, and the context of the surrounding sequence play crucial roles in influencing mutation detection accuracy, when analysing HIV genomic data (Jiang *et al.*, 2020).

4.5.1 Study Limitations and Future Research directions:

Although this study provides valuable insights into the performance of multiple HIVDR bioinformatics pipelines, several limitations should be acknowledged. Firstly, the analysis was based on secondary data from a single national surveillance dataset, which may not fully capture the variability of sample quality or viral subtypes observed in routine diagnostic settings. Secondly, the evaluation was limited to eight key drug resistance mutations, and therefore does not represent the full spectrum of resistance-associated mutations across all drug classes. Thirdly, the reference standard used, a median-derived comparative measure, served as a practical proxy but not an absolute biological truth. Finally, differences in pipeline accessibility and version updates at the time of analysis may have influenced specific performance outcomes. Despite these constraints, the study design and comparative framework provide a robust

foundation for evaluating inter-pipeline variability and identifying critical factors influencing HIVDR detection accuracy.

Future research should aim to extend this comparative framework across larger, more diverse datasets representing multiple subtypes and treatment cohorts. Experimental validation using known reference panels would help establish biological benchmarks for mutation detection accuracy. Additionally, systematic evaluation of emerging deep-learning-based variant callers and hybrid assembly tools could offer further refinement of HIVDR interpretation pipelines.

Collaborative efforts between developers, public health institutions, and global surveillance networks are essential to harmonize data processing standards and promote open-source benchmarking resources. Expanding standardization efforts will strengthen the clinical and epidemiological value of NGS-based HIVDR testing, particularly in resource-limited settings where consistency and cost-effectiveness are crucial.

4.5.2 Conclusion:

Taken together, these findings provide a comprehensive evaluation of six NGS pipelines (PASeq, HyDRA, MiCall, DeepChek HIV, Stanford HIVdb-NGS (beta) and Exatype) in the detection of HIV-1 drug resistance mutations using national surveillance data. While all pipelines demonstrated high accuracy and specificity, subtle yet important discrepancies were observed in the detection of low-frequency variants, particularly at thresholds below 5%. Pipelines incorporating stringent quality-score filtering and Bowtie-based alignment (such as HyDRA and Exatype) exhibited greater stability and specificity across both default and Winnipeg settings. In contrast, pipelines without adaptive filtering (e.g., DeepChek and MiCall) were more sensitive to parameter changes, occasionally yielding inconsistent mutation calls.

The analysis underscores that no single pipeline consistently outperforms others; rather, their performance depends on the interaction between algorithm design, reference mapping strategy, and user-defined parameters. Establishing standardized quality control metrics—such as minimum read depth, variant quality score thresholds, and consensus cut-offs—is therefore essential to enhance reproducibility and ensure reliable data interpretation across laboratories.

From a clinical and public health perspective, overcalling of minor variants can inflate estimates of resistance prevalence, leading to unnecessary regimen modifications, whereas undercalling genuine variants risks delayed detection of emerging resistance. Achieving a balance between sensitivity and specificity remains critical for both individual patient management and population-level surveillance.

Future efforts should prioritize harmonizing pipeline parameters, improving algorithm transparency, and developing benchmarking datasets reflective of subtype C viral diversity common in Southern Africa. Incorporating automated quality control alerts and adaptive thresholding would further promote consistency and robustness in HIV drug resistance detection.

This study demonstrates that while the Winnipeg-optimized settings marginally improved inter-pipeline concordance, broader standardization initiatives are needed to strengthen the reliability of NGS-based HIV drug resistance reporting. Such efforts will ensure that genomic surveillance continues to provide actionable insights for treatment optimization and public health policy.

Ultimately, standardizing pipeline parameters and quality control practices will enhance the reliability of HIVDR surveillance and strengthen treatment decision-making in real-world settings.

CHAPTER 5 REFERENCES

Alidjinou, E.K. *et al.* (2017) 'RNA and DNA Sanger sequencing versus next-generation sequencing for HIV-1 drug resistance testing in treatment-naive patients', *The Journal of antimicrobial chemotherapy*, 72(10), pp. 2823–2830. Available at: https://doi.org/10.1093/JAC/DKX232.

Armenia, D. *et al.* (2024) 'Comparison of Different HIV-1 Resistance Interpretation Tools for Next-Generation Sequencing in Italy', *Viruses*, 16(9), p. 1422. Available at: https://doi.org/10.3390/V16091422.

Ávila-Ríos, S. *et al.* (2020) 'Next-generation sequencing for HIV drug resistance testing: Laboratory, clinical, and implementation considerations', *Viruses*, 12(6). Available at: https://doi.org/10.3390/v12060617.

Bani Baker, Q. *et al.* (2020) 'Comprehensive comparison of cloud-based NGS data analysis and alignment tools', *Informatics in Medicine Unlocked*, 18, p. 100296. Available at: https://doi.org/10.1016/j.imu.2020.100296.

Barbitoff, Y.A. *et al.* (2022) 'Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery', *BMC Genomics*, 23(1), pp. 1–17. Available at: https://doi.org/10.1186/S12864-022-08365-3/FIGURES/5.

Bennett, D.E. *et al.* (2009) 'Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update', *PloS one*, 4(3). Available at: https://doi.org/10.1371/JOURNAL.PONE.0004724.

Blassel, L. *et al.* (2021) 'Drug resistance mutations in HIV: new bioinformatics approaches and challenges', *Current Opinion in Virology*, 51, pp. 56–64. Available at: https://doi.org/10.1016/j.coviro.2021.09.009.

Boender, T.S. *et al.* (2015) 'Long-term Virological Outcomes of First-Line Antiretroviral Therapy for HIV-1 in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 61(9), pp. 1453–1461. Available at: https://doi.org/10.1093/CID/CIV556.

Carr, A. *et al.* (2023) ‘HIV drug resistance in the era of contemporary antiretroviral therapy: A clinical perspective’, *Antiviral therapy*, 28(5). Available at: https://doi.org/10.1177/13596535231201162.

Chen, S. *et al.* (2017) ‘AfterQC: Automatic filtering, trimming, error removing and quality control for fastq data’, *BMC Bioinformatics*, 18(Suppl 3). Available at: https://doi.org/10.1186/s12859-017-1469-3.

Chou, T.C., Maggirwar, N.S. and Marsden, M.D. (2024) ‘HIV Persistence, Latency, and Cure Approaches: Where Are We Now?’, *Viruses*, 16(7). Available at: https://doi.org/10.3390/V16071163.

Clutter, D.S. *et al.* (2016) ‘HIV-1 drug resistance and resistance testing’, *Infection,Genetics and Evolution*, 46, pp. 292–307. Available at: https://doi.org/10.1016/j.meegid.2016.08.031.

Donato, L. *et al.* (2021) ‘New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies’, *Neural Computing and Applications*, 33(22), pp. 15669–15692. Available at: https://doi.org/10.1007/s00521-021-06188-z.

Foley, B. *et al.* (2024) *HIV Circulating Recombinant Forms (CRFs)*, *Database*. Available at: https://www.hiv.lanl.gov/components/sequence/HIV/crfdb/crfs.comp (Accessed: 3 January 2025).

Gibson, R.M. *et al.* (2014) ‘Sensitive deep-sequencing-based HIV-1 genotyping assay to simultaneously determine susceptibility to protease, reverse transcriptase, integrase, and maturation inhibitors, as well as HIV-1 coreceptor tropism’, *Antimicrobial Agents and Chemotherapy*, 58(4), pp. 2167–2185. Available at: https://doi.org/10.1128/AAC.02710-13.

Gougeon, M.L. *et al.* (1998) ‘T Cell Apoptosis in HIV Infection: Mechanisms and Relevance for AIDS Pathogenesis’, *Results and problems in cell differentiation*, 24, pp. 233–248. Available at: https://doi.org/10.1007/978-3-540-69185-3_11.

Gu, W., Miller, S. and Chiu, C.Y. (2019) ‘Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection’, *Annu Rev Pathol.*, 14, pp. 319–338. Available at:

https://doi.org/10.1146/annurev-pathmechdis-012418-012751.

Hakimzadeh, A. *et al.* (2023) ‘A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses’, *Molecular Ecology Resources*, (July), pp. 1–17. Available at: https://doi.org/10.1111/1755-0998.13847.

Hemelaar, J. *et al.* (2019) ‘Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis’, *The Lancet Infectious Diseases*, 19(2), pp. 143–155. Available at: https://doi.org/10.1016/S1473-3099(18)30647-9.

Hirsch, H.H. *et al.* (2005) ‘Genotypic and phenotypic resistance testing of HIV-1 in routine clinical care’, *European Journal of Clinical Microbiology and Infectious Diseases*, 24(11), pp. 733–738. Available at: https://doi.org/10.1007/S10096-005-0044-4/METRICS.

Hirsch, M.S. *et al.* (1998) ‘Antiretroviral drug resistance testing in adults with HIV infection: implications for clinical management. International AIDS Society--USA Panel’, *JAMA*, 279(24), pp. 1984–1991. Available at: https://doi.org/10.1001/JAMA.279.24.1984.

Hu, T. *et al.* (2021) ‘Next-generation sequencing technologies: An overview’, *Human immunology*, 82(11), pp. 801–811. Available at: https://doi.org/10.1016/J.HUMIMM.2021.02.012.

Inzaule, S.C. *et al.* (2017) ‘The Evolving Landscape of HIV Drug Resistance Diagnostics for Expanding Testing in Resource-Limited Settings.’, *AIDS Reviews*, 19(4), pp. 219–230.

Inzaule, S.C. *et al.* (2018) ‘Clinically relevant thresholds for ultrasensitive HIV drug resistance testing: a multi-country nested case-control study’, *The lancet. HIV*, 5(11), pp. e638–e646. Available at: https://doi.org/10.1016/S2352-3018(18)30177-2.

Ji, H. *et al.* (2018) ‘Bioinformatic data processing pipelines in support of next-generation sequencing-based HIV drug resistance testing: the Winnipeg Consensus’, *Journal of the International AIDS Society*, 21(10), pp. 1–14. Available at: https://doi.org/10.1002/jia2.25193.

Ji, H., Sandstrom, P., *et al.* (2020) ‘Are We Ready for NGS HIV Drug Resistance Testing? The Second “Winnipeg Consensus” Symposium’, *Viruses*, 12(6), p. 586. Available at: https://doi.org/10.3390/V12060586.

Ji, H., Parkin, N., *et al.* (2020) ‘External Quality Assessment Program for Next-Generation Sequencing-Based HIV Drug Resistance Testing: Logistical Considerations’, *Viruses*, 12(5), p. 556. Available at: https://doi.org/10.3390/V12050556.

Jiang, C. *et al.* (2020) ‘A unique viral reservoir landscape in HIV-1 elite controllers’, *Nature*, 585(7824), p. 261. Available at: https://doi.org/10.1038/S41586-020-2651-8.

Kanyerezi, S. *et al.* (2024) ‘HIV-DRIVES: HIV drug resistance identification, variant evaluation, and surveillance pipeline’, *Access microbiology*, 6(7). Available at: https://doi.org/10.1099/ACMI.0.000815.V3.

Kanzi, A.M. *et al.* (2020) ‘Next-generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance’, *Frontiers in Genetics*, 11(October), pp. 1–18. Available at: https://doi.org/10.3389/fgene.2020.544162.

Kayani, M.U.R. *et al.* (2021) ‘Genome-resolved metagenomics using environmental and clinical samples’, *Briefings in Bioinformatics*, 22(5), pp. 1–20. Available at: https://doi.org/10.1093/bib/bbab030.

Kim, J., Ji, M. and Yi, G. (2020) ‘A review on sequence alignment algorithms for short reads based on next-generation sequencing’, *IEEE Access*, 8, pp. 189811–189822. Available at: https://doi.org/10.1109/ACCESS.2020.3031159.

Koboldt, D.C. (2020) ‘Best practices for variant calling in clinical sequencing’, *Genome Medicine*, 12(1), pp. 1–13. Available at: https://doi.org/10.1186/s13073-020-00791-w.

Kumar, K.R., Cowley, M.J. and Davis, R.L. (2019) ‘Next-Generation Sequencing and Emerging Technologies’, *Seminars in Thrombosis and Hemostasis*, 45(7), pp. 661–673. Available at: https://doi.org/10.1055/s-0039-1688446.

Lahens, N.F. *et al.* (2017) ‘A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression’, *BMC genomics*, 18(1), p. 602. Available at: https://doi.org/10.1186/s12864-017-4011-0.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods 2012 9:4*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Lee, E.R. *et al.* (2020) 'Performance comparison of next-generation sequencing analysis pipelines for HIV-1 drug resistance testing', *Scientific Reports*, 10(1634), pp. 1–10. Available at: https://doi.org/10.1038/s41598-020-58544-z.

Lee, J.H. *et al.* (2013) 'Restriction fragment mass polymorphism (RFMP) analysis based on MALDI-TOF mass spectrometry for detecting antiretroviral resistance in HIV-1 infected patients', *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(6). Available at: https://doi.org/10.1111/1469-0691.12167.

Li, G. and Clercq, E. De (2016) 'HIV Genome-Wide Protein Associations: a Review of 30 Years of Research', *Microbiology and Molecular Biology Reviews : MMBR*, 80(3), p. 679. Available at: https://doi.org/10.1128/MMBR.00065-15.

Liu, T.F. and Shafer, R.W. (2006) 'Web resources for HIV type 1 genotypic-resistance test interpretation', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 42(11), pp. 1608–1618. Available at: https://doi.org/10.1086/503914.

Liu, X. *et al.* (2013) 'Variant Callers for Next-Generation Sequencing Data: A Comparison Study', *PLoS ONE*, 8(9), pp. 1–11. Available at: https://doi.org/10.1371/journal.pone.0075619.

Lo, C.C. and Chain, P.S.G. (2014) 'Rapid evaluation and quality control of next-generation sequencing data with FaQCs', *BMC Bioinformatics*, 15(1), pp. 1–8. Available at: https://doi.org/10.1186/s12859-014-0366-2.

MacConaill, L.E. *et al.* (2018) 'Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing', *BMC Genomics*, 19(1), pp. 1–10. Available at: https://doi.org/10.1186/s12864-017-4428-5.

Maertens, G.N., Engelman, A.N. and Cherepanov, P. (2022) 'Structure and function of

retroviral integrase', *Nature reviews. Microbiology*, 20(1), pp. 20–34. Available at: https://doi.org/10.1038/S41579-021-00586-9.

Magi, A. *et al.* (2010) 'Bioinformatics for next-generation sequencing data', *Genes*, 1(2), pp. 294–307. Available at: https://doi.org/10.3390/genes1020294.

Manyana, S. *et al.* (2021) 'Hiv-1 drug resistance genotyping in resource-limited settings: Current and future perspectives in sequencing technologies', *Viruses*, 13(6), pp. 1–9. Available at: https://doi.org/10.3390/v13061125.

Martín, V. *et al.* (2016) 'An Efficient Microarray-Based Genotyping Platform for the Identification of Drug-Resistance Mutations in Majority and Minority Subpopulations of HIV-1 Quasispecies', *PloS one*, 11(12). Available at: https://doi.org/10.1371/JOURNAL.PONE.0166902.

Masenga, S.K. *et al.* (2023) 'HIV-Host Cell Interactions', *Cells*, 12(10). Available at: https://doi.org/10.3390/CELLS12101351.

Menéndez-Arias, L., Sebastián-Martín, A. and Álvarez, M. (2017) 'Viral reverse transcriptases', *Virus research*, 234, pp. 153–176. Available at: https://doi.org/10.1016/J.VIRUSRES.2016.12.019.

Moyo, S. *et al.* (2020) 'HIV drug resistance profile in South Africa: Findings and implications from the 2017 national HIV household survey', *PLoS ONE*, 15(11 November), pp. 1–13. Available at: https://doi.org/10.1371/journal.pone.0241071.

Musumeci, D., Riccardi, C. and Montesarchio, D. (2015) 'G-Quadruplex forming oligonucleotides as Anti-HIV agents', *Molecules*, 20(9), pp. 17511–17532. Available at: https://doi.org/10.3390/molecules200917511.

Ndashimye, E., Reyes, P.S. and Arts, E.J. (2023) 'New antiretroviral inhibitors and HIV-1 drug resistance: more focus on 90% HIV-1 isolates?', *FEMS microbiology reviews*, 47(1). Available at: https://doi.org/10.1093/FEMSRE/FUAC040.

Oliver, G.R., Hart, S.N. and Klee, E.W. (2015) 'Bioinformatics for clinical next-generation sequencing', *Clinical Chemistry*, 61(1), pp. 124–135. Available at: https://doi.org/10.1373/clinchem.2014.224360.

Olson, N.D. *et al.* (2022) 'PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions', *Cell Genomics*, 2(5), p. 100129. Available at: https://doi.org/10.1016/J.XGEN.2022.100129.

Pabinger, S. *et al.* (2014) 'A survey of tools for variant analysis of next-generation genome sequencing data', *Briefings in Bioinformatics*, 15(2), pp. 256–278. Available at: https://doi.org/10.1093/bib/bbs086.

Panpradist, N. *et al.* (2019) 'OLA-Simple: A software-guided HIV-1 drug resistance test for low-resource laboratories', *EBioMedicine*, 50, p. 34. Available at: https://doi.org/10.1016/J.EBIOM.2019.11.002.

Parikh, U.M. *et al.* (2017) 'Future technologies for monitoring HIV drug resistance and cure', *Current opinion in HIV and Aids*, 12(2), pp. 182–189. Available at: https://doi.org/10.1097/COH.0000000000000344.Future.

Parkin, N. *et al.* (2023) 'Need assessment for HIV drug resistance testing and landscape of current and future technologies in low- and middle-income countries', *PLOS global public health*, 3(10). Available at: https://doi.org/10.1371/JOURNAL.PGPH.0001948.

Parkin, N.T. *et al.* (2020) 'Multi-Laboratory Comparison of Next-Generation to Sanger-Based Sequencing for HIV-1 Drug Resistance Genotyping', *Viruses*, 12(7). Available at: https://doi.org/10.3390/V12070694.

Peng, Y. *et al.* (2023) 'Current drugs for HIV-1: from challenges to potential in HIV/AIDS', *Frontiers in Pharmacology*, 14(October), pp. 1–13. Available at: https://doi.org/10.3389/fphar.2023.1294966.

Pereira, R., Oliveira, J. and Sousa, M. (2020) 'Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics', *Journal of Clinical Medicine*, 9(1). Available at: https://doi.org/10.3390/jcm9010132.

Pfeifer, S.P. (2017) 'From next-generation resequencing reads to a high-quality variant data set', *Heredity*, 118(2), pp. 111–124. Available at: https://doi.org/10.1038/hdy.2016.102.

Quail, M.A. *et al.* (2012) 'A tale of three next-generation sequencing platforms:

comparison of Ion Torrent , Pacific Biosciences and Illumina MiSeq sequencers', *BMC Genomics*, 13(1), p. 1. Available at: https://doi.org/10.1186/1471-2164-13-341.

Sever, B. *et al.* (2024) 'A Review of FDA-Approved Anti-HIV-1 Drugs, Anti-Gag Compounds, and Potential Strategies for HIV-1 Eradication', *International Journal of Molecular Sciences 2024, Vol. 25, Page 3659*, 25(7), p. 3659. Available at: https://doi.org/10.3390/IJMS25073659.

Sezerman, O.U. *et al.* (2019) 'Bioinformatics Workflows for Genomic Variant Discovery, Interpretation and Prioritization', *Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations* [Preprint]. Available at: https://doi.org/10.5772/INTECHOPEN.85524.

Shafer, R.W. (2002) 'Genotypic testing for human immunodeficiency virus type 1 drug resistance', *Clinical Microbiology Reviews*, 15(2), pp. 247–277. Available at: https://doi.org/10.1128/CMR.15.2.247-277.2002.

Sharp, P.M. and Hahn, B.H. (2011) 'Origins of HIV and the AIDS pandemic', *Cold Spring Harbor Perspectives in Medicine*, 1(1), pp. 1–22. Available at: https://doi.org/10.1101/cshperspect.a006841.

Singh, Y. (2017) 'Machine Learning to Improve the Effectiveness of ANRS in Predicting HIV Drug Resistance', *Healthcare Informatics Research*, 23(4), p. 271. Available at: https://doi.org/10.4258/HIR.2017.23.4.271.

SoRelle, J.A., Wachsmann, M. and Cantarel, B.L. (2020) 'Assembling and Validating Bioinformatic Pipelines for Next-Generation Sequencing Clinical Assays', *Archives of pathology & laboratory medicine*, 144(9), pp. 1118–1130. Available at: https://doi.org/10.5858/ARPA.2019-0476-RA.

South African National Department of Health (2019) *2019 ART Clinical Guidelines for the Management of HIV in Adults, Pregnancy, Adolescents, Children, Infants and Neonates*, *South African HIV clinicians Society*. Pretoria. Available at: https://www.health.gov.za/wp-content/uploads/2020/11/2019-art-guideline.pdf (Accessed: 3 January 2025).

Ssekagiri, A. *et al.* (2022) 'QuasiFlow: a Nextflow pipeline for analysis of NGS-based

HIV-1 drug resistance data', *Bioinformatics Advances*, 2(1). Available at: https://doi.org/10.1093/BIOADV/VBAC089.

Steegen, K. *et al.* (2024) *National HIV Drug Resistance Surveillance Among Adults with Viral Non-Suppression in South Africa*. Johannesburg. Available at: https://www.nicd.ac.za/national-hiv-drug-resistance-surveillance-among-adults-with-viral-non-suppression-in-south-africa/ (Accessed: 26 December 2024).

Tang, M.W., Liu, T.F. and Shafer, R.W. (2012) 'The HIVdb System for HIV-1 Genotypic Resistance Interpretation', *Intervirology*, 55(2), pp. 98–101. Available at: https://doi.org/10.1159/000331998.

Temereanca, A. and Ruta, S. (2023) 'Strategies to overcome HIV drug resistance-current and future perspectives', *Frontiers in Microbiology*, 14, p. 1133407. Available at: https://doi.org/10.3389/FMICB.2023.1133407/BIBTEX.

UNAIDS (2024a) *Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic. | UNAIDS*, *https://www.unaids.org/en/resources/documents/2024/UNAIDS_FactSheet*. Available at: https://www.unaids.org/en/resources/documents/2024/UNAIDS_FactSheet (Accessed: 3 January 2025).

UNAIDS (2024b) *UNAIDS Country Fact Sheets | South Africa | 2023*, *https://www.unaids.org/en/regionscountries/countries/southafrica*. Available at: https://www.unaids.org/en/regionscountries/countries/southafrica (Accessed: 26 December 2024).

Vercauteren, J. *et al.* (2013) 'Clinical Evaluation of Rega 8: An Updated Genotypic Interpretation System That Significantly Predicts HIV-Therapy Response', *PLoS ONE*, 8(4), p. e61436. Available at: https://doi.org/10.1371/JOURNAL.PONE.0061436.

Weber, I.T., Wang, Y.F. and Harrison, R.W. (2021) 'HIV Protease: Historical Perspective and Current Research', *Viruses*, 13(5), p. 839. Available at: https://doi.org/10.3390/V13050839.

Woodham, A.W. *et al.* (2016) 'Human Immunodeficiency Virus Immune Cell Receptors, Coreceptors, and Cofactors: Implications for Prevention and Treatment',

AIDS patient care and STDs, 30(7), pp. 291–306. Available at: https://doi.org/10.1089/APC.2016.0100.

World Health Organization (2018) *Updated recommendations on first-line and second-line antiretroviral regimens and post-exposure prophylaxis and recommendations on early infant diagnosis of HIV*. Available at: https://www.who.int/publications/i/item/WHO-CDS-HIV-18.51 (Accessed: 3 January 2025).

World Health Organization (2019) *Update of recommendations on first- and second-line antiretroviral regimens*. Available at: https://www.who.int/publications/i/item/WHO-CDS-HIV-19.15 (Accessed: 3 January 2025).

World Health Organization (2021a) *Consolidated guidelines on HIV prevention, testing, treatment, service delivery and monitoring: recommendations for a public health approach*. Available at: https://www.who.int/publications/i/item/9789240031593 (Accessed: 5 January 2025).

World Health Organization (2021b) *HIV drug resistance report 2021*, *https://www.who.int/publications/i/item/978924*. Available at: https://www.who.int/publications/i/item/9789240038608 (Accessed: 29 January 2025).

World Health Organization (2024) *HIV Drug Resistance - brief report 2024*. Geneva. Available at: https://www.who.int/publications/i/item/9789240086319 (Accessed: 26 December 2024).

Yao, Z. *et al.* (2020) 'Evaluation of variant calling tools for large plant genome re-sequencing', *BMC Bioinformatics*, 21(1), pp. 1–16. Available at: https://doi.org/10.1186/S12859-020-03704-1/FIGURES/6.

Yohe, S. and Thyagarajan, B. (2017) 'Review of clinical next-generation sequencing', *Archives of Pathology and Laboratory Medicine*, 141(11), pp. 1544–1557. Available at: https://doi.org/10.5858/arpa.2016-0501-RA.

Zook, J.M. *et al.* (2019) 'An open resource for accurately benchmarking small variant and reference calls', *Nature biotechnology*, 37(5), pp. 561–566. Available at: https://doi.org/10.1038/S41587-019-0074-6.

CHAPTER 6 APPENDIX 1

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/48 Miss Vibha Kana

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M210525

NAME: (Principal Investigator)	Miss Vibha Kana
DEPARTMENT:	Virology National Institute for Communicable Diseases
PROJECT TITLE:	Evaluation of HIV-Drug Resistance Next Generation Sequencing data analysis pipelines
DATE CONSIDERED:	28/05/2021
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Dr Gillian Hunt and Dr Cathrine Scheepers
APPROVED BY:	Dr CB Penny, Chairperson, HREC (Medical)
DATE OF APPROVAL:	24/06/2021

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on the Third Floor, Faculty of Health Sciences, Phillip Tobias Building, 28 Princess of Wales Terrace, Parktown, 2193, University of the Witwatersrand. I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions.Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to the Committee. **I agree to submit a yearly progress report.** The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **May** and will therefore be due in the month of **May** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES