

**Molecular dynamic simulation studies of the South African HIV-1
Integrase subtype C protein to understand the structural impact of
naturally occurring polymorphisms**



**UNIVERSITY of the
WESTERN CAPE**

Darren Matthew Isaacs

**A thesis submitted in fulfilment of the requirements for the degree of
Master of Science (Bioinformatics), at the South African National
Bioinformatics Institute, University of the Western Cape**

Supervisor: Dr Ruben Earl Ashley Cloete¹

Co-supervisor: Dr Graeme Brendon Jacobs²

**1. Lecturer and Research Scientist South African Medical Research Council Bioinformatics Unit, South African
National Bioinformatics Institute, University of the Western Cape, Cape Town 8000, South Africa**

**2. Senior Lecturer and Research Scientist Division of Medical Virology, Department of Pathology, Faculty of
Medicine and Health Sciences, Stellenbosch University, Francie van Zijl Avenue, P.O. Box 241, Cape Town 8000**

Acknowledgements

I would firstly like to acknowledge and thank my supervisor Dr Ruben Cloete for having given me the opportunity to pursue a master's degree under his supervision. Under his eye, I have developed new skills and strengthened as an independent problem solver in scientific research. I am grateful for the "firsts" that I experienced during this masters project under Dr Ruben and with SANBI, memorably I for the first time visited the city of Durban and gave my first presentation at a conference. I want to express my appreciation to my Co-Supervisor Dr Graeme Brendon Jacobs and in particular thank him for his encouragement over the course of my masters. I thank my colleagues Rumbidzai Chitongo and Maryam Hassan for their friendship, assistance and at times much needed conversation as welcome distractions before focusing again. I would like to thank and show appreciation to my friends, who started this journey of science with me all the way back in 2014 as first year students, thank you for the great support and all the laughs shared. I would like to thank my Stellenbosch colleague's Dr Emmanuel Obasa and Dr Mikasi for all their input and constructive criticism. I would like to thank my landlady Mrs Carnow for her kindness, encouragement and understanding of my late nights during the course of my masters. I would like to thank Prof Alan Christoffels for the masters funding I received and the opportunity to join the SANBI team. Finally I appreciate the University of the Western Cape for how it has shaped me as a person over the years from teenager to now and I am grateful to every staff member who has lectured me and every admin member who assisted me, I am especially grateful to the lady who switched my intended field of study from first choice Bachelor of Arts to second choice Bachelor of Science during first year orientation, I have enjoyed this journey that her assistance has helped bring about.

Scientific Contributions

Publication(s):

1. **Isaacs, D.**, Mikasi, S. G., Obasa, A. E., Ikomey, G. M., Shityakov, S., Cloete, R., & Jacobs, G. B. (2020). Structural comparison of diverse HIV-1 subtypes using Molecular Modelling and Docking analyses of Integrase inhibitors. *Viruses*, **12(9)**, 936.
2. Mikasi, S. G., **Isaacs, D.**, Ikomey, G. M., Shimba, H., Cloete, R., & Jacobs, G. B. (2021). HIV-1 Drug Resistance Mutation Analyses of Cameroon-Derived Integrase Sequences. *AIDS Research and Human Retroviruses*, **37(1)**, 54-56.
3. Mikasi, S. G., **Isaacs, D.**, Chitongo, R., Ikomey, G. M., Jacobs, G. B., & Cloete, R. (2021). Interaction analysis of statistically enriched mutations identified in Cameroon recombinant subtype CRF02_AG that can influence the development of Dolutegravir drug resistance mutations. *BMC Infectious Diseases*, **21(1)**, 1-12.

Conferences attended and Presentations:

1. 9th SAAIDS CONFERENCE 2019 (11 – 14 June)

Title: *In-Silico Structural Comparison of HIV-1C and HIV-1B Integrase*

Co-authors: *Graeme Brendon Jacobs, Emmanuel Obasa, Ujjwal Neogi, Ruben Cloete*

2. Virology Africa 2020 (11 – 13 February)

Title: *Structural comparison of diverse HIV-1 subtypes using Molecular modelling and docking studies of integrase inhibitors*

Co-authors: *Graeme Brendon Jacobs, Ruben Cloete*

Abstract

The viral Integrase (IN) protein is an essential enzyme of all known retroviruses, including HIV-1. It is responsible for the insertion of viral DNA into the human genome. It is known that HIV-1 is highly diverse with a high mutation rate as evidenced by the presence of a large number of subtypes and even strains that have become resistant to antiretroviral drugs. It remains inconclusive what effect this diversity in the form of naturally occurring polymorphisms/variants exert on IN in terms of its function, structure and susceptibility to IN inhibitory antiretroviral drugs. South Africa is home to the largest HIV-1 infected population, with (group M) subtype C being the most prevalent subtype. An investigation into IN is therefore pertinent, even more so with the introduction of the IN strand-transfer inhibitor (INSTI) Dolutegravir (DTG). This study makes use of computational methods to determine any structural and DTG drug binding differences between the South African subtype C IN protein and the subtype B IN protein. The methods employed included homology modelling to predict a three-dimensional model for HIV-1C IN, calculating the change in protein stability after variant introduction and molecular dynamics simulation analysis to understand protein dynamics. Here we compared subtype C and B IN complexes without DTG and with DTG. Stability predictions showed that the variants/polymorphisms present in subtype C were negligible in effect upon protein stability. Molecular dynamics of the IN complexes without DTG confirmed that the proteins behave similarly. Simulations performed with the drug DTG indicated that both complexes stabilize with the drug DTG remaining bound. Minimal differences were calculated in the binding energies between each IN complex and DTG. Finally, PCA analysis indicated that DTG binding induces conformational changes that destabilize the protein structures of HIV-1C and 1B resulting in fewer local energy minima clusters. In summary, our main findings from this study showed that the wild type (WT) HIV-1 Subtype C IN derived from South Africa behaves similarly to the WT Subtype B IN even with the presence of naturally occurring polymorphisms and that DTG remains effectively bound to both IN subtypes.

Table of Contents

1.1 Introduction.....	1
1.2 HIV-1 Life Cycle.....	2
1.3 HIV-1 Integrase.....	5
1.3.1 HIV-1 Integrase Structure.....	6
1.3.2 IN Polymorphic variation within HIV-1 subtypes.....	7
1.4 Integrase function.....	8
1.5 Inhibiting integration.....	8
1.5.1 First generation INSTI's.....	9
.....	11
1.5.2 Second generation INSTI's.....	11
1.5.3 Non-catalytic site inhibitors.....	14
1.5.4 IN Drug Resistance.....	14
1.6 <i>in-silico</i> studies.....	16
1.6.1 Homology Modelling and Structure validation.....	17
1.6.2 Molecular dynamic simulations.....	20
1.6.3 Structural computational studies of HIV-1 IN.....	21
1.7 Rationale of present research work.....	23
Chapter 2: Methods and Materials.....	25
1.1 Data preparation.....	25
1.2 Consensus sequence generation.....	25
2.1 Protein modelling.....	26

2.2 Model Validation	27
2.3 Complex structure generation (IN-DNA-MG without or with DTG)	28
2.4 Structural Comparative Analysis	28
3. Molecular Dynamic simulations	29
3.1 Structural preparation of systems	29
3.2 CHARMM-GUI	30
3.3 Energy minimisation	30
3.4 Equilibration	31
3.5 Production simulation.....	31
3.6 MD analysis.....	31
Chapter 3: Results:.....	33
1.1 Protein Homology Modelling:	33
1.1.2 Protein Model Quality Assessment:	34
1.2 Variant stability change calculations:.....	35
1.3 Molecular dynamic simulations:.....	36
1.3.1 IN-DNA-MG complex	37
1.3.2 IN-DNA-MG-DTG complex.....	40
Chapter 4: Discussion and Conclusion	49
Supplementary Material.....	55
References.....	58
Appendices.....	78

1.1 Introduction

The viral Integrase (IN) protein is an essential enzyme and plays a critical role in the lifecycle of retroviruses. It is responsible for a defining feature of the retroviridae family of viruses, which is the integration of reverse transcribed viral DNA into the host cell genome (Lesbats, Engelman and Cherepanov, 2016). Understanding the molecular biology of IN is an important focus of ongoing medical research, due to the high prevalence of the Acquired Immune Deficiency Syndrome (AIDS) pandemic, particularly in developing countries. Human Immunodeficiency Virus (HIV) belongs to the retroviridae family and is the causative pathogenic agent of AIDS (Lesbats, Engelman and Cherepanov, 2016; Clavel et al., 1986).

South Africa (SA) is particularly burdened by HIV and AIDS. In the latest census studies conducted in 2018, an HIV prevalence of 13.06% was reported for the total population, with approximately 7,52 million people living with the disease (Stats SA, 2018). In 2017, 126 755 deaths were attributed to HIV/AIDS, a significant improvement over the year 2006 census, which had the highest HIV/AIDS related death rate at 345 185. The improvement is a consequence of the antiretroviral treatment (ART) programme that was nationally implemented. South Africa subsequently now boasts the largest HIV programme in the world (Stats SA, 2018). The latest recommendations published by the World Health Organisation (WHO) are that first line anti-viral regimens should include an IN strand-transfer inhibitor (INSTI) agent (World Health Organisation, 2016). South Africa is currently in the process of introducing Dolutegravir (DTG), a potent INSTI agent that may increase the effectiveness of the local HIV programme, while being predicted to significantly reduce costs as reported in a media statement by the health department in 2017 (Department of Health, 2017). The tender for DTG containing antiretroviral (ARV) treatment was awarded in late 2018 and as of August 2020 according to media reports more than 1 million patients have received DTG containing ARV treatment in SA (Treasury Department, 2018; Green, 2020).

DTG was developed in first world nations, where the prevailing HIV-1 variant is subtype B, whereas in South Africa HIV-1 group M subtype C is most prevalent, prompting us to investigate DTG efficacy against SA HIV-1 group M subtype C (Wainberg, 2004; Keyhani *et al.*, 2010). Recent studies have shown evidence for subtypes and naturally occurring

polymorphisms (NOPs) playing a role in drug resistance and failure prompting us to investigate this further (Lessells, Katzenstein and de Oliveira, 2012; Chehadeh *et al.*, 2017; Brado *et al.*, 2018; Rogers *et al.*, 2018). The following paragraphs gives a brief overview of the HIV-1 life cycle and in particular the structure and function of the HIV-1 IN protein as well as the use of computational tools to understand HIV-1 IN drug resistance and INSTI development.

1.2 HIV-1 Life Cycle

HIV-1 has a replicative cycle highly similar to other known retroviruses. The retroviral life cycle has been arbitrarily divided into two phases by researchers; namely the early phase consisting of cell binding, fusion, reverse transcription and integration, while the late phase is composed of transcription, assembly and budding (Nisole and Saib, 2004). Figure 1 graphically displays the steps of the HIV-1 life cycle which are further discussed below.

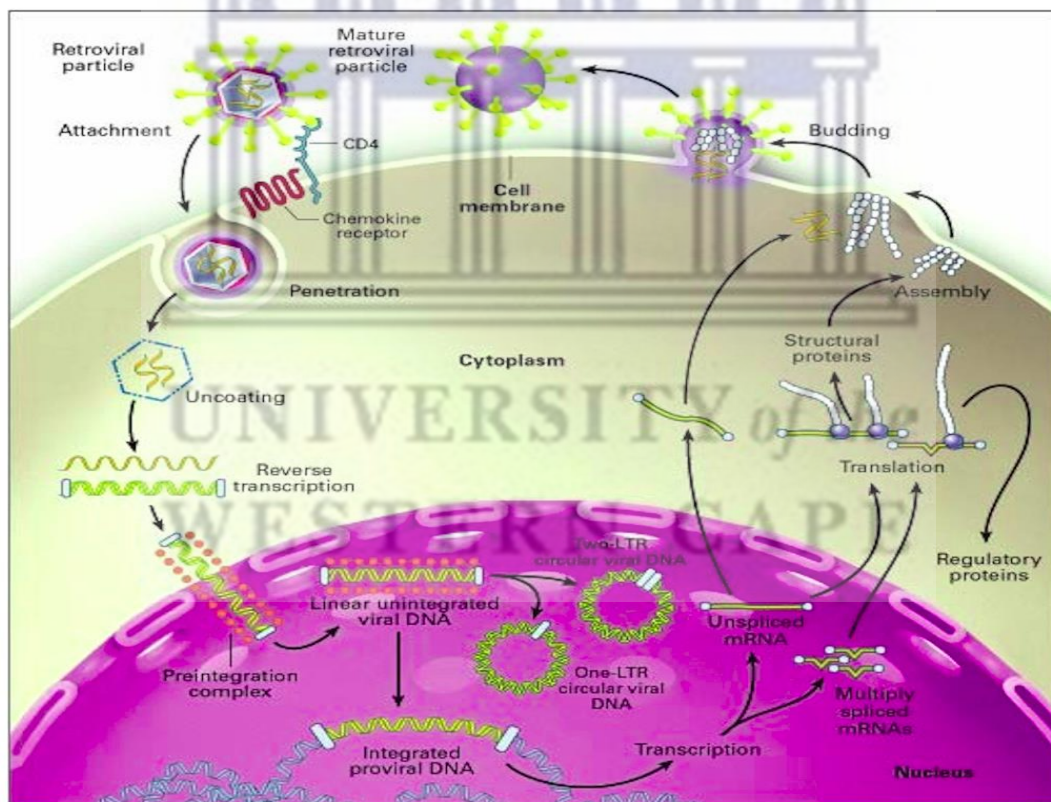


Figure 1: Essential steps of the HIV-1 replication cycle. The cycle initiates at the attachment step, where the viral particle attaches to the CD4 receptor on the cell wall and enters the cytoplasm. The next major step is reverse transcription of the viral RNA genome and formation of the Preintegration complex. This complex then enters the nucleus and subsequently integrates into the host cell DNA with cases of unsuccessful integration also taking place. The provirus DNA is then transcribed and translation of viral proteins then occur followed by viral particle assembly. The final step is the budding process whereby the viral particle exits the cell and becomes fully mature with the acquiring of cellular surface proteins. Adapted from (Furtado, *et al.*, 1999).

Cell binding and fusion is the initial step, virions bind to the surface of the host cell. Binding occurs to the CD4+ T-cell host protein receptor and is facilitated by the viral protein Envelope (Env) (Wilensky, Tilton and Doms, 2012). Viral binding to CD4 induces subsequent attachment and conformational changes in host chemokine-receptors, which may be either CC chemokine receptor 5 (CCR5) or C-X-C chemokine receptor type 4 (CXCR4). Conformational changes allow for the cellular and viral membrane to come into contact and form a fusion pore through which the entry of the viral core into the cellular cytoplasm is facilitated (Wilensky, Tilton and Doms, 2012). The next step reverse transcription, is the generation of viral DNA from viral RNA. This process is primarily mediated by the viral Reverse Transcriptase (RT) protein. RT performs its function in different stages; first synthesis of a DNA strand from the viral RNA genome occurs, followed by cleavage of RNA from the resulting DNA-RNA duplex to form single stranded DNA (ssDNA), finally synthesis of complementary DNA is performed using the ssDNA as template (Hu and Hughes, 2012; Tekeste *et al.*, 2015). Completing the early phase is the process of integration. Integration is the insertion of the reverse transcribed viral DNA into the host genome. Integration is primarily mediated by the IN enzyme. The integration process is subdivided into two distinct successive steps. The first one is termed 3' processing, the ends of the viral DNA (vDNA) are subjected to endonucleolytic cleavage, in which a dinucleotide is removed from both ends, leaving exposed 3'- hydroxyl groups (Delelis *et al.*, 2008; Hare, Maertens and Cherepanov, 2012). The resulting processed viral DNA (vDNA) then undergoes the second step, referred to as the strand transfer reaction. The 3'- hydroxyl groups are used in a nucleophilic attack upon host DNA phosphodiester bonds and subsequently the proviral containing host genome is repaired by the host cell DNA repair mechanisms (Delelis *et al.*, 2008; Hare, Maertens and Cherepanov, 2012). The early phase is now complete and the provirus may remain latent, lysogenically replicating with the host or it can induce the lytic life cycle.

The transcribing of viral RNA from proviral DNA is the next step, if the lytic life cycle is induced. The proviral HIV genome encodes for the transcription factor proteins, Transcriptional activator (Tat) and Anti-repression trans-activator (Rev), however transcription is primarily performed by cellular RNA polymerase II. Transcription is initiated at the 3'-end repeat junction continuing up to the 5'-end repeat junction in the downstream long terminal repeat. The resulting transcribed RNA is spliced into varying lengths prior to transport to the cytoplasm for protein synthesis (Roebuck and Saifuddin, 1999; Friedrich *et al.*, 2011; Karn and Stoltzfus, 2012). The synthesized messenger RNA (mRNA) transcripts, including both spliced and un-spliced are transported to the cytoplasm for protein synthesis by host cell translation mechanisms. The full-length mRNA transcript is translated into the capsid forming Gag polyprotein and by a ribosomal frame shift the Gag-Pol polyprotein, serving additionally as the genome of resulting viral progeny (Karn and Stoltzfus, 2012). Partially spliced mRNA transcripts produce the viral proteins Viral infectivity factor (Vif), Tat, Viral protein U (Vpu) and Envelope (Env) and fully spliced mRNA transcripts produce Viral protein R (Vpr), Tat, Rev and Negative Factor (Nef). Several transcripts may code for the same protein (de Breyne and Ohlmann, 2018). Once all viral components are synthesized and accumulated, assembly of new virions is possible, this is mediated primarily by the Gag polyprotein. The final step in the HIV-1 viral life cycle is virion release from the host cell, referred to as budding. The budding process of HIV-1 virions is facilitated by the endosomal sorting complexes required for transport (ESCRT) machinery. Budding may be sorted into two distinct processes the first is the envelopment of the virion by the cellular membrane and the second is the fission of the viral particle from the membrane. In the process HIV-1 acquires a cellular envelope which also contains a number of cellular proteins. The newly formed HIV-1 virions may now bind to CD4+ receptors on other immune cells and continue infection (Votteler and Sundquist, 2013).

Therapeutic agents have been developed which target and inhibit a number of steps within the HIV-1 replication cycle, collectively referred to as anti-retrovirals (ARVs). ARVs are divided into 6 distinct classes. These are nucleoside analogue Reverse Transcriptase inhibitors, non-nucleoside Reverse Transcriptase inhibitors, Integrase inhibitors, Protease inhibitors, Fusion inhibitors and finally co-receptor antagonists (Arts and Hazuda, 2012). It has by now been well established that therapy is more effective when more than 1 ARV class is combined resulting in combination anti-retroviral therapy (cART), which has since become standard (Holtzer and

Roland, 1999). The remainder of this review focuses solely on the HIV-1 IN protein structure and its inhibition via INSTIs.

1.3 HIV-1 Integrase

The polymerase gene (*pol*) encodes the Integrase (IN) protein and two other indispensable proteins, Reverse Transcriptase (RT) and Protease (PR). The IN protein is the principal enzyme involved in catalyzing the integration process. IN is a 32 kilodalton protein of 288 residue length, subdivided into 3 distinct domains (Figure 2); the N-terminal domain (NTD), catalytic core domain (CCD) and C-terminal domain (CTD) (Delelis *et al.*, 2008). IN represents a rational target for therapeutic intervention, due to its essential role in the HIV-1 replication cycle and with no known analogue in human cells (Nair, 2002).

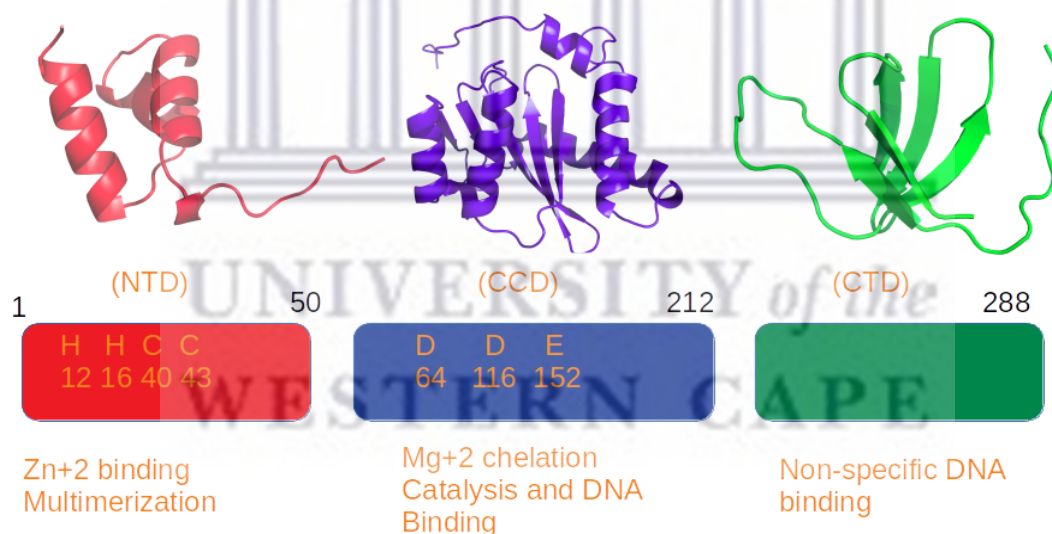


Figure 2: The different domains of the IN protein shown in cartoon depiction. In red is shown the N-terminal domain which is responsible for Zn ion binding at the HHCC motif and protein multimerization. In blue is shown the catalytic core domain which is involved in DNA binding at the DDE motif. In green is shown the C-terminal domain which contributes to non-specific DNA binding and IN-DNA stabilization. Figure produced by author generated using PyMol.

1.3.1 HIV-1 Integrase Structure

The function of HIV-1 IN has been studied before at the biochemical and cellular level. However, structural data has been restricted to only either a single domain, two domain or a homology model of the related Prototype Foamy Virus (PFV) IN protein. Biochemical studies performed on IN has shown it to consist of three functionally distinct domains. The N-terminal domain composes the first 1-49 amino acids, this domain contains a HHCC (His12, His16, Cys40, Cys43) motif, which is analogous to a zinc finger and is responsible for binding Zn^{2+} ions (Delelis *et al.*, 2008). Furthermore, the N-terminal domain is known to play a role in protein multimerization, and forms IN tetramers more readily when bound to Zn^{2+} ions (Zheng, Jenkins and Craigie, 1996; Delelis *et al.*, 2008). The catalytic core domain makes up amino acids 50-212. It includes the DDE motif, which is required for catalytic activity and is highly conserved among retroviruses. The DDE motif residues are located at positions 64, 116 and 152, respectively. The catalytic activity is dependent upon metallic cofactors, which is coordinated by residues D64 and D116 (Maignan *et al.*, 1998). The C-terminal domain encompasses residues 213-288, this domain is responsible for non-specific binding to DNA and therefore plays a role in stabilizing the resulting nucleoprotein complex (Chen *et al.*, 2000). IN multimerization is key to performing its catalytic function, to perform 3' processing, it assumes a dimeric form at the viral DNA molecule ends, the dimer pairs are then brought together to form a tetramer, it has been shown that tetramers will only form in the presence of viral DNA ends (Feng *et al.*, 2015).

Generating atomic resolution structures of the complete IN protein structure proved elusive due to poor solubility and the propensity for the protein to aggregate (Passos *et al.*, 2017). To resolve the issue of aggregation, it was found that engineering a fusion protein consisting of HIV-1 IN and binding at its N-terminus, Sso7D a small DNA binding protein derived from *Sulfolobus solfataricus* results in a hyperactive IN protein exhibiting characteristics, which make it more amenable to structure determining experimental techniques, such as Cryo-Electron microscopy (Li *et al.*, 2014; Passos *et al.*, 2017). Using the engineered fusion IN protein, the nearly complete HIV-1B intasome (tetrameric form in complex with DNA) (PDB

ID: 5U1C) structure was resolved using the Cryo-Electron microscopy method and was made publicly available (Passos *et al.*, 2017). The experimentally solved HIV-1 subtype B IN had a resolution of 3.9 Å, the lower the value the higher the confidence in the accuracy of the atomic ordering within the structure. Each protein chain for the tetrameric HIV-1 subtype B IN contained 12 helices comprised of 103 residues in total and 13 beta sheet strands comprising 55 residues in total (Passos *et al.*, 2017).

1.3.2 IN Polymorphic variation within HIV-1 subtypes

One study assessed the effect of polymorphisms on IN structure across 5 different subtypes (A, B, C, 01_AE, 02_AG). It was observed that different polymorphisms occurred at different frequencies across the various HIV-1 subtype IN structures. IN protein 3D structure models were generated for each subtype. It was found that some polymorphisms alter structural properties and thereby possibly exert an effect upon IN structure and viral DNA binding, as well as potentially upon drug binding propensity (Rogers *et al.*, 2018). The majority of HIV research reported has focused on subtype B, with comparatively limited investigation of other subtypes. It is thought the high genetic variability of HIV-1 may impact the functional efficiency of IN across the various subtypes, circulating recombinant forms (CRFs) and polymorphic varieties, although the amino acid level variability between the different IN's are relatively low at between 8-12% (Bar-Magen *et al.*, 2009; Llácer Delicado, Torrecilla and Holguín, 2016). The most common subtype of HIV-1 is subtype C accounting for approximately 50% of global infections it is therefore particularly important to determine if this variability has an impact on the IN structure and function of subtype C (Llácer Delicado, Torrecilla and Holguín, 2016; Gartner *et al.*, 2020).

A study conducted by Bar-Magen et al in (2009), biochemically compared subtype B and C Integrases. Time course experimental assays were conducted showing that 3' processing and strand transfer activity are comparable (Bar-Magen *et al.*, 2009). HIV-1 group O has been shown to have comparable 3' processing ability and reduced strand transfer activity when compared with HIV-1 group M subtype B (Depatureaux *et al.*, 2014). Group O is a rare variant characterised by a high number of polymorphisms, two polymorphisms namely, 74I and 153A

were implicated in this study as being likely responsible for the observed effect of reduced strand transfer activity by molecular modelling (Depatureaux *et al.*, 2014)

1.4 Integrase function

Integration occurs via two distinct sequential steps, namely 3' processing and strand transfer. Both reactions make use of a metallic cofactor, which are Mg²⁺ ions, however *in-vitro* manganese ions have been shown to be viable for both of these reactions. The integration process has been reproduced *in-vitro* by using short double-stranded oligonucleotides as viral DNA substitutes allowing for their study and investigation (Delelis *et al.*, 2008). Three prime 3' end processing commences when the IN dimer proteins bind to each end of the viral DNA. These ends are called long terminal repeats (LTR). The LTR's end with a CAGT segment, which is recognized by IN. An endonucleotide cleavage reaction is catalyzed by the IN dimers in which a GT dinucleotide is removed from either end exposing CA with 3' hydroxyl ends (Delelis *et al.*, 2008). Strand transfer occurs simultaneously at both ends of the viral DNA, with a five base pair offset between the two points of insertion. The reaction is a single-step transesterification in which the exposed 3' hydroxyl ends generated by 3' processing disrupt the phosphodiester bond on host DNA by nucleophilic attack. Removing a 5' dinucleotide overhang and ligating the viral DNA into the host DNA, host cell repair mechanisms are utilized. It was previously thought that IN had polymerase and ligation activity, but this has not been confirmed (Delelis *et al.*, 2008).

1.5 Inhibiting integration

Initial attempts at developing viable IN inhibitors resulted in experimentation with peptides, nucleotides, DNA complexes and polyhydroxylated aromatic compounds as potential therapeutic IN inhibitors (Di Santo, 2014). These were derived from either natural products or drug design strategies, however, none of these small molecules were developed into an effective IN inhibitor due to issues such as either low antiviral activity or high cytotoxicity (Di Santo, 2014). The discovery of aryl diketo acids (DKA) and their derivatives, represented a major breakthrough (Hazuda, 2000; Di Santo, 2014). It was found that in nanomolar concentrations these compounds can inhibit integration, while leaving vDNA synthesis unaffected (Hazuda, 2000; Di Santo, 2014). The DKA's only inhibit the strand transfer reaction of integration if in the presence of viral DNA (Espeseth *et al.*, 2000). Their mechanism of

activity was to bind to the catalytic domain in competition with host DNA (Espeseth *et al.*, 2000). The inhibition was determined to be metal dependent (Marchand *et al.*, 2003). Furthermore, DKA's were found to only inhibit the strand transfer reaction of integration thus referred to as Integrase Strand Transfer Inhibitors (INSTI) (Marchand *et al.*, 2003). The DKA's were further developed and modified by pharmaceutical companies, with some entering clinical trials. A team by Merck developed DKA compounds with a di-hydroxy-pyrimidine-carboxylic moiety. These compounds exhibited low nanomolar concentration activity and good pharmacokinetics in preclinical trials. The developed lead compounds would result in the discovery of the first FDA approved INSTI Raltegravir (Summa *et al.*, 2008; Pace *et al.*, 2007; Di Santo, 2014).

1.5.1 First generation INSTI's

Raltegravir (RAL) was the first integrase inhibitor to be approved by the FDA, in 2007 (Hicks and Gulick, 2009). RAL is a derivative of the original DKAs which showed potential during development. RAL is a pyrimidine carboxamide with the chemical formula $C_{20}H_{21}FN_6O_5$ (Anker and Corales, 2008; Summa *et al.*, 2008; Hicks and Gulick, 2009). Figure 3 depicts the chemical structure of RAL. The inhibition activity of RAL depend on the interaction with the DDE motif within the catalytic core of IN via binding to divalent magnesium cations (Mouscadet and Tchertanov, 2009). Once RAL is bound to the IN catalytic site and present within the intasome/pre-integration complex, host DNA binding is blocked (Anker and Corales, 2008; Hazuda, 2000; Hicks and Gulick, 2009). Studies conducted regarding the pharmacodynamics of RAL *in vitro* found that RAL is 1000 times more specific for HIV-1 IN in comparison to other polymerase proteins. RAL was reported to inhibit HIV-1 IN with an IC_{50} of 10nmol/L, furthermore in SupT1 cells infected with HIV-1, 1 micromol/L of RAL decreased integration by 50-fold (Hicks and Gulick, 2009). The clinical trials which led to eventual FDA approval found RAL to exhibit potent antiviral activity, favourable tolerability, low drug-drug interactions and relatively convenient twice a day dosing (Grinsztejn *et al.*, 2007; Iwamoto *et al.*, 2008; Wenning *et al.*, 2008; Hanley *et al.*, 2009). Furthermore, the previously mentioned trials found that RAL can be effectively combined with other anti-retroviral drugs in a regimen for superior results, even in treatment experienced patients.

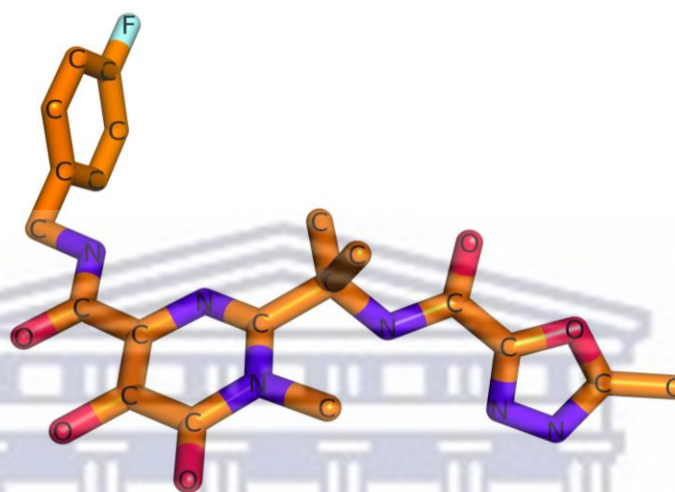


Figure 3: Chemical structure of Raltegravir. The first FDA approved INSTI. The full chemical name is *N*-(2-(4-(4-fluorobenzylcarbamoyl)-5-hydroxy-1-methyl-6-oxo-1,6-dihydropyrimidin-2-yl) propan-2-yl)-5-methyl-1,3,4-oxadiazole-2-carboxamide. A β -hydroxy-ketone structural motif is present and has

Elvitegravir (EVG) is the second first generation INSTI approved by the USA Food and Drug Administration (FDA). Like RAL, EVG was also developed with insights gained from the discovery and development of DKAs. EVG is a 4-quinolone-3-carboxylic acid with two functional groups that are coplanar (Shimura and Kodama, 2009) The chemical structure of EVG is displayed in Figure 4. The researchers found that a coplanar monoketo acid motif in 4-quinolone-3-carboxylic acid was a viable alternative to a diketo acid motif (Shimura and Kodama, 2009). EVG exhibits a 50% effective concentration in the nanomolar to sub nanomolar range, including activity against HIV isolates with known resistance to other anti-retroviral therapeutic classes (Shimura *et al.*, 2008). EVG inhibits the strand transfer reaction with minimal inhibition of the 3' processing step based on a previous report (Shimura *et al.*,

2008). EVG has been shown to exhibit 2 to 3 times greater potency in comparison to RAL, however it requires boosting with a CYP3A inhibitor (Marinello et al., 2008; Molina et al., 2012). The mechanism of activity is similar to that of RAL with EVG interacting with the DDE motif in the integrase core domain, also preferentially in the presence of viral DNA (Marinello et al., 2008).

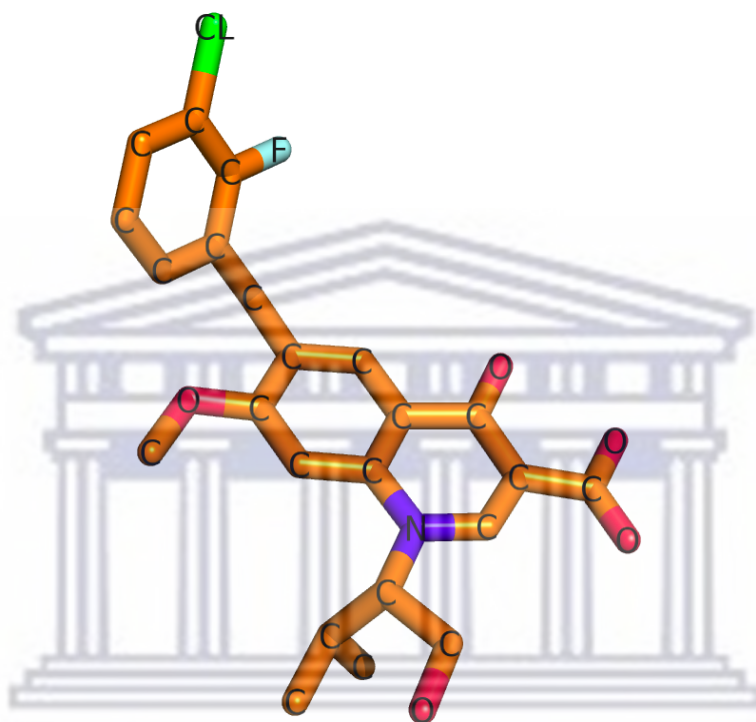


Figure 4: Chemical structure of Elvitegravir. The second FDA approved INSTI. The full chemical name is 6-[(3-chloro-2-fluorophenyl)methyl]-1-[(2S)-1-hydroxy-3-methylbutan-2-yl]-7-methoxy-4-oxoquinoline-3-carboxylic acid

1.5.2 Second generation INSTI's

The second generation of INSTI's are distinguished from the first generation of IN inhibitors due to having minimal cross resistance with first generation INSTI's resistance pathways, higher genetic barrier to resistance and an improved profile of efficacy and toxicity (Dow and Bartlett, 2014). Second generation INSTI's remain efficacious in first generation INSTI

treatment experienced patients who have shown resistance to first generation INSTI treatment because of the development of resistance mutation pathways (Castagna et al., 2014; Di Santo, 2014). DTG is the first FDA approved second generation INSTI (Figure 5). DTG like the first generation INSTI's target the strand transfer step of integration preferentially (Di Santo, 2014). It is highly potent against HIV integration, however of importance is that it has been found to remain effective against viral clones that contain RAL and EVG resistant mutations (Seki et al., 2015; Hare *et al.*, 2011; Di Santo, 2014). *In vitro* antiviral studies investigating the effect of combining DTG with other anti-retroviral classes of drugs found that DTG does not cause an increase in cytotoxicity and demonstrated synergism with EFV, nevirapine, stavudine, abacavir, lopinavir, amprenavir and enfuvirtide, with an additive effect when in combination with maraviroc (Kobayashi *et al.*, 2011; Di Santo, 2014). The ability of DTG to remain effective against RAL/EVG resistant HIV-1 strains has been attributed to DTG binding tighter to IN than RAL or EVG resulting in a higher genetic resistance barrier with it being able to subtly change its conformation in response to IN conformational changes (Hare *et al.*, 2011). The binding half-life of DTG is reported to be 71 hours which is significantly greater than RAL or EVG which are reported to have a binding half-life of 8.8 and 2.7 hours, respectively (Hightower et al., 2011; Hare *et al.*, 2011).



UNIVERSITY of the
WESTERN CAPE

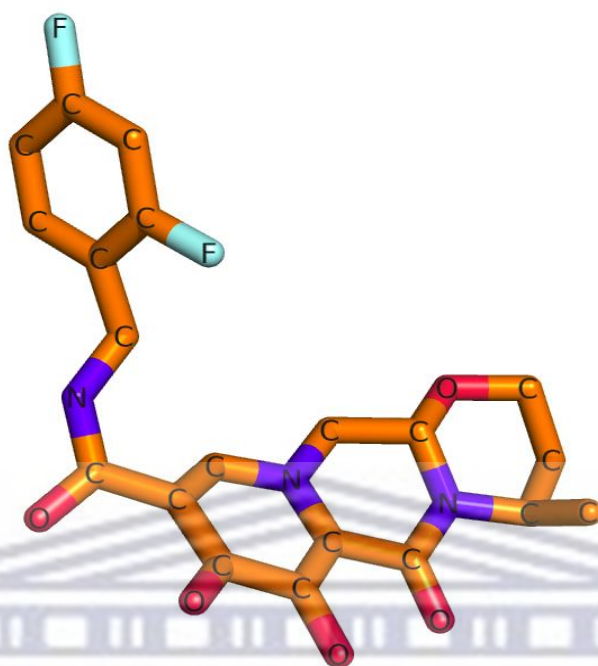


Figure 5: Chemical structure of Dolutegravir.
 Full chemical name is (3*S*,7*R*)-*N*-[(2,4-difluorophenyl)methyl]-11-hydroxy-7-methyl-9,12-dioxo-4-oxa-1,8

Bictegravir (BIC) and Cabotegravir (CBT) are the latest INSTI's with BIC having received FDA approval in 2018 and CBT currently in phase 3 clinical trials (Gulick, 2018). They are comparable with DTG as they possess a similar genetic resistance barrier. A comparative study found BIC to be more potent than DTG against viral isolates displaying resistance to first generation INSTI's and isolates which contain mutations which reduce DTG susceptibility. DTG, however, showed broader superiority against isolates containing known resistance mutations as compared with CBT (Smith *et al.*, 2018). CBT has a unique feature compared to the other INSTI's, the option of an injectable long acting form of CBT is possible, this may have major advantages particularly pertaining to patient adherence and usage for the prevention of HIV infection. One recent study investigated the viability of an injection dose once every 12 weeks (Murray *et al.*, 2018). The study found that while pain was experienced at the location of the injection site, most patients remained satisfied with CBT as an alternative to daily pre-

exposure prophylaxis (PrEP) (Murray *et al.*, 2018). A non-inferiority clinical trial of HIV-1 infected patients conducted over 96 weeks compared a CBT containing 2 drug regimen administered either every 4 or 8 weeks to a daily oral 3 drug regimen. The researchers found that viral suppression maintenance was comparable for the trial duration (Margolis *et al.*, 2017). At present DTG is the preferred INSTI for first line treatment, particularly for low to middle income nations. Inclusion of DTG into the South African AIDS treatment programme is currently in progress as of 2020 after initial delays due to side-effect concerns (Mendelsohn and Ritchwood, 2020).

1.5.3 Non-catalytic site inhibitors

All currently approved FDA INSTI's target the catalytic site and the strand transfer reaction. Allosteric inhibition is an alternative to active site inhibition and involves a ligand binding to a target enzyme at a location other than the active site (Monod *et al.*, 1965). Allosteric inhibitors are being developed that inhibit integration by inducing a protein conformational change(s) (Di Santo, 2014). A particularly promising target is the LEDGF/p75 a transcriptional co-activator (Di Santo, 2014). LEDGF/p75 serves as a co-factor for IN, responsible for the tethering and correct integration of the viral genome into the host genome. Therapeutic agents referred to as ledgins have been developed which bind to the inner core dimer interface of IN (Christ and Debyser, 2013). Many ledgins or ledgin analogues simultaneously inhibit the catalytic ability of IN allosterically and the LEDGF/p75-IN protein-protein interaction (Christ *et al.*, 2012; Kessl *et al.*, 2012; Tsiang *et al.*, 2012).

1.5.4 IN Drug Resistance

INSTI's while highly effective, have been shown to be vulnerable to treatment failure like the other anti-retroviral drug classes due to the development of drug induced resistance mutations. The first generation INSTI's RAL and EVG are known to fail during treatment due to drug resistance. Furthermore, they can cause cross resistance to one another (Van Wesenbeeck *et al.*, 2011). DTG, while it has been shown to select for resistance substitutions, remains effective (Anstett *et al.*, 2017). Drug pressure induced mutations are grouped into primary resistance substitutions and secondary resistance substitutions (Anstett *et al.*, 2017). Primary resistance substitutions arise in response to INSTI treatment and cause a decrease in drug susceptibility,

these substitutions mostly occur in the vicinity of the active site where INSTI binding would take place. Primary resistance substitutions/mutations often have an impact on viral fitness, these lead to compensatory or secondary resistance substitutions developing which alleviate the negative impacts of the primary resistance substitutions and may also further contribute to drug resistant activity (Delelis *et al.*, 2008; Hare *et al.*, 2010; Anstett *et al.*, 2017).

The three major and most common resistance mutation/substitution positions known at present are Q148H/R, N155H and Y143C/H/R with the latter only occurring in response to RAL treatment (Fransen *et al.*, 2009; Delelis *et al.*, 2010). Compensatory mutations which commonly co-occur with the major mutations have been found at positions; G140S, T97A and E92Q, the drug resistant effect is compounded when these additional substitutions occur alongside a major resistance mutation/substitution (Fransen *et al.*, 2009). Resistance against DTG has been reported both *in-vitro* and *in-vivo*, particularly for the mutations R263K and G118R. However, DTG maintains efficacy as only low level resistance is conferred and these mutations negatively influence viral fitness including decreased integration (Quashie *et al.*, 2012; Anstett *et al.*, 2017). DTG monotherapy failure in patients has been reported which contained the G118R mutation, this is a rare occurrence and researchers concluded the possibility that monotherapy and additional polymorphisms facilitated the acquisition of the mutation G118R (Brenner *et al.*, 2016).

Evidence suggesting that subtype differences and naturally occurring polymorphisms impact drug susceptibility remain inconclusive with further studies being required (Han, Mesplède and Wainberg, 2016). Polymorphisms which contribute to reduced drug susceptibility have been reported in treatment naïve patients but occur at a very low frequency in several populations (Lataillade, Chiarella and Kozal, 2007; Ambrosioni *et al.*, 2017; Chehadeh *et al.*, 2017). A study which conducted an analysis into the IN protein's natural variability, across 6706 INSTI treatment naïve patients, found that polymorphisms/mutations associated with INSTI resistance occur at a low rate of <2% for the tested subtypes B, C, D and recombinant CRF01_AE (Han, Mesplède and Wainberg, 2016; Llácer Delicado, Torrecilla and Holguín,

2016). Differing natural polymorphisms in the various subtypes may favour the development of different resistant pathways leading to varying levels of drug resistance (Bar-Magen *et al.*, 2010; Han, Mesplède and Wainberg, 2016).

A study conducted using 127 treatment naïve patients in SA, confirmed that HIV-1 subtype C is the most prevalent subtype in SA. Furthermore, an examination for mutations revealed that polymorphisms/mutations which confer <5-fold resistance occurred in about 7% of the sequences (Bessong and Nwobegahay, 2013; Han, Mesplède and Wainberg, 2016). A more recent study also focusing on a South African cohort, found no resistance associated mutations present prior to the anticipated introduction of DTG. Molecular modelling of the naturally occurring polymorphisms in the IN protein suggests that some polymorphisms may affect intasome complex stability and IN structure which in turn may have an effect on drug binding (Brado *et al.*, 2018).

1.6 *in-silico* studies

Computational approaches such as homology modelling, molecular docking and molecular dynamics have been very instrumental in studying the functional and structural details of HIV-1 IN. Furthermore, *in-silico* studies played a major role in the development of the first FDA approved INSTI RAL. In the following sections we introduce homology modelling and molecular dynamic simulation methods.

1.6.1 Homology Modelling and Structure validation

Homology modelling also referred to as comparative modelling, is a computational method used to generate a three dimensional (3D) structure of a protein (Cavasotto and Phatak, 2009). This method relies upon the observation that similar sequence implies similar structure but not always. It is possible to generate a reliable homologous structure from an experimentally solved structure that has a sequence similarity of 30% or greater (Cavasotto and Phatak, 2009). Figure 6 shows the various steps involved in the construction of a protein's 3D structure. It is an efficient and cost effective alternative in the event a 3D structure for a protein cannot be resolved experimentally. Homology modelling has played a major role in drug discovery and further refinement of experimentally solved structures (Cavasotto and Phatak, 2009).

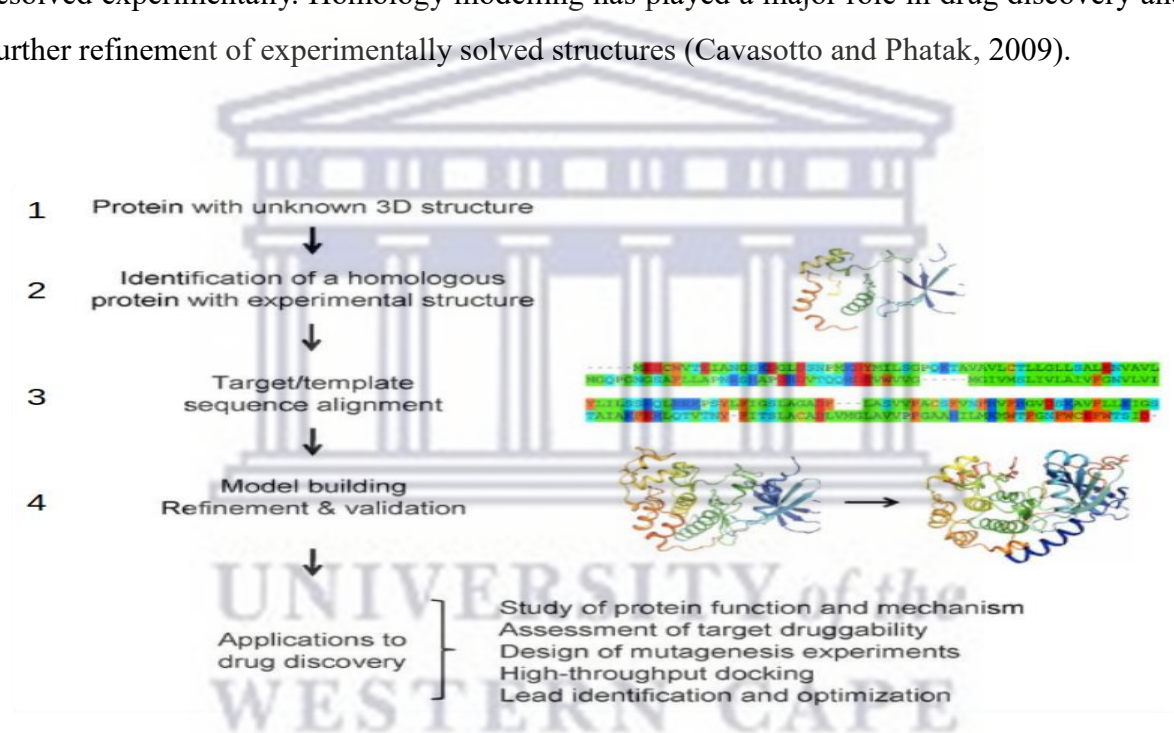


Figure 6: Process of homology modelling the workflow shows the steps involved. 1: Obtain target protein sequence of unknown structure. 2: Identification of a closely related protein sequence/s with an experimentally solved structure to the target sequence. 3: Alignment of the target sequence to the homologous protein sequence. 4: Building of the target sequence structure using the alignment between the target sequence and the homologous protein sequence. Adapted from (Cavasotto and Phatak, 2009)

To ensure the generated 3D protein model is of high quality, it is necessary to compare several parameters of the protein model to experimentally verified native proteins of high accuracy or resolution. A variety of software tools exist for this purpose, here I will briefly introduce three widely used quality tests namely Verify3D, PROCHECK and ProSA.

Verify3D software measures compatibility between the protein model and corresponding amino acid sequence, it does this by creating a 3D model (Eisenberg, Lüthy and Bowie, 1997). Each residue in the 3D model has its properties defined by its environment and gets represented by a row of 20 numbers in the profile. These numbers represent statistical preferences and are called 3D – 1D scores. The environmental parameters which affect residue properties are: area of the residue that is buried; fraction of side-chain covered by polar atoms and finally the local secondary structure. The 3D profile S score is the sum of all residue positions. The model is considered more accurate the greater the 3D profile S score is, with the quality thresholds being 0.2 per residue and the accuracy of a whole model being deemed reliable if over 80% of residues meet or surpass this threshold.

PROCHECK is a software which provides a highly detailed analysis of the stereochemistry of protein structures. It assesses the overall structural quality of the protein model and compares it with well refined structures and highlights regions of the structure which may warrant additional refinement (Laskowski *et al.*, 1993). The stereochemical parameters tested for, are derived from the work of Morris *et al* (1992) along with bond-length and bond angle data from the work of Engh and Huber (1991). PROCHECK is divided into 5 distinct programs, the first is “CLEAN.F”, responsible for fixing formatting issues and for adherence to standard convention (Laskowski *et al.*, 1993). The second program is called “SECTR.F”, this is responsible for the secondary structure assignment of individual residues and derived from a modified method of Kabsch and Sander (1983). The third program called “NB.C”, identifies all nonbonded interactions between the different pairs of residues. The non-bonded interaction is defined as the closest atom-atom contact between two residues within 4.0Å distance and atoms which are four or more bonds apart. The fourth program “ANGLN.F”, calculates all main-chain bond lengths and bond angles. The final program “PPLOT.F” is responsible for the final output. The PROCHECK final output consists of a series of plots together with a residue-

by-residue listing. The residue listing shows the calculated stereochemical value for each individual residue and their deviation from 'ideal' (Laskowski *et al.*, 1993). The main chain Ramachandran plot shows four quadrants divided by psi against phi dihedral angles and groups each individual residue into one of the quadrants dependent upon the actual dihedral angles calculated as compared with the expected dihedral angle for a given residue.

Protein Structure Analysis or ProSA. This program focuses on the usage of the Boltzmann's principle (Sippl, 1993). It places emphasis on determining the correct arrangement of the protein chains and the forces responsible for stabilizing native folds within solution spatially as opposed to assessing structural correctness via protein stereochemistry. To accomplish the intended goal, knowledge based mean fields are used to determine the stabilizing forces and energy distribution within protein structures (Sippl, 1993). The ProSA tool interrogates a precompiled database of known proteins and extracts the forces from the database using the Boltzmann's principle in the form of potentials of mean force. The potential of mean force is defined as a statistical average derived from the summation of amino acid pairwise contacts or distances. The forcefields of known or unknown proteins can then be obtained via recombination's of these potentials of mean force as a function of the amino acid sequence. Once the energy of a given structure has been evaluated, a Z-score is generated which indicates overall model quality. Obtained Z-scores outside the range characteristic of native structures are deemed to be erroneous, with a negative Z-score indicating that the result is below the expected mean of all possible values. A plot is generated which displays the obtained Z-score in comparison to all experimentally resolved native structures within the PDB (Wiederstein and Sippl, 2007). ProSA requires only carbon alpha atoms of a protein structure and may thus be used for structures in which refinement of other characteristics such as side-chain orientations are still required. Furthermore, a web-based version is available at (<https://prosa.services.came.sbg.ac.at/prosa.php>) (Wiederstein and Sippl, 2007).

1.6.2 Molecular dynamic simulations

Molecular dynamics simulations aim to predict the relative movement of atoms in space over time. The atomic motion of atoms in simulations are calculated by solving Newton's second law of motion (Durrant and McCammon, 2011; Gelpi *et al.*, 2015; Hollingsworth and Dror, 2018; Braun *et al.*, 2019) defined as Force equals mass times acceleration or ($F = ma$). Here the Forces involved are derived from non-bonded interactions and the bonded interactions. Bonded interactions: atomic angles, atomic bonds and dihedral angles are calculated using a set of equations as shown in Figure 7 (Durrant and McCammon, 2011; Gelpi *et al.*, 2015; Hollingsworth and Dror, 2018; Braun *et al.*, 2019). While non-bonded forces are the result of van der Waals interactions, which is modelled using the Lenard Jones 6-12 potential model and electrostatic interactions are derived from Coulomb's law. All of these parameters are collectively referred to as a forcefield (Durrant and McCammon, 2011; Gelpi *et al.*, 2015; Hollingsworth and Dror, 2018; Braun *et al.*, 2019). A forcefield is defined as a mathematical formula which calculates all forces exerting effects on atoms within a given simulation system. Different forcefields are parameterized differently, in that the magnitude and/or distance cut offs of every interaction/force may differ. The forcefield weightings used are fitted to quantum mechanical calculations and experimental results (Hollingsworth and Dror, 2018).

$$E_{total} = \underbrace{\sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2} [1 + \cos(n\phi - \gamma)]}_{\text{Bonded}} + \underbrace{\sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right]}_{\text{Non-bonded}}$$

Figure 7: Force Field Parameters

Shown are the physics derived equations for the determination of atomic motions, subdivided into bonded and non-bonded parameters. Bonded parameters are from left to right atomic bonds, atomic angles and dihedrals. Non-bonded parameters are made up of Electrostatic and Van der Waals forces.

Molecular dynamic simulations have been useful in studying the movement of biomolecular structures and elucidating the function of proteins. However, several limitations are known and these include the following. Firstly, computational processing power remains a barrier and has not improved to the point where simulations of protein folding/unfolding or ligand binding can be meaningfully elucidated. Secondly, molecular dynamic simulations do not account for covalent bond formation as these types of studies would require quantum mechanics simulations. Thirdly, while forcefields have substantially improved to the point where most are reproducibly accurate and align well with experimental results, they are inherently approximate as they are derived from classical mechanics and not quantum mechanics. Finally, a high quality simulation experiment requires high quality experimentally solved structures or homology models as a prerequisite (Hollingsworth and Dror, 2018).

1.6.3 Structural computational studies of HIV-1 IN

Homology modelling has been used to generate a 3D structure for HIV-1 integrase to facilitate drug development and understanding HIV-1 IN DNA and drug binding as well as drug resistance. Early studies exploited the homology shared between the bacterial transposon TN5 transposase protein and HIV-1 IN to generate a homology model of IN (Barreca *et al.*, 2006, 2007, p. 5). One such TN5 based homology model study found that residues comprising 140-149 represents a flexible catalytic loop responsible for stabilizing the integration complex by operating as a barrier between the two ends of the viral DNA, and they also showed that residues responsible for DNA binding are highly conserved (Wielens, Crosby and Chalmers, 2005). This flexible catalytic loop plays an important role in drug resistance with mutations reported at positions 140 and 148 which confer Raltegravir resistance (Dewdney *et al.*, 2013).

The experimentally resolved structure of the Prototype Foamy Virus (PFV) IN protein provided the first homologous structure from which HIV-1 IN protein structures could be predicted (Hare

et al., 2010, 2011). The IN active site in the catalytic core of each IN is nearly identical. As a result of this high sequence identity, it is possible to bind INSTI's to PFV IN, this enabled an accurate understanding of how INSTI's bind, their predicted binding mechanism to HIV-1 IN and the effect of resistance mutations upon this binding (Cutillas *et al.*, 2015; Quashie *et al.*, 2015). Homology protein models were also generated using domains of the IN proteins as the starting homologous structure from which to model (Wang, 2001). In one particular study, it was found that there is a dimer interface between the IN N-terminals (Wang, 2001).

Cryogenic electron microscopy (Cryo-EM), an alternative technique to x-ray crystallography and nuclear magnetic resonance applied to macromolecular structure determination. The Cryo-EM technique/technology involves imaging radiation-sensitive specimen samples with a transmission electron microscope where the samples have been cooled to cryogenic temperatures (Milne *et al.*, 2013). Cryo-EM allowed for the determination of the full length tetrameric HIV-1 IN complex of HIV-1B (Passos *et al.*, 2017). This solved structure allowed for further studies, including homology modelling of other HIV-1 subtype IN proteins. Constructing a homology model of HIV-1C IN, facilitated structural studies to assess the impact of different polymorphisms on the structure of HIV-1C IN subtype protein (Brado *et al.*, 2018; Rogers *et al.*, 2018).

Molecular dynamic simulation has played a significant role in understanding the functional role of HIV-1 IN and the effect of mutations on the structure. Particular focus has been placed on the 140's catalytic loop found close to the active site within the CCD domain of IN (Greenwald *et al.*, 1999). In one simulation study, Dewdney and colleagues (2013), investigated the mechanism of how primary and secondary mutations Q148H/R, G140S/A induces drug resistance. They found that the formation frequency of a transient helix increases in the 140s catalytic loop and helix length is increased from 3 residues to 4 residues in the presence of these mutations (Dewdney *et al.*, 2013). This helix formation resulted in reduced flexibility of the 140's loop relative to the loop within the wild type IN and caused the loop to serve as a gating mechanism which led to restricted RAL access to the active site (Dewdney *et al.*, 2013). Another study focused on the resistance inducing mechanism of the E92Q/N155H double mutation, particularly in relation to EVG resistance (Chen *et al.*, 2015). It was found

the mutation caused positional rearrangement around the active site and this rearrangement induced EVG to adopt a different binding mode with a lower binding affinity as compared to the wild type (Chen *et al.*, 2015). A significant drawback of these two studies is that a full length HIV-1 IN structure was not available for accurate comparisons and had to rely upon related IN structures or incomplete HIV-1 IN structures for modelling (Chen *et al.*, 2013). In drug design studies, simulations have also been used to compare the different binding modes of diketo-acids bound to HIV-1 IN, and these studies showed that some diketo-acids adopted similar conformations while other's displayed distinct conformations (Huang, Grant and Richards, 2011). The resulting interactions energies (van der Waals, Coulomb and hydrogen bonds) were calculated and could be correlated to binding pose(s). It was also possible to investigate IN conformations particularly changes to the 140s loop, informing further development of DKA's into more potent compounds (Huang, Grant and Richards, 2011). Based on the successes of previous studies we applied a similar approach to investigate structural differences between full length HIV-1C and 1B IN proteins.

1.7 Rationale of present research work

The aim of this thesis is to determine if naturally occurring polymorphisms affect IN protein structure and possibly DTG drug binding. We will be analysing the IN amino acid sequence derived from a South African HIV-1C INSTI treatment naïve cohort. To serve as a comparison we will be using HIV-1 subtype B, in particular the sequence and structure reported by Passos *et al.*, 2017. This will allow us to determine what structural and dynamic difference(s) exist between the two subtype IN's and any resulting consequences for DTG binding.

The aim of this study is subdivided into 3 objectives.

- 1) To build an accurate HIV-1C tetrameric protein structure.
- 2) To predict stabilizing/destabilizing effects of the polymorphisms found within HIV-1C_{za}.
- 3) To determine if there are any difference in protein dynamics between HIV-1B and HIV-1C using molecular dynamic simulations and Principal component analysis.

This work is pertinent to SA as a structural study focusing on the genetic diversity of HIV-1C IN in South Africa has not been previously reported. Importantly, DTG is being introduced into our HIV-1 treatment programme as part of first line therapy. This study addresses and overcomes multiple prior limitations of previous HIV-1 IN studies, as previous studies had been conducted prior to the solving of a full-length HIV-1 IN complex in 2017 (Passos *et al*, 2017). Several other differing aspects are that it is now known that the active site contains two magnesium ions and we are conducting longer MD simulations than previously reported, enabling us to potentially find and study HIV-1 IN conformational changes not previously described.



Chapter 2: Methods and Materials

1.1 Data preparation

The HIV-1C IN protein sequences were acquired from our collaborators based at the Division of Medical Virology, Stellenbosch University, Tygerberg campus who obtained the sequence data as described by Brado *et al.*, 2018 from South African ARV treatment naïve patients. All sequences have been made publicly available and obtainable from Genbank. The corresponding dataset ID for the sequences are 1475996009 (Brado et al, 2018).

1.2 Consensus sequence generation

The consensus sequence for HIV-1 subtype C was derived from South African IN cohort sequences (n = 91), with accession numbers ranging from [MH161467.1](#) to [MH161557.1](#). Nucleotide sequences were verified for stop codons, insertion and deletions using an online quality control program on the Los Alamos National Laboratory HIV database (HIVLANL) (<https://www.hiv.lanl.gov/content/sequence/QC/index.htm>). Multiple sequence alignments were done with the software tool Multiple Alignment using Fast Fourier Transform (MAFFT) version 7, from which the consensus sequence was derived (Katoh and Standley, 2013). As part of quality control, each of the viral sequences were inferred on a phylogenetic tree to eliminate possible contamination. The amino acid sequence alignment was extensively screened for the presence of primary and secondary resistance associated mutations (RAMs) and polymorphisms associated with resistance to known INSTIs. The screening was conducted using the Stanford database (Shafer, 2006). Briefly, its algorithm assesses and calculates a total drug penalty score by adding all the scores of mutations associated with a particular drug and classes it in 5 tiers from 1 indicating susceptible to 5 indicating high resistance. We only interrogated the 91 sequences for the presence of known RAM's (Liu and Shafer, 2006; Shafer, 2006).

2.1 Protein modelling

The consensus protein sequence of HIV-1C_{za} IN was generated as described above and previously reported by (Brado *et al*, 2018). The sequence was used to build a homology model for HIV-1C IN. Schrodinger Prime was used for homology modelling (Jacobson *et al.*, 2002, 2004). The complete Schrodinger package is a commercial software, available to researchers free of charge who make use of the Council for Scientific and Industrial Research (CSIR) Centre for High Performance (CHPC) supercomputing facilities. Prime was chosen because of high tertiary structure prediction accuracy and ease of use with a number of tutorials and case studies available from their website. A particular advantage it has over many other popular protein structure prediction software is the wide host of features and post-modelling refinement options.

Schrodinger Prime software has built-in access to the NCBI blast program and the Protein Data Bank (PDB). The first step was to identify homologous sequences available with experimentally solved structures deposited within the PDB by performing a BLAST search using the HIV-1C consensus sequence as input. The search produced a list of homologous templates. The templates were ranked according to sequence identity, coverage and sequence similarity to the target sequence used as input. The highest ranked template was 5U1C. 5U1C is the only resolved near full length HIV-1 IN tetramer structure reported to date, with the highest sequence identity, coverage and similarity to our target HIV-1C IN sequence.

5U1C had several unresolved loop regions within its structure, these were resolved by remodelling 5U1C with Prime along with converting the mutation glutamine present at position 152 to glutamate to form the standard DDE motif of HIV-1 IN. The model was then re-generated using default options. The newly modelled 5U1C/HIV-1B IN structure was used as the template to generate a HIV-1C_{za} IN three dimensional structure.

2.2 Model Validation

To determine the accuracy of the predicted protein structures, numerous quality assessments were performed. A variety of assessment tools available at the Structural Analysis and Verification Server (SAVES) which are PROCHECK, VERIFY3D, ERRAT, PROVE and WHATCHECK, was used to validate the quality of the predicted protein structures (Colovos and Yeates, 1993; Laskowski *et al.*, 1993; Hooft *et al.*, 1996; Pontius, Richelle and Wodak, 1996; Eisenberg, Lüthy and Bowie, 1997). PROCHECK assesses the stereochemical quality of the protein model to determine the orientation of phi and psi dihedral angle distributions of residues within the protein model. Verify3D determines the compatibility between the 3D model and its amino acid sequence by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar). ERRAT analyses the non-bonded interactions which occur between different atom types and compares this data with that of highly refined structures. PROVE calculates atomic volumes within macromolecules according to an algorithm which treats atoms like hard spheres and calculates a statistical Z-score deviation for the 3D model from well resolved structures (2Å or less). WHATCHECK assesses several stereochemical factors of residues within the 3D model. The PDB files of the 3D models were used as the input data. The HIV-1B/5U1C and HIV-1C_{za} IN 3D models were validated for their accuracy using SAVES webservice, HIV-1C_{za} IN was compared to the newly generated HIV-1B 3D structure to determine any structural deviation. The HIV-1C_{za} IN protein model was considered reliable, if the scores were similar to the scores obtained for the HIV-1B IN model. Cryo-EM resolved structures are not yet comparable in resolution to high quality x-ray crystallisation or nuclear magnetic resonance resolved structures upon which the various SAVES tools have been optimized and may thus not pass the standard thresholds used.

2.3 Complex structure generation (IN-DNA-MG without or with DTG)

HIV and other similar retroviral IN's function is dependent on one or two magnesium (Mg) cations within the active site (Miri *et al.*, 2014). To insert these Mg cations into the predicted structures we aligned the models to the homologous Prototype Foamy Virus (PFV) IN structure (ID: 3S3M) which contained the necessary Mg cations. These cations were then extracted and copied into our predicted 3D models. PFV IN active site is highly identical to HIV-1 IN active site and the structure has been experimentally solved in complex with two Mg cations and INSTI's (Hare *et al.*, 2010, 2011). Similarly, DNA present within 5U1C was added back into our predicted 3D models through structural alignment and subsequent extraction to generate IN-DNA-MG.

To create IN-DNA-MG-DTG complexes, DTG was extracted along with Mg cations from 3S3M. The alignments and subsequent structural feature extractions and copying was performed using the molecular visualisation software PyMol (Delano, 2002).

2.4 Structural Comparative Analysis

To determine if any structural changes occurred as a result of the polymorphisms present in HIV-1C_{za} secondary structure compared to HIV-1B a structural alignment was performed in PyMol. The HIV-1C_{za} model was superimposed onto the newly generated HIV-1B model and the root mean square deviation (RMSD) calculated using the align function within PyMol. This function allowed for the measurement of backbone atom difference in angstroms (Å) between the HIV-1B and HIV-1C_{za} IN proteins. An RMSD value closer to 0 indicates higher similarity between carbon backbone atoms of two structures. PyMol was used to visualize differences in the secondary structural features such as alpha helices, beta-sheets and coil regions for each IN protein structure.

Stability impact assessment of polymorphisms

To predict the effects that polymorphisms may have on HIV-1C_{za} integrase protein structure stability the machine learning based approach: mutation Cut-off Scanning Matrix (mCSM) and the knowledge-based approach Site Directed Mutator (SDM) web-based software tools were

used (Worth et al, 2011; Pires et al, 2014). mCSM uses pharmacophore properties such as aromatic rings, hydrogen bond acceptors and donators of the mutating residues and calculates resulting changes in Gibbs free energy. SDM calculates the thermal stability impact of mutations/polymorphisms on protein folding. Mutations in specific residues are considered and based on their local structural environment's substitution probabilities are calculated from analyses of families of protein homologues. The HIV-1B/5U1C IN protein model in PDB format and a text document containing the list of polymorphic differences present in HIV-1C_{za} was used as input. HIV-1B IN was used as the wild type (WT) structure to calculate the differences which occur when these polymorphisms are present. Both mCSM and SDM classify mutations/polymorphisms as either stabilizing or destabilizing. A negative value indicates reduced stability while a positive value indicates increased stability. We also calculated the loss or gain of polar interactions within 3.5Å distance caused by the polymorphisms in their immediate environment using PyMol. Usually, the loss of polar interactions is associated with a reduction in protein structural stability and gain of polar interactions conversely with an increase of protein stability.

3. Molecular Dynamic simulations

The Gromacs 2018.2 MD package coupled with CHARMM36M forcefield and the webserver CHARMM-GUI was used to prepare simulation systems and run molecular dynamic simulations (Van Der Spoel *et al.*, 2005; Jo *et al.*, 2008; Huang and MacKerell, 2013). A total of four systems were prepared, two IN-DNA-MG complexes and two IN-DNA-MG-DTG complexes.

3.1 Structural preparation of systems

Schrodinger Maestro Protein Preparation wizard was used following IN complex building with PyMol. This automated the process of bond order assignment, addition of the correct hydrogens, disulphide bonds and structural refinement such as bond removal or addition of missing bonds.

3.2 CHARMM-GUI

Following structural preparation, the structures were used as input into the CHARMM-GUI webserver's solution builder function. The DTG containing complexes required an extra step. DTG was extracted and copied to a separate file in PDB format and converted to sdf format using OBABEL (O'Boyle *et al.*, 2011), as sdf or mol2 formats are required as input for the ligand when uploading systems containing ligands. The IN-DNA-MG-DTG complexes required this extra step as forcefields are generally not parameterized for non-standard molecules and required the usage of a small molecule forcefield parameterization software (Vanommeslaeghe *et al.*, 2009).

The webserver automatically calculated the simulation water box size, the amount of water molecules, number of ions to neutralize the system and finally generated parameter files (mdp) as input for succeeding steps. The chosen box shape for the system was rectangular with the size calculated by CHARMM-GUI at 14.5Å in all 3 planes. The selected forcefield was CHARMM36M which is an additive all-atom forcefield able to handle proteins, DNA and ligands and has been further optimized to better deal with disordered proteins (Huang and MacKerell, 2013; Huang *et al.*, 2017). The systems were solvated in a TIP3 water box. All the systems were neutralized by the addition of sodium (Na) ions. A total of 90 294 TIP3 water molecules and 58 sodium ions were added to the IN-DNA systems, while a total of 70 599 TIP3 water molecules and 60 sodium ions were added to the IN-DTG-DNA systems each. No chloride ions were needed for neutralization of all the systems.

3.3 Energy minimisation

The four systems were energy minimized on a DELL Inspiron I5 3000 machine. The system was minimized over 5000 steps to an emtol of 1000.0 KJ/mol using the "steep" integrator method, with a step size of 0.01.

3.4 Equilibration

Following energy minimisation, the systems were equilibrated using the same machine. The CHARMM-GUI equilibration mdp file was used, which had a timestep of 0.001 and a total step number of 125000, equivalent to 125 picoseconds(ps). The aim was to equilibrate the system at NVT conditions to a temperature of 303.15 Kelvin. The thermostat algorithm used was Nose-Hoover (Nosé, 1984; Hoover, 1985). Constraints were placed on the h-bonds with the algorithm Linear Constraint solver LINCS (Hess *et al.*, 1997). The cut-off scheme algorithm used was the Verlet algorithm (Verlet, 1967). This cut-off scheme algorithm was chosen as it offers superior computational performance and is now the standard cut-off scheme algorithm implemented in Gromacs. The pressure regulator barometer was switched on during the production simulation therefore no NPT ensemble was performed.

3.5 Production simulation

After all, four systems were equilibrated. Production simulations were ran on the Centre for High Performance Computing (CHPC) Lengau cluster using Gromacs 2018.2. This facility allows jobs to be run in parallel on the central processing unit (CPU)/graphics processing unit (GPU) clusters improving the speed of the simulations. The parameters within the md.mdp file generated by CHARRM-GUI was kept with one modification, change of simulation duration. The production simulations were performed over a 300 ns time interval and repeated to validate reproducibility of trajectory results. The Parrinello-Rahman barometer algorithm was switched on to allow the system to be at a constant pressure of 1 atmosphere for the production simulation (Parrinello and Rahman, 1980, 1981). The simulations were conducted at random seed values a condition that ensures a pseudorandom starting conformation of the structure and no restraints were applied to the systems.

3.6 MD analysis

Trajectory analysis was performed using Gromacs tools. The tool *gmx trjconv* was used to eliminate periodic boundary condition artefacts or “jumps” within the simulation system and to repair any broken molecules. Backbone RMSD was calculated to confirm if the systems

reached equilibrium using *gmx rms*. The radius of Gyration is a measurement of protein compactness, was calculated using *gmx gyrate* for the protein atoms to determine if the protein's remained folded during the simulation. The last 50ns of the trajectory corresponding to the equilibrated part of the simulation were used for further analysis. To extract the equilibrated part of the trajectory the tool *gmx trjconv* was used and the start and end frames specified. RMSF fluctuation was calculated to assess flexibility of the protein residues with the tool *gmx rmsf*. The average number of Hydrogen bonds was calculated between IN-DTG and DTG-DNA using the tool *gmx hbond*. Distance measurements between DTG and the magnesium ions were calculated using *gmx mindist*. The non-bonded interaction energy between DTG and the IN complex was calculated via a number of steps; firstly, a copy of the production simulation mdp file was made with a modification as additional energy groups were defined (IN complex and DTG), secondly *gmx mdrun* was used in conjunction with the *-rerun* parameter and finally the energy components were extracted using *gmx energy*. The two components that compose the total interaction energy are the Coulomb and Van der Waals contributions. The thermodynamic properties of the systems were calculated using *gmx energy*, with the selected properties of temperature, potential and total energy. These selected properties allow for confirming whether the system reached convergence. Clustering analysis was done to identify similar structural conformations/clusters sampled during the simulation run. The gromacs tool *gmx cluster* was used to perform clustering at an average RMSD cut-off value. Determination of the RMSD cut off value was done by an initial clustering run with the default cut-off value of 0.1nm. The average RMSD value obtained was used as the cut-off value.

Principal Component Analysis (PCA) is a useful statistical method used to reduce the complexity of the generated data set to extract the most important biologically relevant movements of the IN protein (David and Jacobs, 2014). The gromacs tools that were used included *gmx covar* to construct the covariance matrix and *gmx anaeig* to calculate the eigenvectors and eigenvalues by diagonalizing the matrix. The output of *gmx covar* were *eigenvec.trr* and *eigenval.xvg* files, *eigenval.xvg* listed all the eigenvectors and their contribution to protein motion. We selected the most dominant eigenvectors; namely vector 1 and 2 using *gmx anaeig*. The results were then plotted in a 2D dotplot using the plotting software GNU PLOT (Janert, 2016)

Chapter 3: Results:

1.1 Protein Homology Modelling:

The HIV-1C_{za} IN model was generated by alignment of the HIV-1C consensus sequence with the sequence of the HIV-1B solved structure (ID:5U1C). The sequence identity was calculated to be 88% and sequence similarity 92% when aligned to the 5U1C sequence. The alignment showed missing residues within 5U1C and these residues were added into the 3D protein model by re-modelling (Figure 8). Residue Q152 was converted from glutamine to glutamate (E152) to reconstitute the naturally occurring DDE motif (Figure 8). The sequence gaps or missing unresolved structural sections were remodelled with Schrodinger Prime. The remodelled HIV-1B IN served as the template to build a HIV-1C_{za} model with Schrodinger Prime. All four chains can be seen comprising the tetrameric IN structure in Figure 9. The two inner dimers are responsible for interaction with DNA and the catalytic activity of IN. The two outer dimers serve to stabilize the tetramer by a protein-protein interaction with the inner dimers.

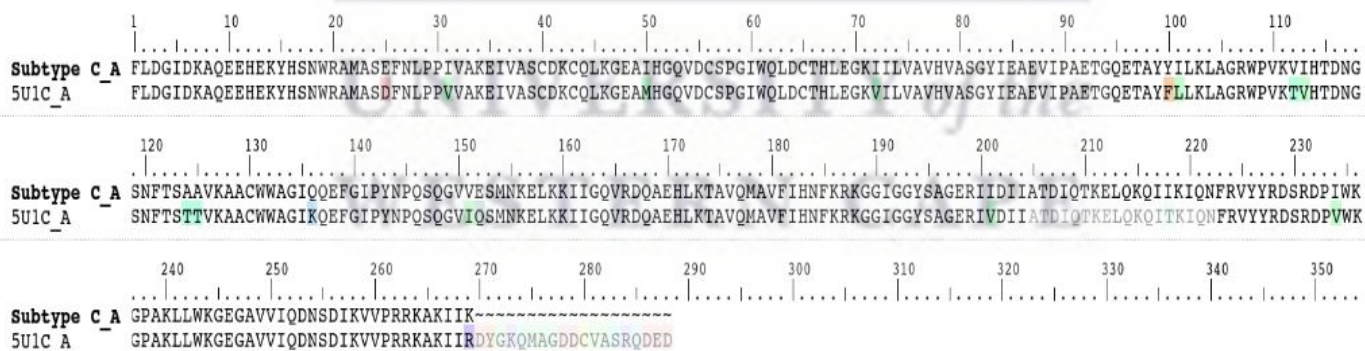


Figure 8: Pair-wise sequence alignment between HIV-1C consensus sequence and HIV-1B amino acid sequence. The sequences of HIV-1C_{za} compared with the 5U1C/HIV-1B structure prior to modification. Shown in colour(s) are the differing polymorphisms between the two sequences. It can also be seen the missing regions in 5U1C which were fixed by re-modelling, shown in grey and at position 152 the residue glutamine is present instead of glutamate for 5U1C.

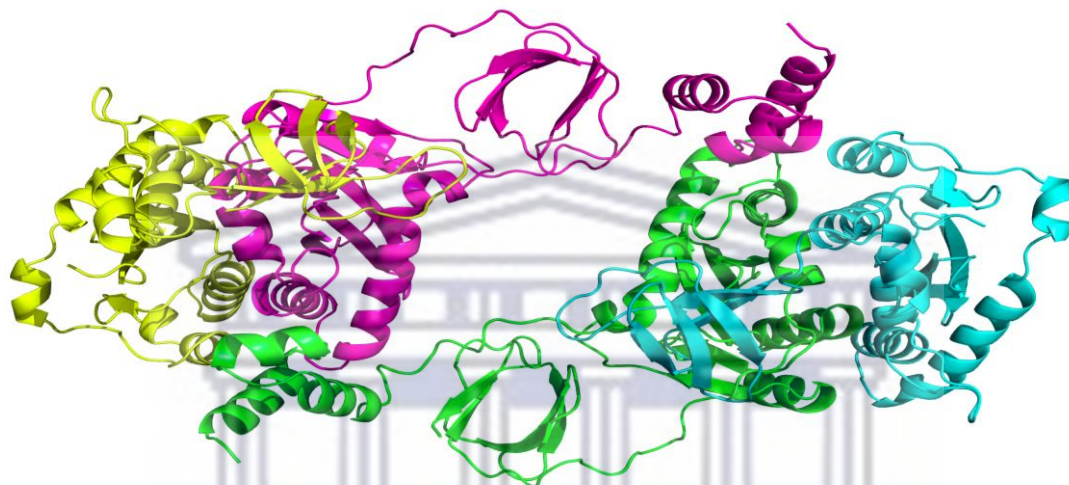


Figure 9: Cartoon representation of the predicted tetrameric 3D structure of HIV-1C. The secondary structure for each chain is shown in a different colour. The purple and green chains comprise the inner dimer while the blue and yellow chains comprise the outer dimer.

1.1.2 Protein Model Quality Assessment:

Protein model quality assessment consisted of a number of computational tests to assess the reliability of the generated models. Firstly, the backbone structure of HIV-1C_{za} IN was superimposed onto HIV-1B IN, indicating an RMSD value of 0.03Å. Secondly, the tools located at the SAVES webservice assessed the quality of the generated model and this was compared to the HIV-1B template structure. The Ramachandran plot generated by the PROCHECK tool in SAVES showed 84.7% of residues are located in most favoured regions of the plot and none in disallowed regions, compared with 87.3% of residues located in most favoured regions of the Ramachandran plot and 0.8% in disallowed regions for HIV-1B IN (Figure S1A). The ERRAT overall energy analysis score was found to be 90.5% for HIV-1C_{za} IN structure compared to 97% for HIV-1B, while the Verify3D score indicated that 75.84% of

the residues had an average 3D-1D score ≥ 0.2 for HIV-1C_{za} IN compared to the 83.4% for HIV-1B IN. In summary, the two models are comparable and suitable for further analysis.

1.2 Variant stability change calculations:

Natural occurring polymorphisms (variants) identified in the amino acid sequence of HIV-1C IN were interrogated structurally to determine their phenotypic effect on the protein structure using the webserver SDM and mCSM. The remodelled HIV-1B template was used as the WT input structure. Furthermore, the loss or gain of polar interaction due to the variant was calculated using the find polar contacts function in PyMol. In Table 1, mCSM classified all variants as destabilizing, while SDM showed contradictory results with eight of the 12 variants being classified as stabilizing and the remaining four being destabilizing. Furthermore, 3 variants showed a loss in polar contacts and one showed a gain in the number of polar contacts with neighbouring amino acids. To further clarify the effect of the variants we conducted MD to assess the overall effect of the natural occurring variants on the protein's dynamics.

Table 1: The predicted Gibbs free energy change due to the introduction of 12 natural occurring polymorphisms in HIV-1B IN.

Polymorphism	SDM (kcal/mol)	mCSM (kcal/mol)	#Polar interactions	
			HIV-1B	HIV-1C _{za}
D25E	1.41	-0.191	2(ALA21, LYS188)	1(ALA21)
V31I	0.19	-0.433	2(LYS34, GLU35)	2(LYS34, GLU35)
M50I	0.70	-0.22	0	0
V72I	-0.28	-0.902	1(ILE89)	1(ILE89)
F100Y	-1.08	-0.785	2(GLU96, LEU104)	3(GLU85, GLU96, LEU104)

L101I	-0.39	-0.685	3(ALA98, ALA105, THR97)	3(ALA98, ALA105, THR97)
T112V	2.17	-0.825	4(GLY59, TRP61, LYS136, VAL113)	2(GLY59, TRP61)
T124A	1.53	-0.296	1(ALA128)	1(ALA128)
T125A	1.53	-0.638	2(ALA128, ALA129)	2(ALA128, ALA129)
K136Q	-0.43	-0.204	3(LYS111, THR112, VAL113)	2(LYS111, ILE113)
V201I	0.19	-0.287	1(GLY197)	1(GLY197)
T218I	0.46	-0.057	0	0

Negative numbers indicate the variant has a destabilizing effect and a positive number a stabilizing effect.

indicates number of polar interactions. Amino acid abbreviations used: ALA: Alanine, LYS: Lysine, GLU: Glutamate, ILE: Isoleucine, GLY: Glycine, VAL: Valine, TRP: Tryptophan, LEU: Leucine

1.3 Molecular dynamic simulations:

In this section, the simulation results are split into two separate sections. One section focuses on the IN-DNA complexes and compares the complexes using *RMSD*, *GYRATION*, *RMSF* and *PCA* solely. The second section focuses on results produced for IN-DNA-DTG complexes, additional analyses presented here are DTG-MG distance, Hbond analysis, *clustering* and non-bonded interaction-analysis. IN-DNA complex simulations allowed us to analyze the structural differences between IN B and IN C in respect to folding compactness, residue flexibility and essential dynamics. IN-DNA-DTG complexes allowed us to determine whether structural conformational changes can account for differences in DTG binding to IN B and IN C subtypes.

In the supplementary material we provide the RMSD and Radius of gyration results of the repeated simulations, as well as a table showing the thermodynamic properties for each system.

1.3.1 IN-DNA-MG complex

Firstly, the backbone RMSD deviation for each of the IN protein structures indicates whether the protein backbone of the two systems are stable. Both IN C and IN B structures deviate around 0.3 nm and reaches equilibrium after 225 ns (Figure 10). Significant RMSD overlap is observed for both structures in the last 50 ns. The mean and standard deviation values calculated for the backbone RMSD was 0.325 ± 0.025 nm for IN B and 0.316 ± 0.036 nm for IN C (Figure 10), respectively. Secondly, the Radius of gyration (Rg) measures the compactness of the protein structures and if they remain folded throughout the simulation. Figure 11 shows the Rg for the backbone atoms of both IN structures measured over the final 50ns of the trajectories, which corresponds to the most stable part of the trajectory. Both structures deviate between 3.6 and 3.7 nm with a mean and standard deviation of 3.670 ± 0.012 nm for IN B and 3.680 ± 0.017 nm for IN C. Thirdly, the RMSF analysis indicate average residue flexibility for the protein residues over the course of the simulation. The RMSF values are shown for the final 50 ns for each IN protein system. In Figure 12 it is shown that RMSF values for IN B and C overlap, but in a few regions there are higher flexibility. The mean and standard deviation values calculated for C-alpha RMSF atoms was 0.11 ± 0.06 nm for IN B and for IN C it was 0.12 ± 0.057 nm. The regions of high flexibility occur primarily in the loop regions, one such region which comprises the 140s loop or catalytic loop displayed greater flexibility with values reaching 0.3 nm. IN B displays large amounts of flexibility at residue region 200-225, this occurs in a disordered coil region and therefore is expected to display higher flexibility, it peaked at a value over 0.5 nm. Furthermore, this represents an unresolved region of the original 5U1C template which was remodelled.

Finally, Principal component analysis was used to determine the contribution of the first 5 eigenvectors to the variation in the data set (Figure 13). In IN C the contributions were as follows for the first five eigenvectors; 1 contributed 51%, 2 contributed 16.6%, 3 contributed 13.2%, 4 contributed 10.1% and finally eigenvector 5 was responsible for 9.1% of the variation. In IN B the contributions were as follows; 1 contributed 34.65%, 2 contributed 22.5%, 3 contributed 18%, 4 contributed 14% and finally eigenvector 5 was responsible for 10.7% of the variation. Based on the PCA results we plotted the first two eigenvectors as they contributed to the majority of the protein movement. The first two eigenvectors of IN C showed more randomized movements with two clusters throughout the phase space as compared with IN B which demonstrated more concerted movements and adopting three distinct clusters suggesting that the IN B system was more stable than IN C (Figure 13). For the repeated simulation system the RMSD was calculated to be 0.355 ± 0.031 nm for IN B and 0.308 ± 0.030 nm for IN C (Figure S2A). Similarly, For the Rg values the mean and standard deviations were 3.647 ± 0.018 nm for IN B and 3.692 ± 0.018 nm for IN C (Figure S2B). The findings are in agreement with the first production run indicating that the IN B system is more compact than IN C.

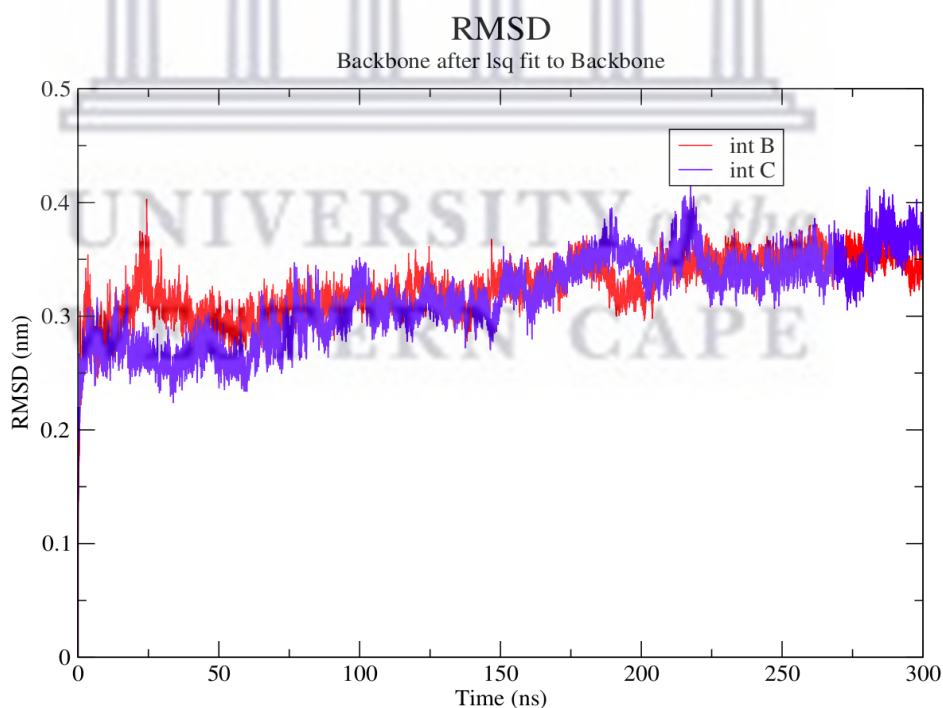


Figure 10: Change in backbone RMSD for the HIV-1 subtypes C and B IN proteins plotted over 300ns

Radius of gyration (total and around axes)

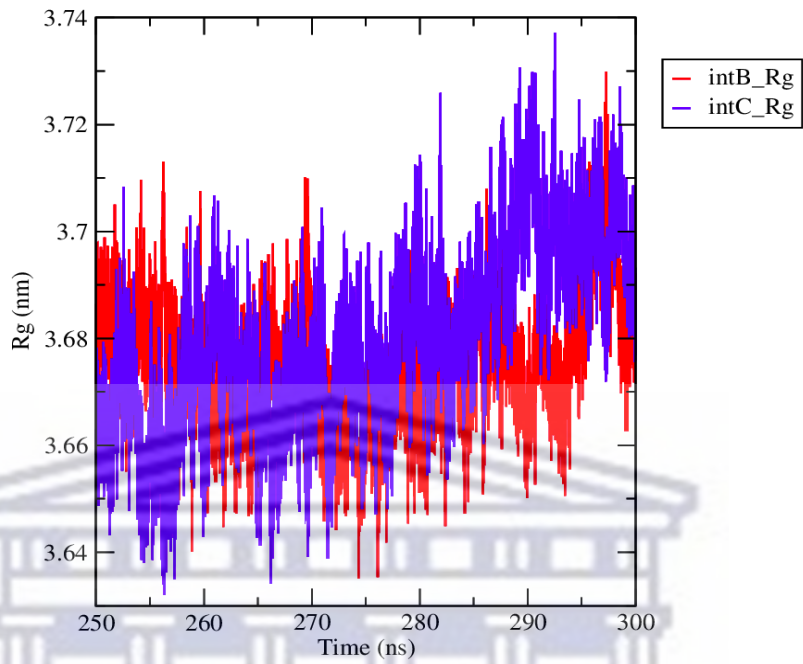


Figure 11: Radius of gyration measured for backbone atoms for both HIV-1 subtype B and C IN proteins plotted over the last 50ns

RMS fluctuation

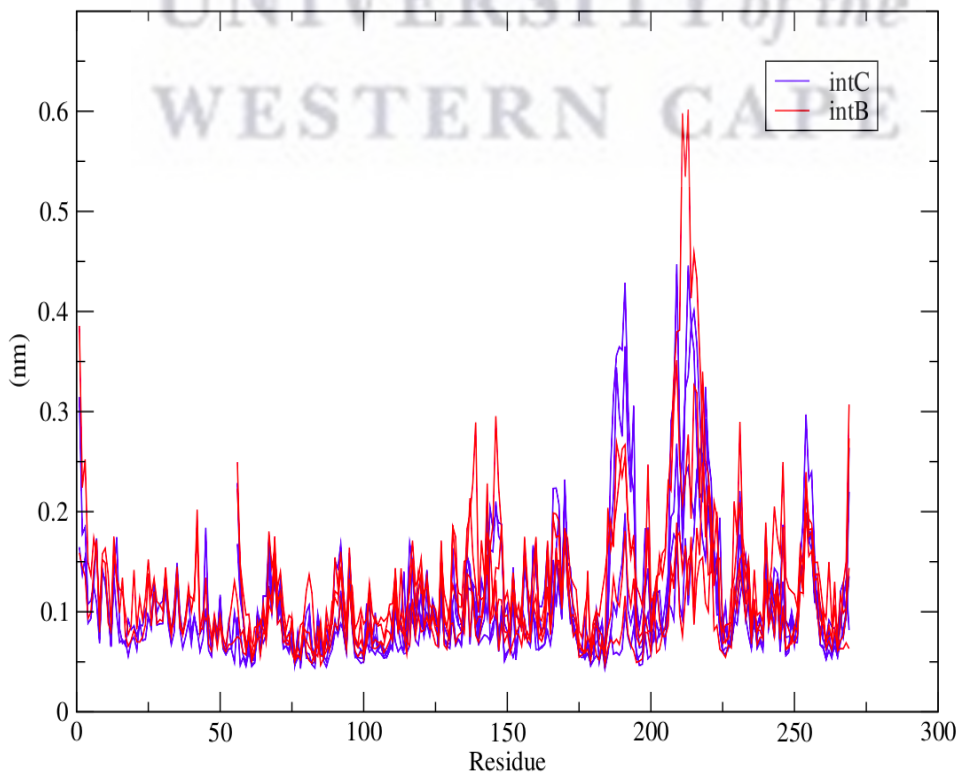


Figure 12: Change in RMSF for the C_α residues of the IN protein models plotted over the last 50ns of the trajectory.

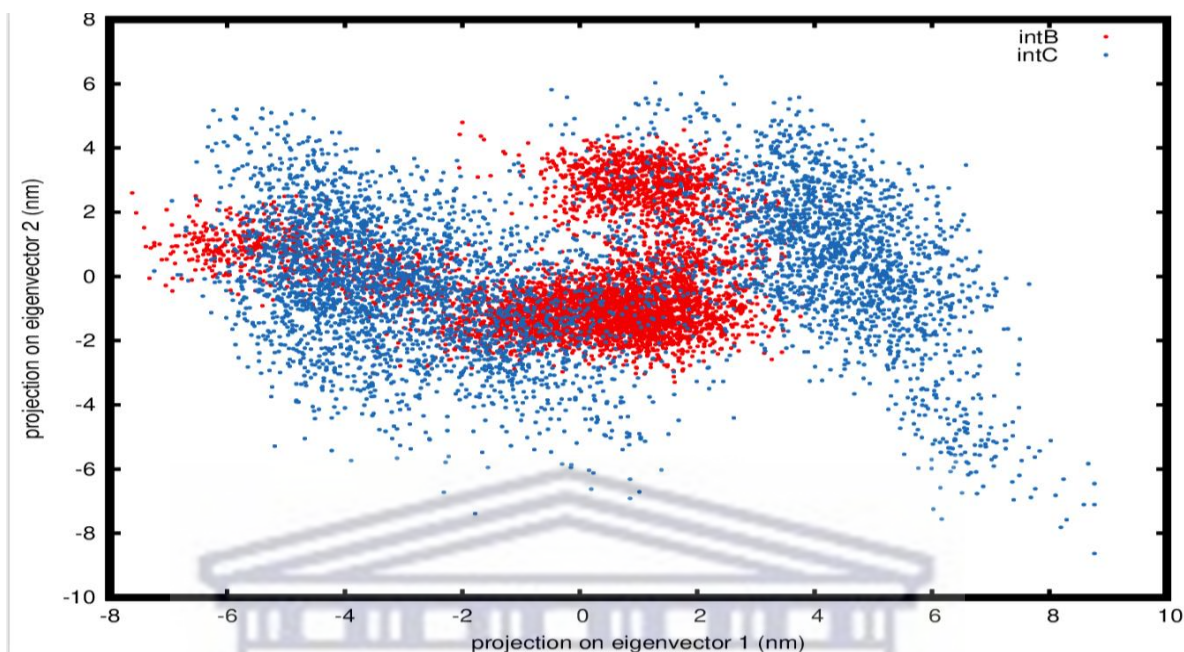


Figure 13: Two-dimensional (2D) PCA Plot for eigenvector 1 vs eigenvector 2 for both HIV-1 C and B DNA-MG systems plotted over the last 50 ns.

1.3.2 IN-DNA-MG-DTG complex

The complexes are shown in Figures S3A and S3B. The proteins backbone RMSD values showed that the two IN proteins in complex with the ligand DTG followed initially differing trajectories toward stabilization but equilibrated at similar values (Figure 14). The mean and standard deviation values for the backbone RMSD were 0.330 ± 0.031 nm for IN B and for IN C was 0.340 ± 0.034 nm. The ligand RMSD was calculated to assess the stability of the ligand DTG during the simulation (Figure 15). Over the course of the full 300 ns trajectory, the ligand DTG in complex with IN C stabilizes after 50 ns and remains stable throughout the remainder of the simulation. DTG in complex with IN B remains in a given conformation until the ~ 75 ns and then adopts a new conformation with seemingly large deviations for the remainder of the trajectory. Conformational changes of the DTG ligand within the IN B complex are shown in figures S4A and S4B. IN B complex residues form polar contacts between DTG and MG

ions. The major structural difference is that at timepoint 50ns the DTG remains more linear while at timepoint 200ns the benzene ring on DTG adopts a greater flexed conformation pointing away from the catalytic DDE motif. The flexed away conformation allows for one extra polar interaction to occur with DNA at 200ns. The mean and standard deviations for the heavy chain atoms of the ligand was 0.18 ± 0.062 nm for IN B and for IN C was 0.14 ± 0.012 nm. The R_g analyses showed that both complexes remain folded during the final 50 ns (Figure 16). HIV-1 IN C is more compactly folded than IN B, with a mean and standard deviation of 3.632 ± 0.012 nm compared with 3.69 ± 0.014 nm for IN B.

The RMSF analysis showed that the flexibility of the two complexes remained well conserved (Figure 17). The mean and standard deviations for the RMS fluctuation of IN B was 0.109 ± 0.058 nm and 0.096 ± 0.063 nm for IN C. The distance in space between DTG and MG ions are important as these ions are exploited by all current INSTI's for their mechanism of inhibition. Gromacs minimum distance analysis conducted over the last 50 ns showed both complexes with DTG deviate around different values (Figure 18). The mean and standard deviations of minimum distance calculated were 0.35 ± 0.023 nm for IN B and 0.44 ± 0.021 nm for IN C. Hydrogen bonds are important for both structural stability and the strength of ligand binding to the IN active site and viral DNA. We performed *Hbond* analysis over the last 50ns of the trajectory and computed the average number of hydrogen bonds forming between DTG – IN and DTG – DNA (Figures 19 and 20). Polar interactions were calculated at 10ns intervals over the last 50ns and summarized in table S5. It shows that DTG in IN B forms bonds with DDE motif residues whereas IN C does not, although bonds are still formed with nearby residues. The mean and standard deviations calculated for hydrogen bond formation between DTG – IN were 0.44 ± 0.539 nm for IN C and 1.27 ± 0.558 nm for IN B. The mean and standard deviations calculated between DTG – DNA were 0.75 ± 0.428 nm and 0.36 ± 0.486 nm for IN's C and B respectively. Clustering allowed for the determination of the number of different conformations that the complexes assume at an average RMSD. To determine the average RMSD, the command was first executed without a cut-off value. The lowest resulting average RMSD was used as the cut-off value and the cluster command executed once again. The lowest RMSD average was 0.26nm. At this cut-off value both IN C and B formed 4 clusters.

Principal component analysis was conducted, and the percentage contribution for the first five components were calculated. For HIV-1C IN the contributions were as follows for the first 5 eigenvectors; 1 contributed 38.68%, 2 contributed 24.59%, 3 contributed 16.11%, 4 contributed 12.11% and finally eigenvector 5 was responsible for 8.51% of variation. For HIV-1B IN the contributions were as follows; 1 contributed 34.09%, 2 contributed 24.34%, 3 contributed 19.43%, 4 contributed 12.48% and finally eigenvector 5 was responsible for 9.66% of variation. In Figure 21 we plotted the two main contributors to the dynamics of the protein namely eigenvector 1 and 2.

The 2D plot showed IN C having more randomized movements in the protein's phase space adopting two clusters. However, both systems displayed wide range of movements throughout the phase space indicating that DTG binding results in fewer concerted movements and possibly fewer low energy minima states. The Non-bonded Interaction energy method calculates the short-range Coulombic interaction energy and the Lennard-jones interaction energy. The sum of these components represents the total interaction energy. The interaction energy between DTG and the IN-DNA-MG groups were calculated over the last 50 ns trajectory. The mean values and standard deviations are shown in Table 2. Interestingly, both components contributed equally to the total interaction energy. However, HIV-1C IN showed stronger affinity for DTG compared to HIV-1B (gold standard) suggesting DTG is a feasible option for patients infected with HIV-1 subtype C. For the repeat simulation systems, the mean and standard deviation values for the RMSD was 0.339 ± 0.029 nm for IN B and 0.349 ± 0.032 nm for IN C (Figure S2C). The mean and standard deviation Rg values for the repeat simulation systems was 3.6389 ± 0.013 nm for IN B and 3.625 ± 0.021 nm for IN C (Figure S2D). This finding is similar to the first simulation run whereby the IN C system where found to be more stable compared to IN B based on Rg values.

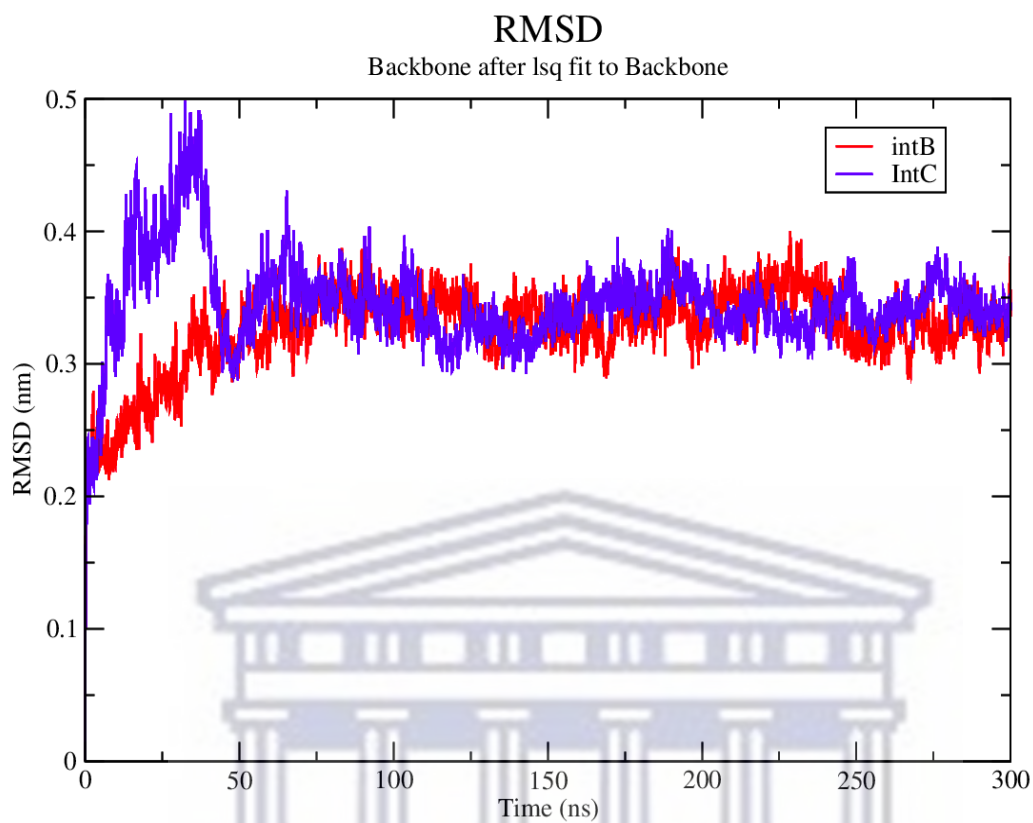


Figure 14: Change in backbone RMSD for the two IN-DNA-DTG complexes plotted over the course of the 300 ns simulation trajectory.

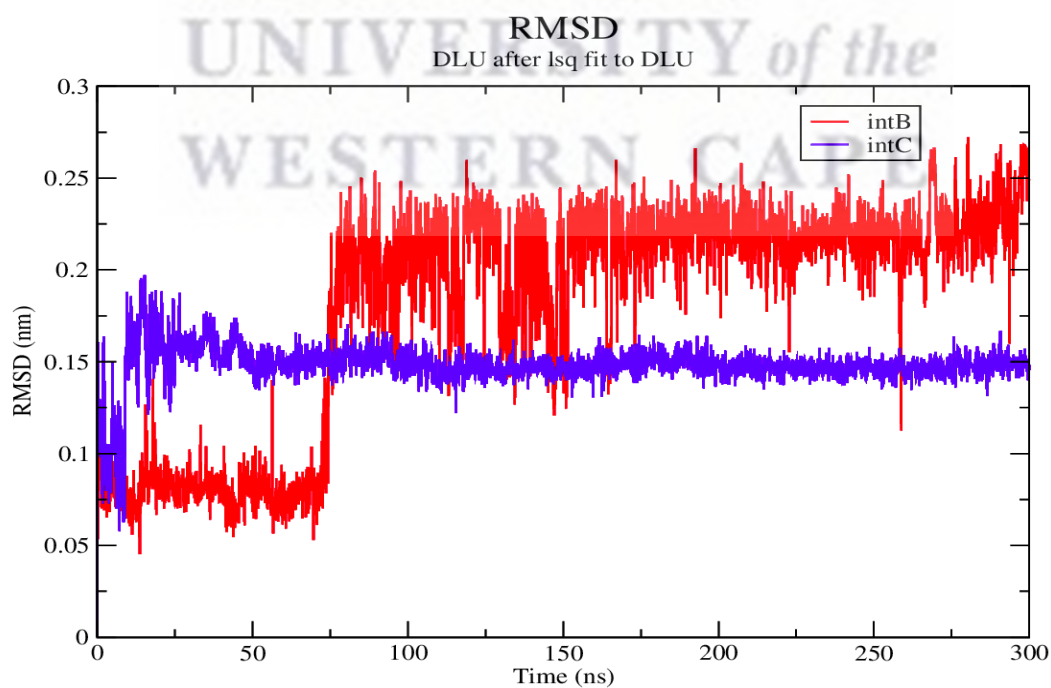


Figure 15: Change in RMSD for the ligand DTG for both complexes over 300ns

Radius of gyration (total and around axes)

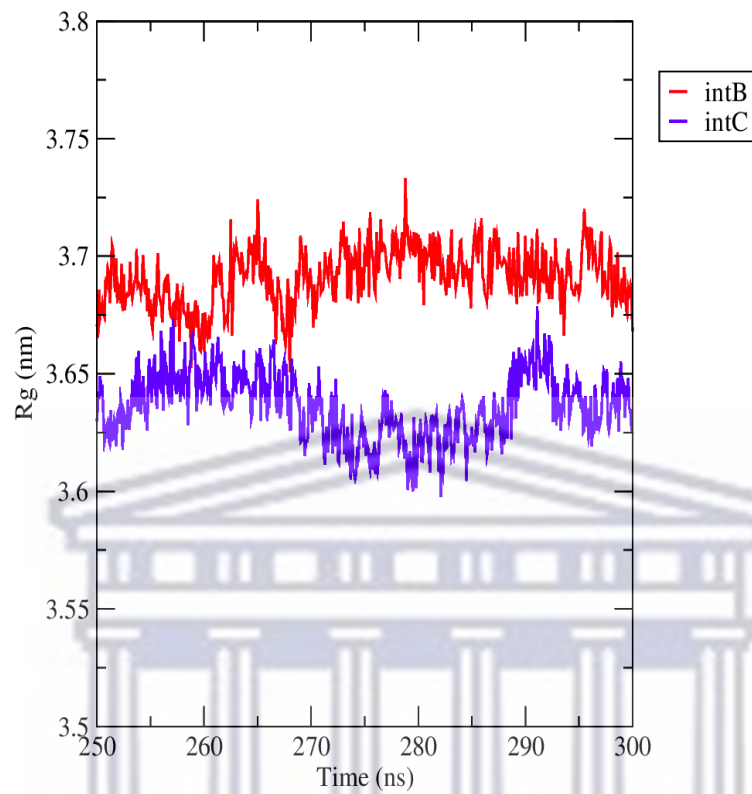


Figure 16: Measure of compactness for the IN-DNA-DTG complexes over the final 50ns of the trajectory

UNIVERSITY of the
WESTERN CAPE

RMS fluctuation

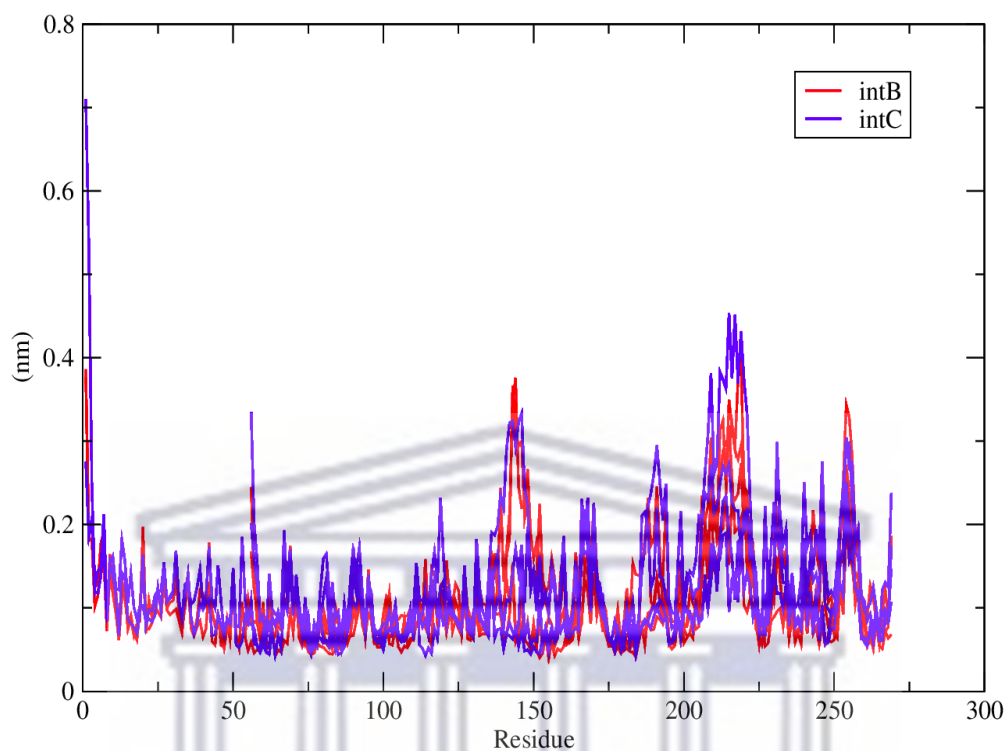


Figure 17: Change in RMSF fluctuation for the C-alpha residues of IN-DNA-DTG complexes plotted over the course of the final 50 ns of the simulation

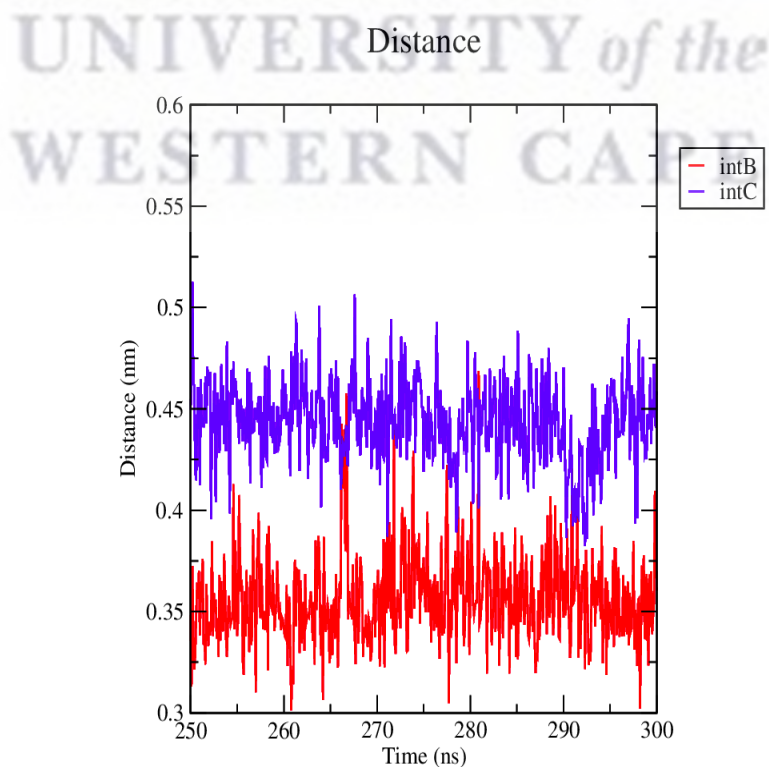


Figure 18: Change in minimum distance between the ligand DTG and MG ion group plotted over the <http://etd.uwc.ac.za/>

Hydrogen Bonds

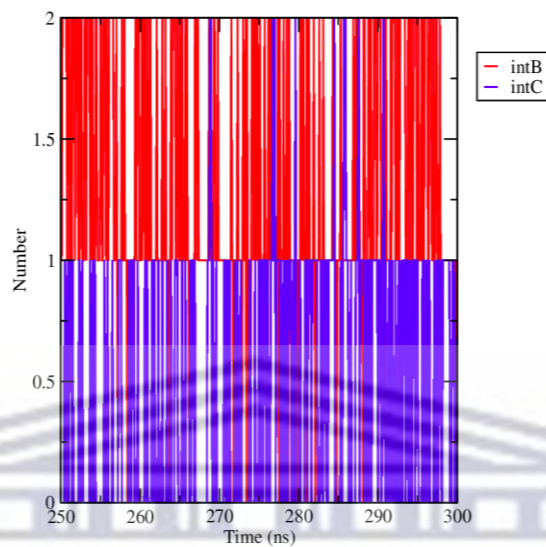


Figure 19: The average number of hydrogen bonds formed between DTG and IN during last 50ns of the trajectory

UNIVERSITY of the
WESTERN CAPE

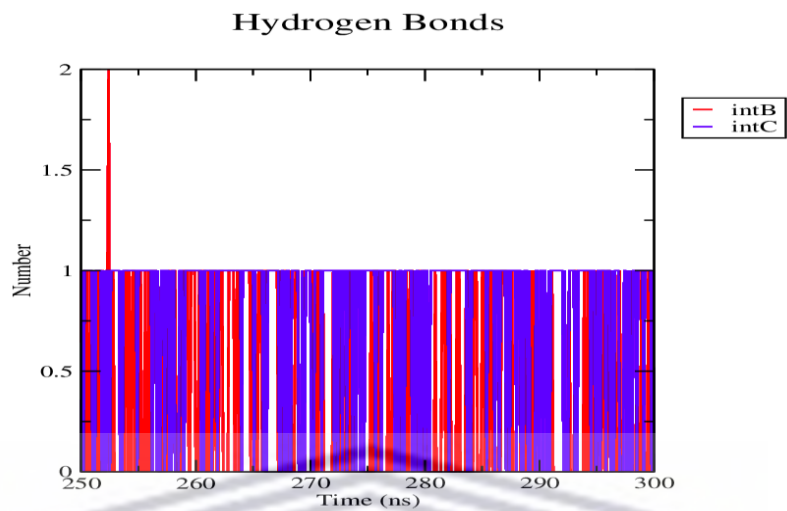


Figure 20: The average number of hydrogen bonds formed between DTG and DNA nucleotides plotted over the last 50 ns

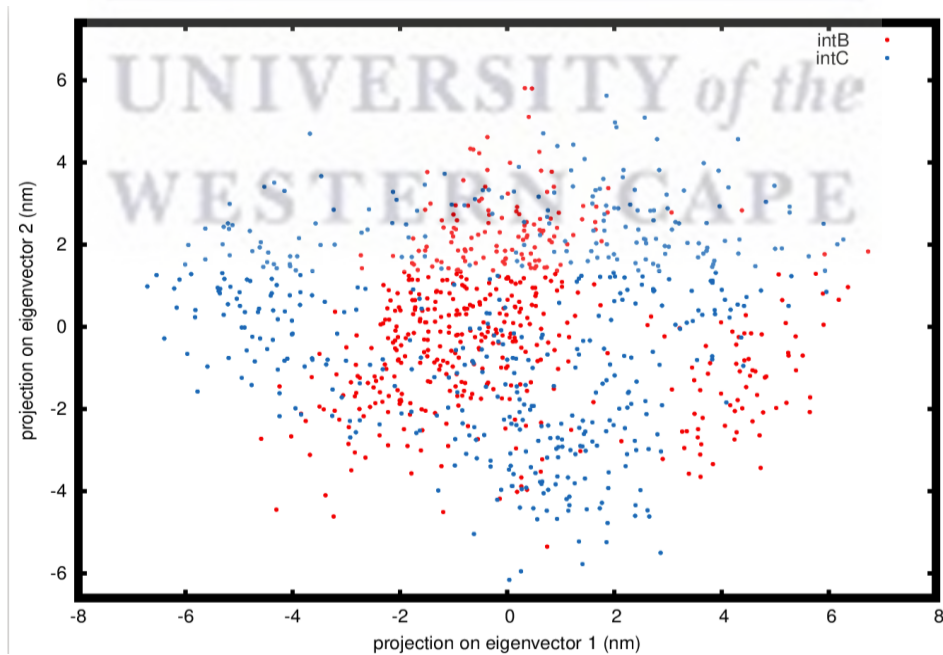
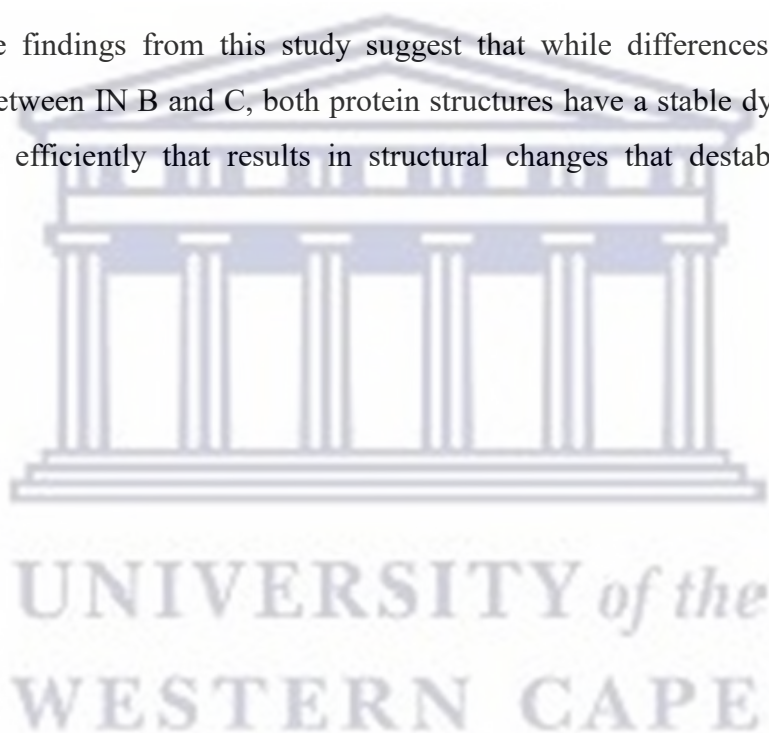


Figure 21: 2D PCA Plot for eigenvector 1 vs eigenvector 2 for both HIV-1 C and B DNA-MG-DTG systems over last 50ns.

Table 2: The interaction energy composition between DTG and each complex is shown.

Complex	Lennard-Jones (KJ/Mol)	Coulombic (KJ/Mol)	Total (KJ/Mol)
IN B	-57.55±12.35	-57.40±17.96	-114.95±15.15
IN C	-68.57±10.26	-69.82±19.21	-138.39±14.735

In summary, the findings from this study suggest that while differences in polymorphic variation exist between IN B and C, both protein structures have a stable dynamic behaviour and binds DTG efficiently that results in structural changes that destabilize the protein structure.



Chapter 4: Discussion and Conclusion

In this thesis we compared two HIV-1 IN proteins from two distinct subtypes namely, subtypes B and C to determine if structural differences would affect protein structure behaviour and drug binding. We investigated the effect of natural occurring polymorphisms present in the South African HIV-1C IN protein, to inform DTG use within the South Africa population. In the following paragraphs we discuss the findings of this study and its implications as well as the limitations and future perspectives.

The 3D structures predicted for HIV-1 IN B and C are highly similar in structural fold, differing only by the variant side chains. Structural visualization of the 3D models showed that the secondary structural features of IN C and B are identical, with RMSD analysis indicating minimal deviation in backbone residues. Noticeably, regions that were missing from template 5U1C are connecting regions between the monomers and was repaired by modelling these connecting regions. These regions correspond to highly disordered regions that remain unamenable to experimental resolving. Furthermore, the C-terminal tail end of HIV-1C and B could not be modelled as it was not resolved in the original 5U1C template model. Subsequent, protein model quality analyses using a variety of tools and software programs confirmed the accuracy of the predicted protein structures of HIV-1C and B. Our modelling results agree with previous reports that IN is highly conserved between subtypes and that the variants identified have a negligible effect on the proteins structure. One study in particular focused on variants within IN from different HIV-1 isolates sourced from the Pacific islands of Jayapura and Papua (Hutapea, Maladan, and Widodo, 2018). They found that although variants slightly altered some characteristics of the protein it did not affect structural stability and INSTI binding, even though the presence of a longer helix region in one of the IN isolates had been identified (Hutapea, Maladan, and Widodo, 2018).

SDM and mCSM predictions showed that none of the individual variants tested, showed a significant impact on IN protein stability. The software mCSM classified all variants as slightly destabilizing while SDM and interaction analysis showed contrasting results. For example, the

D25E was predicted to be stabilizing by SDM but showed a loss of one polar interaction. The V31I variant was predicted as slightly stabilizing by SDM but, no polar interactions were lost with this variant, V31I likely therefore has a negligible or a near neutral effect upon the stability of IN protein structure. The variant M50I has been previously implicated as a secondary mutation in INSTI drug resistance pathways, although it was noted in this study to play a minimal to no role in restoring viral fitness but instead contributes to drug resistance (Wares *et al.*, 2014). SDM predicted M50I to be slightly stabilizing. Furthermore, the variant V72I was predicted to be destabilizing by SDM and showed no loss of polar interactions. SDM predicted the variant F100Y to be destabilizing, however a polar interaction was gained with neighbouring residues. The variant L101I was predicted to have a destabilizing effect, although no polar interaction loss occurred. The T112V variant was predicted to be stabilizing by SDM however 2 polar interactions were lost and therefore it is more likely that T112V has a net destabilizing effect. The variant T124A was predicted as stabilizing by SDM, however no change in the number of polar interactions occur. Similarly, the variants T125A and V201I were predicted to increase stability of the protein structure but with no difference in the number of polar interactions and therefore likely to have a neutral or negligible effect on the protein structure. The K136Q variant resulted in reduced stability with one polar interaction being lost. The T128I variant was predicted to be stabilizing by SDM but like the other variants the change in GIBB's free energy was inconsequential and therefore neutral in effect. In contrast to a study by Brado *et al.*, 2018, we used structural tools; SDM and mCSM to assess changes in protein energy due to the introduction of a variant. In the study by Brado *et al.*, 2018 only 5 variants (D25E, M50I, F100Y, L101I and V201I) were interrogated. They found that M50I is in close proximity to two DNA substrate strands and may therefore be important for DNA binding and that D25E forms an ion pair in a symmetrical fashion between 2 monomers of IN. In Rogers *et al.*, 2018 all variants were investigated with the conclusion that several of these variants may affect IN structural stability, vDNA binding and drug binding. Our findings suggest that these changes (D25E, M50I, F100Y, L101I and V201I) has little to no effect on the IN protein structure and function as measured by a change in GIBBS free energy and number of polar interactions gained or lost between neighbouring residues. We further assessed the effects of the natural occurring variants on the protein structures dynamic behaviour using molecular dynamics. We discuss our findings below.

Molecular dynamic simulations

IN-MG-DNA complexes

The backbone RMSD for the IN-DNA complexes showed that both complexes reached equilibrium after 220 ns to a stable value of ~0.3nm. The high similarity in RMSD deviation values suggest that the backbone structure of HIV-1C IN is nearly identical to that of HIV-1B IN and that the polymorphic differences between the subtypes do not significantly alter the backbone structure of HIV-1C IN. The highest degree of overlap between the complexes was found between 250 ns and 300 ns, as a result further analyses were conducted on the final 50 ns of the simulation. The radius of gyration values also confirmed very little difference in compactness of the two simulation systems and this was supported by comparable RMSF fluctuation values between the two structures suggesting similar behaviour between the two structures.

Furthermore, the flexible regions in HIV-1B showed agreement to other studies (Delelis *et al.*, 2008; Williams and Essex, 2009) w.r.t the flexibility of the G140's loop region. This region is particularly important for INSTI binding, with it acting as a gating mechanism regulating DTG access. It can be seen that HIV-1B IN is more flexible in the G140 region compared to HIV-1C IN, with excessive flexibility caused by mutations in this region being implicated as a cause of INSTI drug resistance (Chen *et al.*, 2013, 2015). HIV-1B IN is more flexible at residues 137 and 138 where the variant substitution K136Q occurs. However, this difference in flexibility is not likely to be significant enough to exert an effect upon drug binding. The PCA analysis indicated that IN B is a more stable system than the IN C as the atomic movements were found to be more concerted.

MD studies of IN-MG-DTG complexes to understand the influence of polymorphisms on drug binding

The backbone RMSD values for the two IN systems did not show any significant differences except for the heavy atoms of the drug that adopted two different conformations for HIV-1B IN. This may be due to flexibility of the G140's loop region affecting the volume and size of

the HIV-1B IN binding pocket. Also, the drug DTG is a planar molecule with a benzene ring that can flip around in the binding pocket. This is further explained as smaller molecules being more flexible than larger biological macromolecules. However, DTG attains stability after a certain time period throughout the simulation suggesting in both systems the drug remains stable. Similarly, the radius of gyration and RMSF analysis indicated that both systems were highly comparable with no significant deviation between the two structures. From the RMSF analysis there were no regions of high flexibility at the substitution/variant sites or at G140 loop suggesting little to no effect of the variants on the drug binding site via the gating mechanism of the G140's loop. The only differences were found in the average number of hydrogen bonds formed between the IN protein and drug DTG and the distance between the drug and Mg ions. In the HIV-1B IN complex the drug DTG makes more hydrogen bonds with the IN protein and is in closer proximity to the Mg ions resulting in a stable protein-drug complex. On the other hand the HIV-1C IN protein formed a lower number of hydrogen bonds with DTG but a higher number of hydrogen bonds with the DNA resulting in a stable drug-DNA complex formation. The importance of ionic bonds with Mg ions within the IN active site has been widely reported on and is crucial for DNA binding (Liao and Nicklaus, 2010; Ribeiro, Ramos and Fernandes, 2012; Miri *et al.*, 2014; Musyoka *et al.*, 2018). The distance between the ions, the drug and the observed overlap for the complexes indicate that in both IN complexes DTG remains within close proximity to the Mg ions and therefore likely remains bound to each IN system. INSTI's exert their effect through interaction with Mg ions and findings from our study shows that during the simulations DTG remained bound to Mg ions and IN protein. Additionally, we find that the same number of distinct clusters were generated at the same cut-off indicating that both systems adopt four major conformations. Furthermore, non-bonded interaction energy confirmed that DTG has a stronger interaction with HIV-1C-DNA-MG compared to the HIV-1B complex suggesting that DTG remains a feasible option for treatment in patients with WT or naturally occurring polymorphic variant sequences in HIV-1 IN proteins. In this regard we believe the present study confirms DTG as a plausible option to inhibit HIV-1C IN protein to prevent viral integration.

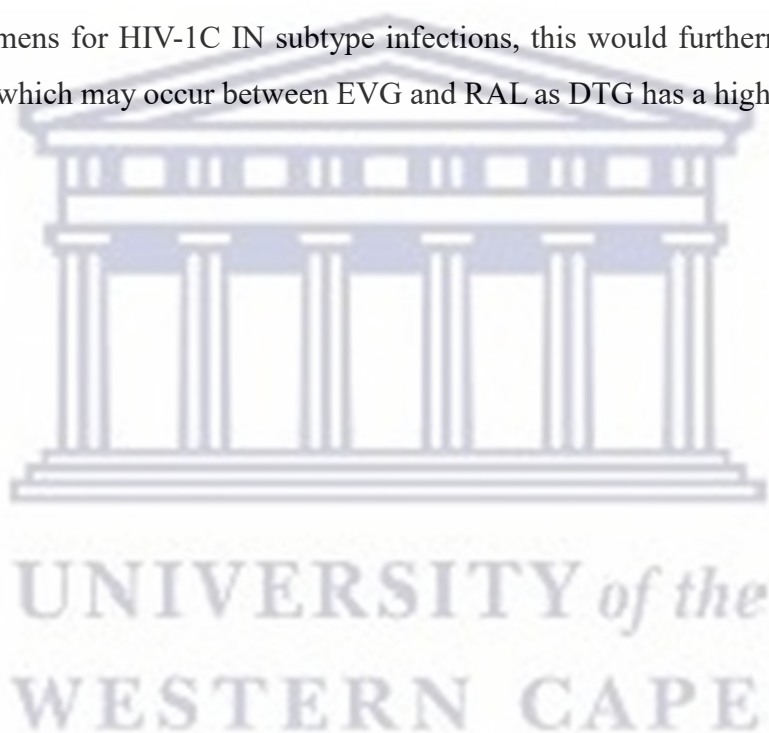
Strengths and limitations

This is the first study to conduct a comparative molecular dynamic simulation study of two HIV-1C and 1B IN protein structures with HIV-1C derived from a near complete experimentally resolved HIV-1 IN protein. In addition, the HIV-1B template structure was repaired by remodelling missing residues to obtain a complete structure. Another strength of this study is the use of SDM and mCSM webservers to assess protein folding for both HIV-1B and 1C. Finally, this was the first molecular dynamic study to investigate naturally found variants within the South African HIV treatment naïve population and predicted their effect on DTG binding. One particular limitation of the study was that it was conducted during mass DTG roll-out, newer studies should include HIV-1 treatment experienced patients or even patients failing DTG or other INSTI based treatments. The experimental model 5U1C had missing residues in the C-terminal end which could not be resolved and this could affect our predicted model due to the presence of variants in the C-terminal end. MD analysis did not include free-energy of binding calculations to determine accurate interaction energy and also the simulations were ran for only 300 ns which did not allow us to exploit the enormous energetic landscape. A portion of this research study was published in the journal *Viruses* (impact factor 4.1) and the first pages of two other published work produced during this Msc is attached in appendices.

Future work would benefit greatly by extending simulations over longer time periods and conducting free-energy of binding calculations such as Molecular mechanics Poisson-Boltzmann surface area method (mmpbsa) to accurately determine DTG binding free energy. Additional experimental validation such as functional assays which compare the activity of HIV-1C and 1B IN proteins and ligand binding assays would be valuable to compare experimental and computational predictions. The inclusion of other second generation INSTI's such as Cabotegravir and Bictegravir would also be advisable to understand drug binding to HIV-1C IN.

Conclusions

Three-dimensional structure prediction was useful in obtaining accurate IN protein models, as confirmed by protein quality assessments. Stability and interaction analysis showed contrast and therefore MD simulations was pursued. MD analysis showed that natural occurring polymorphisms within a South African HIV-1C IN protein does not result in a major change to protein stability nor prevent DTG binding based on RMSD, Rg and non-bonded interaction energy analysis. Interestingly, we did observe that DTG binding destabilizes the protein structure resulting in fewer local energy minima clusters based on PCA analysis. In conclusion, we propose that the second-generation DTG should be added to the antiviral regimens as part of first-line regimens for HIV-1C subtype infections, this would furthermore account for cross-resistance which may occur between EVG and RAL as DTG has a higher genetic barrier to resistance.



Supplementary Material

Quality analyses

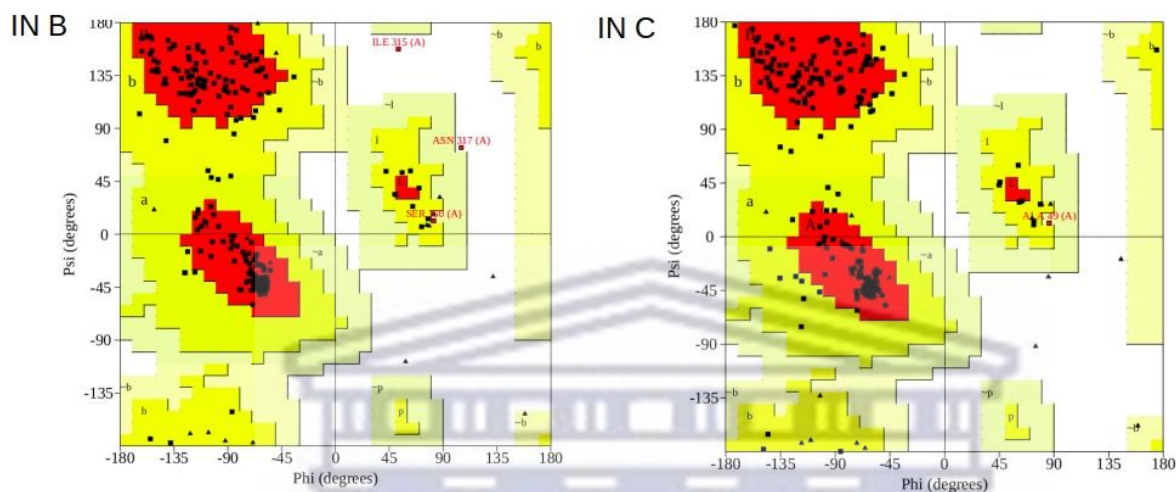


Figure S1A: Ramachandran plots. Small black squares represent residues. Red region indicates most the most favoured regions. Yellow indicates allowable regions. White space with no colour indicates disallowed regions.

Thermodynamic Parameters

Table S1: Thermodynamic parameter values indicating stability of the simulations.

Parameter	INB-DNA-MG-DTG	INC-DNA-MG-DTG	INB-DNA-MG	INC-DNA-MG
Potential	-3.83e+06	-3.82e+06	-3.83e+06	-3.82e+06
Temperature	303.15	303.15	303.15	303.15
Total-energy	-3.09e+06	-3.08e+06	-3.09e+06	-3.09e+06

Simulation Repeats

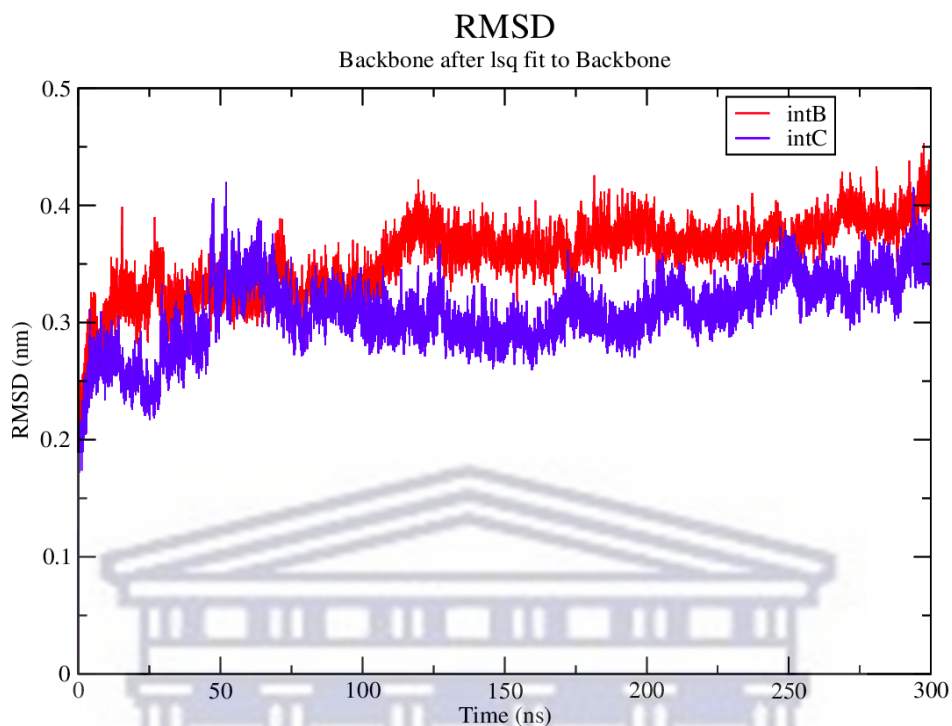


Figure S2A: Change in Backbone RMSD for HIV-1 subtype B and C IN proteins plotted over 300ns. HIV-1 IN-DNA-MG complex

Radius of gyration (total and around axes)

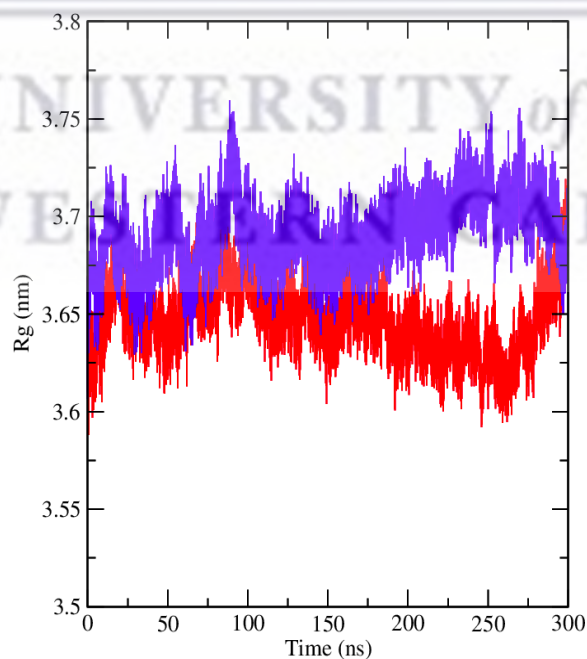


Figure S2B: Radius of gyration measured for backbone atoms for both HIV-1 subtypes B and C IN proteins plotted over 300ns. HIV-1 IN-DNA-MG complex

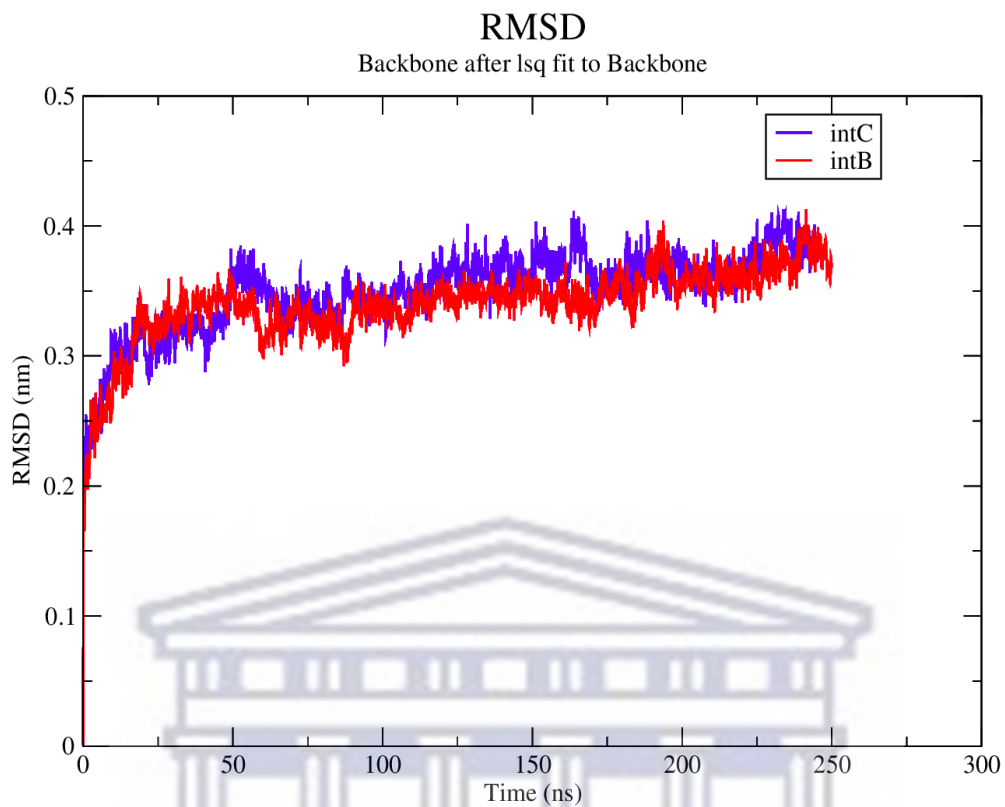


Figure S2C: Change in Backbone RMSD for HIV-1 subtype B and C IN proteins plotted over 350ns. HIV-1 IN-DNA-MG-DTG complex

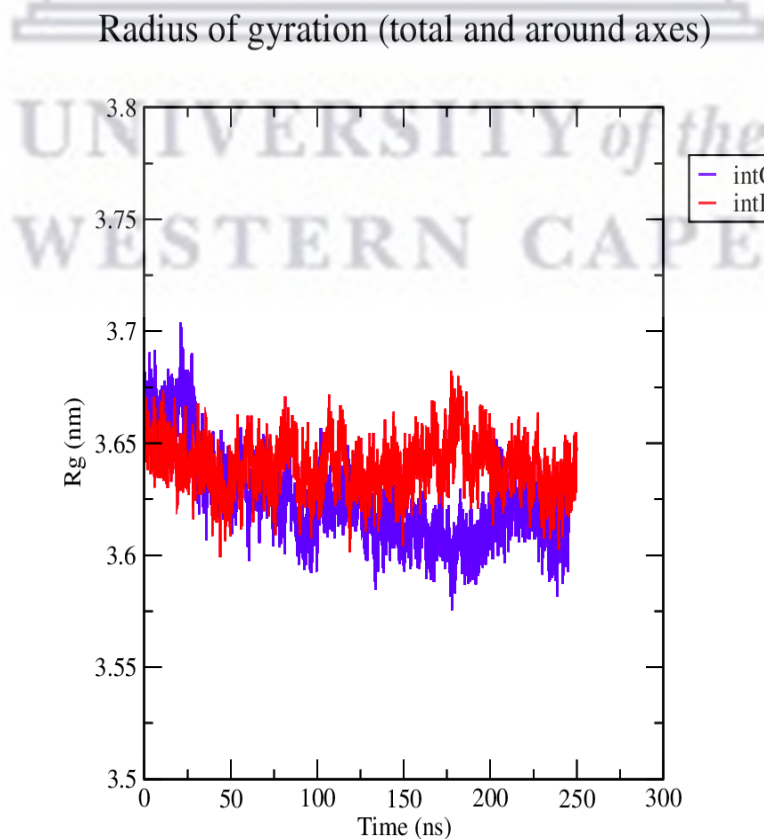


Figure S2D: Radius of gyration measured for backbone atoms for both HIV-1 subtypes B and C IN proteins plotted over 250ns. HIV-1 IN-DNA-MG-DTG complex

Complex Visualisations

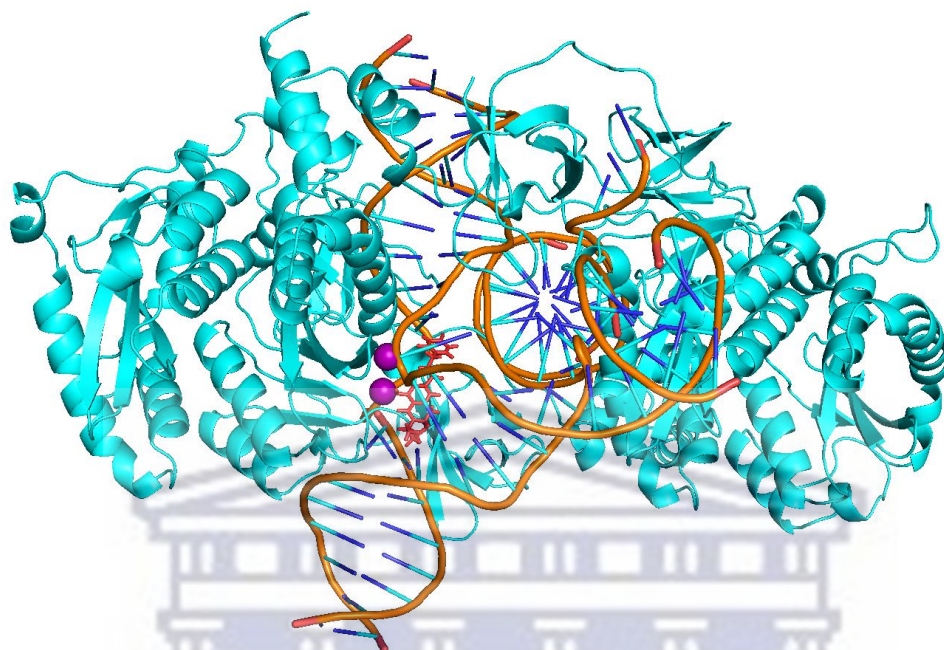


Figure S3A: Cartoon representation of the HIV-1B IN-DNA-MG-DTG complex. Magnesium ions shown as magenta spheres and DTG as red sticks.

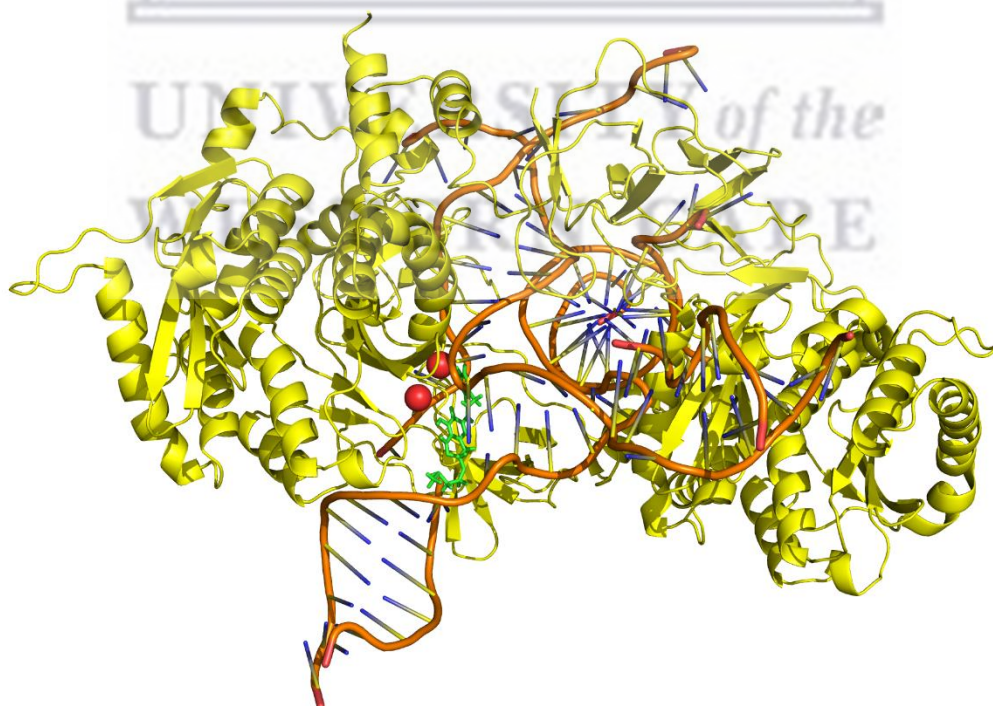


Figure S3B: Cartoon representation of the HIV-1C IN-DNA-MG-DTG complex. Magnesium ions shown as red spheres and DTG as a green stick structure.

HIV-1B IN-DNA-MG-DTG Ligand Conformations

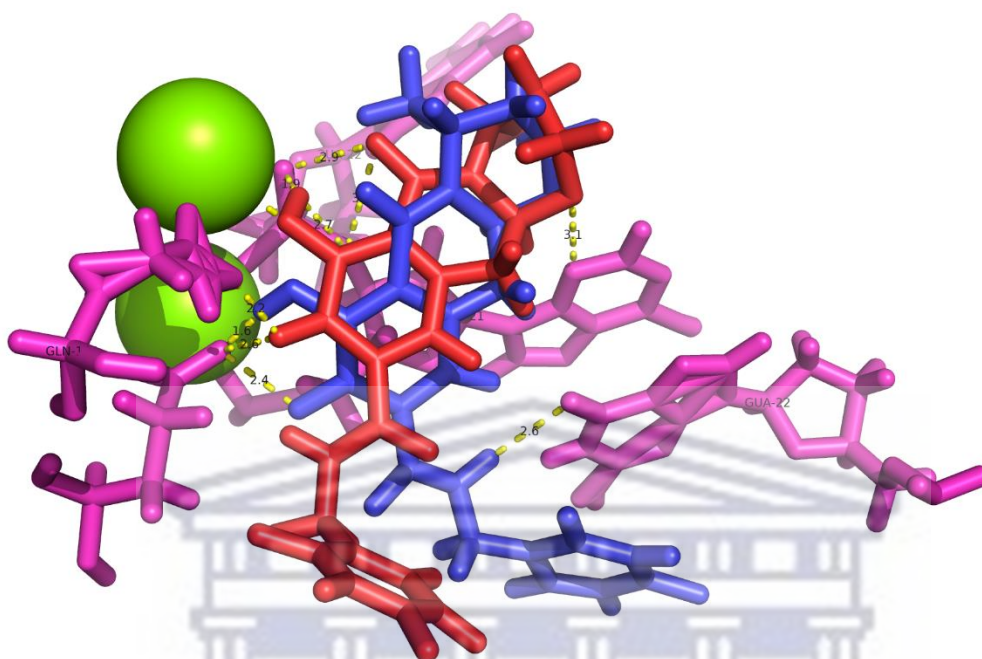


Figure 4A: Overlay of DTG ligands at timepoints 50ns and 200ns. Red Sticks indicate 50ns point and blue 200ns. Yellow dashes indicate polar interactions. Magnesium ions shown as green spheres. Protein and DNA residues shown as magenta sticks

UNIVERSITY of the
WESTERN CAPE

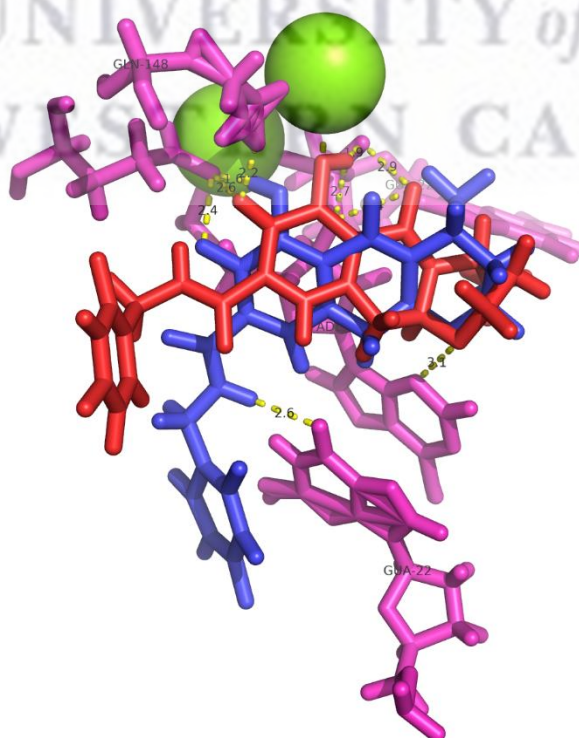


Figure 4B: Alternate angle of overlay of DTG ligands at timepoints 50ns and 200ns. Here it can be seen that in the second conformation the benzene ring of DTG is further away from the magnesium ions. <http://etd.uwc.ac.za/>

Polar Interactions

Table S5: All polar interactions calculated in PyMol over final 50ns.

Time(ns)	Polar Interactions	
	IN B	IN C
250	4(2MG,GLN148, GUA228)	8(GLN148,3GUA22, 3THY11,ASN144)
260	3(2GLU152,ADE21)	4(2THY11,2GUA22)
270	1(GLU152)	5(2THY11,ASN144, 2GUA22)
280	3(2GLU152,ADE21)	3(2GUA22,THY11)
290	4(2GLU152,CYT20,GLN 148)	4(2THY11,GLN148,GUA22)
300	3(GLN148,GLU152)	4(GLN148,GUA22,2THY11)

Number in front of brackets indicate total polar interactions. Number in front of residue indicates number of bonds formed between DTG and that residue.

References

Ambrosioni, J., Nicolás, D., Manzardo, C., Agüero, F., Blanco, J. L., Mosquera, M. M., Peñafiel, J., Gatell, J. M., Marcos, M. A., & Miró, J. M. (2017). Integrase strand-transfer inhibitor polymorphic and accessory resistance substitutions in patients with acute/recent HIV infection. *Journal of Antimicrobial Chemotherapy*, 72(1), 205–209. <https://doi.org/10.1093/jac/dkw376>

Anker, M., & Corales, R. B. (2008). Raltegravir (MK-0518): A novel integrase inhibitor for the treatment of HIV infection. *Expert Opinion on Investigational Drugs*, 17(1), 97–103. <https://doi.org/10.1517/13543784.17.1.97>

Anstett, K., Brenner, B., Mesplede, T., & Wainberg, M. A. (2017). HIV drug resistance against strand transfer integrase inhibitors. *Retrovirology*, 14(1), 36. <https://doi.org/10.1186/s12977-017-0360-7>

Arts, E. J., & Hazuda, D. J. (2012). HIV-1 Antiretroviral Drug Therapy. *Cold Spring Harbor Perspectives in Medicine*, 2(4), a007161–a007161. <https://doi.org/10.1101/cshperspect.a007161>

Bar-Magen, T., Donahue, D. A., McDonough, E. I., Kuhl, B. D., Faltenbacher, V. H., Xu, H., Michaud, V., Sloan, R. D., & Wainberg, M. A. (2010). HIV-1 subtype B and C integrase enzymes exhibit differential patterns of resistance to integrase inhibitors in biochemical assays: *AIDS*, 24(14), 2171–2179. <https://doi.org/10.1097/QAD.0b013e32833cf265>

Bar-Magen, T., Sloan, R. D., Faltenbacher, V. H., Donahue, D. A., Kuhl, B. D., Oliveira, M., Xu, H., & Wainberg, M. A. (2009). Comparative biochemical analysis of HIV-1 subtype B and C integrase enzymes. *Retrovirology*, 6(1), 103. <https://doi.org/10.1186/1742-4690-6-103>

Barreca, M. L., De Luca, L., Iraci, N., & Chimirri, A. (2006). Binding Mode Prediction of Strand Transfer HIV-1 Integrase Inhibitors Using Tn5 Transposase as a Plausible Surrogate

Model for HIV-1 Integrase. *Journal of Medicinal Chemistry*, 49(13), 3994–3997.
<https://doi.org/10.1021/jm060323r>

Barreca, M. L., Ortuso, F., Iraci, N., Luca, L. D., Alcaro, S., & Chimirri, A. (2007). Tn5 transposase as a useful platform to simulate HIV-1 integrase inhibitor binding mode. *Biochemical and Biophysical Research Communications*, 7.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1), 235-242.

Bessong, P., & Nwobegahay, J. (2013). Genetic Analysis of HIV-1 Integrase Sequences from Treatment Naive Individuals in Northeastern South Africa. *International Journal of Molecular Sciences*, 14(3), 5013–5024. <https://doi.org/10.3390/ijms14035013>

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure. *Science, New Series*, 253(5016), 164–170.

Brado, D., Obasa, A. E., Ikomey, G. M., Cloete, R., Singh, K., Engelbrecht, S., Neogi, U., & Jacobs, G. B. (2018). Analyses of HIV-1 integrase sequences prior to South African national HIV-treatment program and availability of integrase inhibitors in Cape Town, South Africa. *Scientific Reports*, 8(1), 4709. <https://doi.org/10.1038/s41598-018-22914-5>

Braun, E., Gilmer, J., Mayes, H. B., Mobley, D. L., Monroe, J. I., Prasad, S., & Zuckerman, D. M. (2019). Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*, 1(1).
<https://doi.org/10.33011/livecoms.1.1.5957>

Brenner, B. G., Thomas, R., Blanco, J. L., Ibanescu, R.-I., Oliveira, M., Mesplède, T., Golubkov, O., Roger, M., Garcia, F., Martinez, E., & Wainberg, M. A. (2016). Development of a G118R mutation in HIV-1 integrase following a switch to dolutegravir monotherapy leading to cross-resistance to integrase inhibitors. *Journal of Antimicrobial Chemotherapy*, 71(7), 1948–1953. <https://doi.org/10.1093/jac/dkw071>

Castagna, A., Maggiolo, F., Penco, G., Wright, D., Mills, A., Grossberg, R., Molina, J.-M., Chas, J., Durant, J., Moreno, S., Doroana, M., Ait-Khaled, M., Huang, J., Min, S., Song, I.,

Vavro, C., Nichols, G., Yeo, J. M., for the VIKING-3 Study Group, ... Uhlenbrauck, G. (2014). Dolutegravir in Antiretroviral-Experienced Patients With Raltegravir- and/or Elvitegravir-Resistant HIV-1: 24-Week Results of the Phase III VIKING-3 Study. *Journal of Infectious Diseases*, 210(3), 354–362. <https://doi.org/10.1093/infdis/jiu051>

Cavasotto, C. N., & Phatak, S. S. (2009). Homology modeling in drug discovery: Current trends and applications. *Drug Discovery Today*, 14(13–14), 676–683. <https://doi.org/10.1016/j.drudis.2009.04.006>

Chehadeh, W., Albaksami, O., John, S. E., & Al-Nakib, W. (2017). Resistance-Associated Mutations and Polymorphisms among Integrase Inhibitor-Naïve HIV-1 Patients in Kuwait. *Intervirology*, 60(4), 131–137. <https://doi.org/10.1159/000484692>

Chen, J. C.-H., Krucinski, J., Miercke, L. J. W., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., & Stroud, R. M. (2000). Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding. *Proceedings of the National Academy of Sciences*, 97(15), 8233–8238. <https://doi.org/10.1073/pnas.150220297>

Chen, Q., Buolamwini, J. K., Smith, J. C., Li, A., Xu, Q., Cheng, X., & Wei, D. (2013). Impact of Resistance Mutations on Inhibitor Binding to HIV-1 Integrase. *Journal of Chemical Information and Modeling*, 53(12), 3297–3307. <https://doi.org/10.1021/ci400537n>

Christ, F., & Debyser, Z. (2013). The LEDGF/p75 integrase interaction, a novel target for anti-HIV therapy. *Virology*, 435(1), 102–109. <https://doi.org/10.1016/j.virol.2012.09.033>

Christ, F., Shaw, S., Demeulemeester, J., Desimmie, B. A., Marchand, A., Butler, S., Smets, W., Chaltin, P., Westby, M., Debyser, Z., & Pickford, C. (2012). Small-Molecule Inhibitors of the LEDGF/p75 Binding Site of Integrase Block HIV Replication and Modulate Integrase Multimerization. *Antimicrobial Agents and Chemotherapy*, 56(8), 4365–4374. <https://doi.org/10.1128/AAC.00717-12>

Clavel, F., Guetard, D., Brun-Vezinet, F., Chamaret, S., Rey, M., Santos-Ferreira, M., Laurent, A., Dauguet, C., Katlama, C., Rouzioux, C., & al., e. (1986). Isolation of a new human retrovirus from West African patients with AIDS. *Science*, 233(4761), 343–346. <https://doi.org/10.1126/science.2425430>

Colovos, C., & Yeates, T. O. (1993). Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, 2(9), 1511–1519. <https://doi.org/10.1002/pro.5560020916>

Cutillas, V., Mesplede, T., Anstett, K., Hassounah, S., & Wainberg, M. A. (2015). The R262K Substitution Combined with H51Y in HIV-1 Subtype B Integrase Confers Low-Level Resistance against Dolutegravir. *Antimicrobial Agents and Chemotherapy*, 59(1), 310–316. <https://doi.org/10.1128/AAC.04274-14>

Daura, X., Gademann, K., Jaun, B., van Gunsteren, W. F., & Mark, A. E. (n.d.). *Peptide Folding: When Simulation Meets Experiment*. 5.

David, C. C., & Jacobs, D. J. (2014). Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. In D. R. Livesay (Ed.), *Protein Dynamics* (Vol. 1084, pp. 193–226). Humana Press. https://doi.org/10.1007/978-1-62703-658-0_11

DeLano, W. L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on protein crystallography*, 40(1), 82–92.

Delelis, O., Malet, I., Na, L., Tchertanov, L., Calvez, V., Marcelin, A.-G., Subra, F., Deprez, E., & Mouscadet, J.-F. (2008). The G140S mutation in HIV integrases from raltegravir-resistant patients rescues catalytic defect due to the resistance Q148H mutation. *Nucleic Acids Research*, 37(4), 1193–1201. <https://doi.org/10.1093/nar/gkn1050>

Delelis, O., Thierry, S., Subra, F., Simon, F., Malet, I., Alloui, C., Sayon, S., Calvez, V., Deprez, E., Marcelin, A.-G., Tchertanov, L., & Mouscadet, J.-F. (2010). Impact of Y143 HIV-1 Integrase Mutations on Resistance to Raltegravir In Vitro and In Vivo. *Antimicrobial Agents and Chemotherapy*, 54(1), 491–501. <https://doi.org/10.1128/AAC.01075-09>

Departments of Clinical Virology, Christian Medical College, Vellore, Sachithanandham, J., Reddy, K. K., SNHRC Vellore and Computer-Aided Drug Design and Molecular Modeling Lab, Solomon, K., Departments of Clinical Virology, Christian Medical College, Vellore, David, S., Departments of Clinical Virology, Christian Medical College, Vellore, Singh, K., SNHRC Vellore and Computer-Aided Drug Design and Molecular Modeling Lab, Ramalingam, V. V., Departments of Clinical Virology, Christian Medical College, Vellore,

Pulimood, S. A., Dermatology Department, Christian Medical College, Vellore, Abraham, O. C., Internal Medicine, Christian Medical College, Vellore, Rupali, P., Internal Medicine, Christian Medical College, Vellore, Sridharan, G., ... Departments of Clinical Virology, Christian Medical College, Vellore. (2016). Effect of HIV-1 Subtype C integrase mutations implied using molecular modeling and docking data. *Bioinformation*, 12(3), 221–230. <https://doi.org/10.6026/97320630012221>

Depatureaux, A., Quashie, P. K., Mesplède, T., Han, Y., Koubi, H., Plantier, J.-C., Oliveira, M., Moisi, D., Brenner, B., & Wainberg, M. A. (2014). HIV-1 Group O Integrase Displays Lower Enzymatic Efficiency and Higher Susceptibility to Raltegravir than HIV-1 Group M Subtype B Integrase. *Antimicrobial Agents and Chemotherapy*, 58(12), 7141–7150. <https://doi.org/10.1128/AAC.03819-14>

Dewdney, T. G., Wang, Y., Kovari, I. A., Reiter, S. J., & Kovari, L. C. (2013). Reduced HIV-1 integrase flexibility as a mechanism for raltegravir resistance. *Journal of Structural Biology*, 184(2), 245–250. <https://doi.org/10.1016/j.jsb.2013.07.008>

Di Santo, R. (2014). Inhibiting the HIV Integration Process: Past, Present, and the Future. *Journal of Medicinal Chemistry*, 57(3), 539–566. <https://doi.org/10.1021/jm400674a>

Dorward, J., Lessells, R., Drain, P. K., Naidoo, K., de Oliveira, T., Pillay, Y., Abdool Karim, S. S., & Garrett, N. (2018). Dolutegravir for first-line antiretroviral therapy in low-income and middle-income countries: Uncertainties and opportunities for implementation and research. *The Lancet HIV*, 5(7), e400–e404. [https://doi.org/10.1016/S2352-3018\(18\)30093-6](https://doi.org/10.1016/S2352-3018(18)30093-6)

Dow, D. E., & Bartlett, J. A. (2014). Dolutegravir, the Second-Generation of Integrase Strand Transfer Inhibitors (INSTIs) for the Treatment of HIV. *Infectious Diseases and Therapy*, 3(2), 83–102. <https://doi.org/10.1007/s40121-014-0029-7>

Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1), 71. <https://doi.org/10.1186/1741-7007-9-71>

Eisenberg, D., Lüthy, R., & Bowie, J. U. (1997). [20] VERIFY3D: Assessment of protein models with three-dimensional profiles. In *Methods in Enzymology* (Vol. 277, pp. 396–404). Elsevier. [https://doi.org/10.1016/S0076-6879\(97\)77022-8](https://doi.org/10.1016/S0076-6879(97)77022-8)

Engh, R. A., & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica Section A Foundations of Crystallography*, 47(4), 392–400. <https://doi.org/10.1107/S0108767391001071>

Espeseth, A. S., Felock, P., Wolfe, A., Witmer, M., Grobler, J., Anthony, N., Egbertson, M., Melamed, J. Y., Young, S., Hamill, T., Cole, J. L., & Hazuda, D. J. (2000). HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique strand transfer conformation for integrase. *Proceedings of the National Academy of Sciences*, 97(21), 11244–11249. <https://doi.org/10.1073/pnas.200139397>

Feng, L., Larue, R. C., Slaughter, A., Kessl, J. J., & Kvaratskhelia, M. (2015). HIV-1 Integrase Multimerization as a Therapeutic Target. In B. E. Torbett, D. S. Goodsell, & D. D. Richman (Eds.), *The Future of HIV-1 Therapeutics* (Vol. 389, pp. 93–119). Springer International Publishing. https://doi.org/10.1007/82_2015_439

Fransen, S., Gupta, S., Danovich, R., Hazuda, D., Miller, M., Witmer, M., Petropoulos, C. J., & Huang, W. (2009). Loss of Raltegravir Susceptibility by Human Immunodeficiency Virus Type 1 Is Conferred via Multiple Nonoverlapping Genetic Pathways. *Journal of Virology*, 83(22), 11440–11446. <https://doi.org/10.1128/JVI.01168-09>

Friedrich, B. M., Dziuba, N., Li, G., Endsley, M. A., Murray, J. L., & Ferguson, M. R. (2011). Host factors mediating HIV-1 replication. *Virus Research*, 161(2), 101–114. <https://doi.org/10.1016/j.virusres.2011.08.001>

Furtado, M. R., Callaway, D. S., Phair, J. P., Kunstman, K. J., Stanton, J. L., Macken, C. A., ... & Wolinsky, S. M. (1999). Persistence of HIV-1 transcription in peripheral-blood mononuclear cells in patients receiving potent antiretroviral therapy. *New England Journal of Medicine*, 340(21), 1614-1622.

Gelpi, J., Hospital, A., Goñi, R., & Orozco, M. (2015). Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, 37. <https://doi.org/10.2147/AABC.S70333>

Green, A. (n.d.). *Millions switched to a new ARV, but now they're gaining weight*. Citypress. Retrieved November 16, 2020, from <https://www.news24.com/citypress/news/millions-switched-to-a-new-arv-but-now-theyre-gaining-weight-20200825>

Greenwald, J., Le, V., Butler, S. L., Bushman, F. D., & Choe, S. (1999). The Mobility of an HIV-1 Integrase Active Site Loop Is Correlated with Catalytic Activity [†] · [‡]. *Biochemistry*, 38(28), 8892–8898. <https://doi.org/10.1021/bi9907173>

Grinsztejn, B., Nguyen, B.-Y., Katlama, C., Gatell, J. M., Lazzarin, A., Gonzalez, C. J., Chen, J., Harvey, C. M., & Isaacs, R. D. (2007). *Safety and efficacy of the HIV-1 integrase inhibitor raltegravir (MK-0518) in treatment-experienced patients with multidrug-resistant virus: A phase II randomised controlled trial*. 369, 9.

Gulick, R. M. (2018). Investigational antiretroviral drugs: what is coming down the pipeline. *Topics in antiviral medicine*, 25(4), 127.

Han, Y.-S., Mesplède, T., & Wainberg, M. A. (2016). Differences among HIV-1 subtypes in drug resistance against integrase inhibitors. *Infection, Genetics and Evolution*, 46, 286–291. <https://doi.org/10.1016/j.meegid.2016.06.047>

Hanley, W. D., Wenning, L. A., Moreau, A., Kost, J. T., Mangin, E., Shamp, T., Stone, J. A., Gottesdiener, K. M., Wagner, J. A., & Iwamoto, M. (2009). Effect of Tipranavir-Ritonavir on Pharmacokinetics of Raltegravir. *Antimicrobial Agents and Chemotherapy*, 53(7), 2752–2755. <https://doi.org/10.1128/AAC.01486-08>

Hare, S., Vos, A. M., Clayton, R. F., Thuring, J. W., Cummings, M. D., & Cherepanov, P. (2010). Molecular mechanisms of retroviral integrase inhibition and the evolution of viral resistance. *Proceedings of the National Academy of Sciences*, 107(46), 20057–20062. <https://doi.org/10.1073/pnas.1010246107>

Hare, Stephen, Maertens, G. N., & Cherepanov, P. (2012). 3'-Processing and strand transfer catalysed by retroviral integrase *in crystallo*: Integration reactions *in crystallo*. *The EMBO Journal*, 31(13), 3020–3028. <https://doi.org/10.1038/emboj.2012.118>

Hare, Stephen, Smith, S. J., Métifiot, M., Jaxa-Chamiec, A., Pommier, Y., Hughes, S. H., & Cherepanov, P. (2011). Structural and Functional Analyses of the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir (S/GSK1349572). *Molecular Pharmacology*, 80(4), 565–572. <https://doi.org/10.1124/mol.111.073189>

Hazuda, D. J. (2000). Inhibitors of Strand Transfer That Prevent Integration and Inhibit HIV-1 Replication in Cells. *Science*, 287(5453), 646–650. <https://doi.org/10.1126/science.287.5453.646>

Hess, B., Bekker, H., & Berendsen, H. J. C. (n.d.). LINC: A linear constraint solver for molecular simulations. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, 18(12), 10.

Hicks, C., & Gulick, R. M. (2009). Raltegravir: The First HIV Type 1 Integrase Inhibitor. *Clinical Infectious Diseases*, 48(7), 931–939. <https://doi.org/10.1086/597290>

Hightower, K. E., Wang, R., DeAnda, F., Johns, B. A., Weaver, K., Shen, Y., Tomberlin, G. H., Carter, H. L., Broderick, T., Sigethy, S., Seki, T., Kobayashi, M., & Underwood, M. R. (2011). Dolutegravir (S/GSK1349572) Exhibits Significantly Slower Dissociation than Raltegravir and Elvitegravir from Wild-Type and Integrase Inhibitor-Resistant HIV-1 Integrase-DNA Complexes. *Antimicrobial Agents and Chemotherapy*, 55(10), 4552–4559. <https://doi.org/10.1128/AAC.00157-11>

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron*, 99(6), 1129–1143. <https://doi.org/10.1016/j.neuron.2018.08.011>

Holtzer, C. D., & Roland, M. (1999). The Use of Combination Antiretroviral Therapy in HIV-Infected Patients. *Annals of Pharmacotherapy*, 33(2), 198–209. <https://doi.org/10.1345/aph.18145>

Hooft, R. W. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature*, 381(6580), 272–272. <https://doi.org/10.1038/381272a0>

Hu, W.-S., & Hughes, S. H. (2012). HIV-1 Reverse Transcription. *Cold Spring Harbor Perspectives in Medicine*, 2(10), a006882–a006882. <https://doi.org/10.1101/cshperspect.a006882>

Huang, J., & MacKerell, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25), 2135–2145. <https://doi.org/10.1002/jcc.23354>

Huang, M., Grant, G. H., & Richards, W. G. (2011). Binding modes of diketo-acid inhibitors of HIV-1 integrase: A comparative molecular dynamics simulation study. *Journal of Molecular Graphics and Modelling*, 29(7), 956–964. <https://doi.org/10.1016/j.jmgm.2011.04.002>

Iwamoto, M., Wenning, L. A., Petry, A. S., Laethem, M., De Smet, M., Kost, J. T., Breidinger, S. A., Mangin, E. C., Azrolan, N., Greenberg, H. E., Haazen, W., Stone, J. A., Gottesdiener, K. M., & Wagner, J. A. (2008). Minimal Effects of Ritonavir and Efavirenz on the Pharmacokinetics of Raltegravir. *Antimicrobial Agents and Chemotherapy*, 52(12), 4338–4343. <https://doi.org/10.1128/AAC.01543-07>

Jacobson, M. P., Friesner, R. A., Xiang, Z., & Honig, B. (2002). On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *Journal of Molecular Biology*, 320(3), 597–608. [https://doi.org/10.1016/S0022-2836\(02\)00470-9](https://doi.org/10.1016/S0022-2836(02)00470-9)

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J. F., Honig, B., Shaw, D. E., & Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2), 351–367. <https://doi.org/10.1002/prot.10613>

Janert, P. K. (2016). *Gnuplot in action* (Second edition). Manning Publications Co.

Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11), 1859–1865. <https://doi.org/10.1002/jcc.20945>

Johnson, B. C., Métifiot, M., Pommier, Y., & Hughes, S. H. (2012). Molecular Dynamics Approaches Estimate the Binding Energy of HIV-1 Integrase Inhibitors and Correlate with *In Vitro* Activity. *Antimicrobial Agents and Chemotherapy*, 56(1), 411–419. <https://doi.org/10.1128/AAC.05292-11>

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12), 2577–2637. <https://doi.org/10.1002/bip.360221211>

Karn, J., & Stoltzfus, C. M. (2012). Transcriptional and Posttranscriptional Regulation of HIV-1 Gene Expression. *Cold Spring Harbor Perspectives in Medicine*, 2(2), a006916–a006916. <https://doi.org/10.1101/cshperspect.a006916>

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>

Kessl, J. J., Jena, N., Koh, Y., Taskent-Sezgin, H., Slaughter, A., Feng, L., de Silva, S., Wu, L., Le Grice, S. F. J., Engelman, A., Fuchs, J. R., & Kvaratskhelia, M. (2012). Multimode, Cooperative Mechanism of Action of Allosteric HIV-1 Integrase Inhibitors. *Journal of Biological Chemistry*, 287(20), 16801–16811. <https://doi.org/10.1074/jbc.M112.354373>

Keyhani, S., Wang, S., Hebert, P., Carpenter, D., & Anderson, G. (2010). US Pharmaceutical Innovation in an International Context. *American Journal of Public Health*, 100(6), 1075–1080. <https://doi.org/10.2105/AJPH.2009.178491>

Kobayashi, M., Yoshinaga, T., Seki, T., Wakasa-Morimoto, C., Brown, K. W., Ferris, R., Foster, S. A., Hazen, R. J., Miki, S., Suyama-Kagitani, A., Kawauchi-Miki, S., Taishi, T., Kawasuji, T., Johns, B. A., Underwood, M. R., Garvey, E. P., Sato, A., & Fujiwara, T. (2011). *In Vitro* Antiretroviral Properties of S/GSK1349572, a Next-Generation HIV Integrase Inhibitor. *Antimicrobial Agents and Chemotherapy*, 55(2), 813–821. <https://doi.org/10.1128/AAC.01209-10>

Koes, D. R., Baumgartner, M. P., & Camacho, C. J. (2013). Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8), 1893–1904. <https://doi.org/10.1021/ci300604z>

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2), 283–291. <https://doi.org/10.1107/S0021889892009944>

Lataillade, M., Chiarella, J., & Kozal, M. J. (2007). Natural polymorphism of the HIV-1 integrase gene and mutations associated with integrase inhibitor resistance. *Antiviral Therapy*, 8.

Lesbats, P., Engelman, A. N., & Cherepanov, P. (2016). Retroviral DNA Integration. *Chemical Reviews*, 116(20), 12730–12757. <https://doi.org/10.1021/acs.chemrev.6b00125>

Lessells, R., Katzenstein, D., & de Oliveira, T. (2012). Are subtype differences important in HIV drug resistance? *Current Opinion in Virology*, 2(5), 636–643. <https://doi.org/10.1016/j.coviro.2012.08.006>

Li, M., Jurado, K. A., Lin, S., Engelman, A., & Craigie, R. (2014). Engineered Hyperactive Integrase for Concerted HIV-1 DNA Integration. *PLoS ONE*, 9(8), e105078. <https://doi.org/10.1371/journal.pone.0105078>

Liu, T. F., & Shafer, R. W. (n.d.). *Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation*. 11.

Llácer Delicado, T., Torrecilla, E., & Holguín, Á. (2016). Deep analysis of HIV-1 natural variability across HIV-1 variants at residues associated with integrase inhibitor (INI) resistance in INI-naive individuals. *Journal of Antimicrobial Chemotherapy*, 71(2), 362–366. <https://doi.org/10.1093/jac/dkv333>

Maignan, S., Guilloteau, J. P., Zhou-Liu, Q., Clément-Mella, C., & Mikol, V. (1998). Crystal structures of the catalytic domain of HIV-1 integrase free and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. *Journal of molecular biology*, 282(2), 359-368.

Marchand, C., Johnson, A. A., Karki, R. G., Pais, G. C. G., Zhang, X., Cowansage, K., Patel, T. A., Nicklaus, M. C., Burke, T. R., & Pommier, Y. (n.d.). *Metal-Dependent Inhibition of HIV-1 Integrase by α -Diketo Acids and Resistance of the Soluble Double-Mutant (F185K)*. 10.

Margolis, D. D. A. (n.d.). *Long-acting intramuscular cabotegravir and rilpivirine in adults with HIV-1 infection (LATTE-2): 96-week results of a randomised, open-label, phase 2b, non-inferiority trial*. 12.

Marinello, J., Marchand, C., Mott, B. T., Bain, A., Thomas, C. J., & Pommier, Y. (2008). Comparison of Raltegravir and Elvitegravir on HIV-1 Integrase Catalytic Reactions and on a Series of Drug-Resistant Integrase Mutants †. *Biochemistry*, 47(36), 9345–9354. <https://doi.org/10.1021/bi800791q>

Meintjes, G., Moorhouse, M. A., Carmona, S., Davies, N., Dlamini, S., van Vuuren, C., Manzini, T., Mathe, M., Moosa, Y., Nash, J., Nel, J., Pakade, Y., Woods, J., Zyl, G. V., Conradie, F., Venter, F., & Moorhouse, M. (n.d.). Adult antiretroviral therapy guidelines 2017. *Open Access*, 24.

Miri, L., Bouvier, G., Kettani, A., Mikou, A., Wakrim, L., Nilges, M., & Malliavin, T. E. (2014). Stabilization of the integrase-DNA complex by Mg²⁺ ions and prediction of key residues for binding HIV-1 integrase inhibitors: Integrase-DNA Complex and Binding of HIV-1 Integrase Inhibitors. *Proteins: Structure, Function, and Bioinformatics*, 82(3), 466–478. <https://doi.org/10.1002/prot.24412>

Molina, J.-M., LaMarca, A., Andrade-Villanueva, J., Clotet, B., Clumeck, N., Liu, Y.-P., Zhong, L., Margot, N., Cheng, A. K., & Chuck, S. L. (2012). Efficacy and safety of once daily elvitegravir versus twice daily raltegravir in treatment-experienced patients with HIV-1 receiving a ritonavir-boosted protease inhibitor: Randomised, double-blind, phase 3, non-inferiority study. *The Lancet Infectious Diseases*, 12(1), 27–35. [https://doi.org/10.1016/S1473-3099\(11\)70249-3](https://doi.org/10.1016/S1473-3099(11)70249-3)

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol*, 12(1), 88-118.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G., & Thornton, J. M. (1992). Stereochemical quality of protein structure coordinates. *Proteins: Structure, Function, and Genetics*, 12(4), 345–364. <https://doi.org/10.1002/prot.340120407>

Mouscadet, J.-F., & Tchertanov, L. (2009). Raltegravir: Molecular basis of its mechanism of action. *European Journal of Medical Research*, 14(Suppl 3), 5. <https://doi.org/10.1186/2047-783X-14-S3-5>

Murray, M. I., Markowitz, M., Frank, I., Grant, R. M., Mayer, K. H., Hudson, K. J., Stancil, B. S., Ford, S. L., Patel, P., Rinehart, A. R., Spreen, W. R., & Margolis, D. A. (2018). Satisfaction and acceptability of cabotegravir long-acting injectable suspension for prevention of HIV: Patient perspectives from the ECLAIR trial. *HIV Clinical Trials*, *19*(4), 129–138. <https://doi.org/10.1080/15284336.2018.1511346>

Nair, V. (2002). HIV integrase as a target for antiviral chemotherapy. *Reviews in Medical Virology*, *12*(3), 179–193. <https://doi.org/10.1002/rmv.350>

Nisole, S., & Saïb, A. (2004). *Early steps of retrovirus replicative cycle*. 20.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, *3*(1), 33. <https://doi.org/10.1186/1758-2946-3-33>

Pace, P., Di Francesco, M. E., Gardelli, C., Harper, S., Muraglia, E., Nizi, E., Orvieto, F., Petrocchi, A., Poma, M., Rowley, M., Scarpelli, R., Laufer, R., Gonzalez Paz, O., Monteagudo, E., Bonelli, F., Hazuda, D., Stillmock, K. A., & Summa, V. (2007). Dihydroxypyrimidine-4-carboxamides as Novel Potent and Selective HIV Integrase Inhibitors. *Journal of Medicinal Chemistry*, *50*(9), 2225–2239. <https://doi.org/10.1021/jm070027u>

Parrinello, M., & Rahman, A. (1980). Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Physical Review Letters*, *45*(14), 1196–1199. <https://doi.org/10.1103/PhysRevLett.45.1196>

Passos, D. O., Li, M., Yang, R., Rebersburg, S. V., Ghirlando, R., Jeon, Y., Shkriabai, N., Kvaratskhelia, M., Craigie, R., & Lyumkis, D. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science*, *355*(6320), 89–92. <https://doi.org/10.1126/science.aah5163>

Pires, D. E., Ascher, D. B., & Blundell, T. L. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, *30*(3), 335–342.

Pontius, J., Richelle, J., & Wodak, S. J. (1996). Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures. *Journal of Molecular Biology*, *264*(1), 121–136. <https://doi.org/10.1006/jmbi.1996.0628>

Quashie, P. K., Mesplede, T., Han, Y.-S., Oliveira, M., Singhroy, D. N., Fujiwara, T., Underwood, M. R., & Wainberg, M. A. (2012). Characterization of the R263K Mutation in HIV-1 Integrase That Confers Low-Level Resistance to the Second-Generation Integrase Strand Transfer Inhibitor Dolutegravir. *Journal of Virology*, 86(5), 2696–2705. <https://doi.org/10.1128/JVI.06591-11>

Quashie, Peter K., Han, Y.-S., Hassounah, S., Mesplède, T., & Wainberg, M. A. (2015). Structural Studies of the HIV-1 Integrase Protein: Compound Screening and Characterization of a DNA-Binding Inhibitor. *PLOS ONE*, 10(6), e0128310. <https://doi.org/10.1371/journal.pone.0128310>

Roebuck, K. A., & Saifuddin, M. (1999). Regulation of HIV-1 transcription. *Gene Expression The Journal of Liver Research*, 8(2), 67-84.

Rogers, L., Obasa, A. E., Jacobs, G. B., Sarafianos, S. G., Sönnnerborg, A., Neogi, U., & Singh, K. (2018). Structural Implications of Genotypic Variations in HIV-1 Integrase From Diverse Subtypes. *Frontiers in Microbiology*, 9, 1754. <https://doi.org/10.3389/fmicb.2018.01754>

Schames, J. R., Henchman, R. H., Siegel, J. S., Sotriffer, C. A., Ni, H., & McCammon, J. A. (2004). Discovery of a Novel Binding Trench in HIV Integrase. *Journal of Medicinal Chemistry*, 47(8), 1879–1881. <https://doi.org/10.1021/jm0341913>

Seki, T., Suyama-Kagitani, A., Kawauchi-Miki, S., Miki, S., Wakasa-Morimoto, C., Akihisa, E., Nakahara, K., Kobayashi, M., Underwood, M. R., Sato, A., Fujiwara, T., & Yoshinaga, T. (2015). Effects of Raltegravir or Elvitegravir Resistance Signature Mutations on the Barrier to Dolutegravir Resistance *In Vitro*. *Antimicrobial Agents and Chemotherapy*, 59(5), 2596–2606. <https://doi.org/10.1128/AAC.04844-14>

Serrao, E., Odde, S., Ramkumar, K., & Neamati, N. (2009). Raltegravir, elvitegravir, and metoogravir: The birth of “me-too” HIV-1 integrase inhibitors. *Retrovirology*, 6(1), 25. <https://doi.org/10.1186/1742-4690-6-25>

Shafer, R. W. (2006). Rationale and Uses of a Public HIV Drug-Resistance Database. *The Journal of Infectious Diseases*, 194(s1), S51–S58. <https://doi.org/10.1086/505356>

Shimura, K., Kodama, E., Sakagami, Y., Matsuzaki, Y., Watanabe, W., Yamataka, K., Watanabe, Y., Ohata, Y., Doi, S., Sato, M., Kano, M., Ikeda, S., & Matsuoka, M. (2008). Broad Antiretroviral Activity and Resistance Profile of the Novel Human Immunodeficiency Virus Integrase Inhibitor Elvitegravir (JTK-303/GS-9137). *Journal of Virology*, 82(2), 764–774. <https://doi.org/10.1128/JVI.01534-07>

Shimura, Kazuya, & Kodama, E. N. (2009). Elvitegravir: A New HIV Integrase Inhibitor. *Antiviral Chemistry and Chemotherapy*, 20(2), 79–85. <https://doi.org/10.3851/IMP1397>

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Genetics*, 17(4), 355–362. <https://doi.org/10.1002/prot.340170404>

Smith, S. J., Zhao, X. Z., Burke, T. R., & Hughes, S. H. (2018). Efficacies of Cabotegravir and Bictegravir against drug-resistant HIV-1 integrase mutants. *Retrovirology*, 15(1), 37. <https://doi.org/10.1186/s12977-018-0420-7>

Stats, S. A. (2018). Statistical release P0302. *Mid-year population estimates*.

Su, M., Tan, J., & Lin, C.-Y. (2015). Development of HIV-1 integrase inhibitors: Recent molecular modeling perspectives. *Drug Discovery Today*, 20(11), 1337–1348. <https://doi.org/10.1016/j.drudis.2015.07.012>

Summa, V., Petrocchi, A., Bonelli, F., Crescenzi, B., Donghi, M., Ferrara, M., Fiore, F., Gardelli, C., Gonzalez Paz, O., Hazuda, D. J., Jones, P., Kinzel, O., Laufer, R., Monteagudo, E., Muraglia, E., Nizi, E., Orvieto, F., Pace, P., Pescatore, G., ... Rowley, M. (2008). Discovery of Raltegravir, a Potent, Selective Orally Bioavailable HIV-Integrase Inhibitor for the Treatment of HIV-AIDS Infection. *Journal of Medicinal Chemistry*, 51(18), 5843–5855. <https://doi.org/10.1021/jm800245z>

Tekeste, S. S., Wilkinson, T. A., Weiner, E. M., Xu, X., Miller, J. T., Le Grice, S. F. J., Clubb, R. T., & Chow, S. A. (2015). Interaction between Reverse Transcriptase and Integrase Is Required for Reverse Transcription during HIV-1 Replication. *Journal of Virology*, 89(23), 12058–12069. <https://doi.org/10.1128/JVI.01471-15>

Tsiang, M., Jones, G. S., Niedziela-Majka, A., Kan, E., Lansdon, E. B., Huang, W., Hung, M., Samuel, D., Novikov, N., Xu, Y., Mitchell, M., Guo, H., Babaoglu, K., Liu, X., Geleziunas, R.,

& Sakowicz, R. (2012). New Class of HIV-1 Integrase (IN) Inhibitors with a Dual Mode of Action. *Journal of Biological Chemistry*, 287(25), 21189–21203. <https://doi.org/10.1074/jbc.M112.347534>

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. C. (2005). GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry*, 26(16), 1701–1718. <https://doi.org/10.1002/jcc.20291>

Van Wesenbeeck, L., Rondelez, E., Feyaerts, M., Verheyen, A., Van der Borgh, K., Smits, V., Cleybergh, C., De Wolf, H., Van Baelen, K., & Stuyver, L. J. (2011). Cross-Resistance Profile Determination of Two Second-Generation HIV-1 Integrase Inhibitors Using a Panel of Recombinant Viruses Derived from Raltegravir-Treated Clinical Isolates. *Antimicrobial Agents and Chemotherapy*, 55(1), 321–325. <https://doi.org/10.1128/AAC.01733-09>

Votteler, J., & Sundquist, W. I. (2013). Virus Budding and the ESCRT Pathway. *Cell Host & Microbe*, 14(3), 232–241. <https://doi.org/10.1016/j.chom.2013.08.012>

Wainberg, M. A. (2004). HIV-1 subtype distribution and the problem of drug resistance: *AIDS*, 18(Supplement 3), S63–S68. <https://doi.org/10.1097/00002030-200406003-00012>

Wang, J.-Y. (2001). Structure of a two-domain fragment of HIV-1 integrase: Implications for domain organization in the intact protein. *The EMBO Journal*, 20(24), 7333–7343. <https://doi.org/10.1093/emboj/20.24.7333>

Wenning, L. A., Friedman, E. J., Kost, J. T., Breidinger, S. A., Stek, J. E., Lasseter, K. C., Gottesdiener, K. M., Chen, J., Teppler, H., Wagner, J. A., Stone, J. A., & Iwamoto, M. (2008). Lack of a Significant Drug Interaction between Raltegravir and Tenofovir. *Antimicrobial Agents and Chemotherapy*, 52(9), 3253–3258. <https://doi.org/10.1128/AAC.00005-08>

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35(Web Server), W407–W410. <https://doi.org/10.1093/nar/gkm290>

Wielens, J., Crosby, I. T., & Chalmers, D. K. (2005). A Three-dimensional Model of the Human Immunodeficiency Virus Type 1 Integration Complex. *Journal of Computer-Aided Molecular Design*, 19(5), 301–317. <https://doi.org/10.1007/s10822-005-5256-2>

Wilens, C. B., Tilton, J. C., & Doms, R. W. (2012). HIV: Cell Binding and Entry. *Cold Spring Harbor Perspectives in Medicine*, 2(8), a006866–a006866. <https://doi.org/10.1101/cshperspect.a006866>

World Health Organization. (2016). *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. World Health Organization.

Worth, C. L., Preissner, R., & Blundell, T. L. (2011). SDM—a server for predicting effects of mutations on protein stability and malfunction. *Nucleic acids research*, 39(suppl_2), W215-W222.

Zheng, R., Jenkins, T. M., & Craigie, R. (1996). Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity. *Proceedings of the National Academy of Sciences*, 93(24), 13659–13664. <https://doi.org/10.1073/pnas.93.24.13659>



Appendices



UNIVERSITEIT
STELLENBOSCH
UNIVERSITY
**Approval Letter
Progress Report**

07/04/2020

Project ID: 2215

Ethics Reference No: N15/08/071

Project Title: Tracking the molecular epidemiology and resistance patterns of HIV-1 in South Africa.

Dear Dr. Graeme Jacobs

We refer to your request for an extension/annual renewal of ethics approval dated 28/03/2020.

The Health Research Ethics Committee reviewed and approved the annual progress report through an expedited review process.

The approval of this project is extended for a further year.

Approval date: 07 April 2020

Expiry date: 06 April 2021

1. Kindly note that although the study has been granted ethics approval, the study may not proceed during the current national lockdown as an embargo has been placed on studies that require interaction with research participants in order to prevent potential harm to participants.
2. HREC will publish on the HREC website a date when the said embargo is to be lifted taking into consideration the best interest of participants and national interests around COVID-19.

Kindly be reminded to submit progress reports two (2) months before expiry date.

Where to submit any documentation

Kindly note that the HREC uses an electronic ethics review management system, *Infonetica*, to manage ethics applications and ethics review process. To submit any documentation to HREC, please click on the following link: <https://appliedethics.sun.ac.za>.

Please remember to use your Project Id 2215 and ethics reference number N15/08/071 on any documents or correspondence with the HREC concerning your research protocol.

Yours sincerely,

Mrs. Brightness Nxumalo
Coordinator: Health Research Ethics Committee 2

National Health Research Ethics Council (NHREC) Registration Number:
REC-130408-012 (HREC1)•REC-230208-010 (HREC2)

Federal Wide Assurance Number: 00001372
Office of Human Research Protections (OHRP) Institutional Review Board (IRB) Number:
IRB0005240 (HREC1)•IRB0005239 (HREC2)

The Health Research Ethics Committee (HREC) complies with the SA National Health Act No. 61 of 2003 as it pertains to health research. The HREC abides by the ethical norms and principles for research, established by the World Medical Association (2013), Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects; the South African Department of Health (2006), Guidelines for Good Practice in the Conduct of Clinical Trials with Human Participants in South Africa (2nd edition); as well as the Department of Health (2015), Ethics in Health Research: Principles, Processes and Structures (2nd edition).

The Health Research Ethics Committee reviews research involving human subjects conducted or supported by the Department of Health and Human Services, or other federal departments or agencies that apply the Federal Policy for the Protection of Human Subjects to such research (United States Code of Federal Regulations Title 45 Part 46); and/or clinical investigations regulated by the Food and Drug Administration (FDA) of the Department of Health and Human Services.



Article

Structural Comparison of Diverse HIV-1 Subtypes using Molecular Modelling and Docking Analyses of Integrase Inhibitors

Darren Isaacs ^{1,†}, Sello Given Mikasi ^{2,†}, Adetayo Emmanuel Obasa ² , George Mondinde Ikomey ³, Sergey Shityakov ^{4,5} , Ruben Cloete ^{1,*} and Graeme Brendon Jacobs ^{2,*}

¹ South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town 8000, South Africa; 3433660@myuwc.ac.za

² Division of Medical Virology, Department of Pathology, Faculty of Medicine and Health Sciences, Stellenbosch University, Francie van Zijl Avenue, P.O. Box 241, Cape Town 8000, South Africa; mikasi@sun.ac.za (S.G.M.); obasa@sun.ac.za (A.E.O.)

³ Centre for the Study and Control of Communicable Diseases (CSCCD), University of Yaoundé 1, Yaoundé P.O. Box 8445, Cameroon; mondinde@yahoo.com

⁴ Department of Psychiatry & Mind-Body Interface Laboratory (MBI-Lab), China Medical University Hospital, Taichung 404, Taiwan; shityakoff@hotmail.com

⁵ Department of Bioinformatics, University of Würzburg, Würzburg 97074, Germany

* Correspondence: ruben@sanbi.ac.za (R.C.); graeme@sun.ac.za (G.B.J.); Tel.: +27-21-938-9744 (G.B.J.)

† These authors contributed equally to this work.

Received: 10 June 2020; Accepted: 5 August 2020; Published: 26 August 2020



Abstract: The process of viral integration into the host genome is an essential step of the HIV-1 life cycle. The viral integrase (IN) enzyme catalyzes integration. IN is an ideal therapeutic enzyme targeted by several drugs; raltegravir (RAL), elvitegravir (EVG), dolutegravir (DTG), and bictegravir (BIC) having been approved by the USA Food and Drug Administration (FDA). Due to high HIV-1 diversity, it is not well understood how specific naturally occurring polymorphisms (NOPs) in IN may affect the structure/function and binding affinity of integrase strand transfer inhibitors (INSTIs). We applied computational methods of molecular modelling and docking to analyze the effect of NOPs on the full-length IN structure and INSTI binding. We identified 13 NOPs within the Cameroonian-derived CRF02_AG IN sequences and further identified 17 NOPs within HIV-1C South African sequences. The NOPs in the IN structures did not show any differences in INSTI binding affinity. However, linear regression analysis revealed a positive correlation between the K_i and EC50 values for DTG and BIC as strong inhibitors of HIV-1 IN subtypes. All INSTIs are clinically effective against diverse HIV-1 strains from INSTI treatment-naïve populations. This study supports the use of second-generation INSTIs such as DTG and BIC as part of first-line combination antiretroviral therapy (cART) regimens, due to a stronger genetic barrier to the emergence of drug resistance.

Keywords: integrase; naturally occurring polymorphisms; HIV-1; molecular modelling; molecular docking; diversity



UNIVERSITY *of the*
WESTERN CAPE

HIV-1 Drug Resistance Mutation Analyses of Cameroon-Derived Integrase Sequences

Sello Given Mikasi,^{1,*} Darren Isaacs,^{2,*} George Mondinde Ikomey,³
Henerico Shimba,^{1,4} Ruben Cloete,³ and Graeme Brendon Jacobs¹

Abstract

HIV-1 integrase (IN) is a primary target for combination antiretroviral therapy. Only a limited number of studies report on the emergence of resistance-associated mutations (RAMs) in Cameroon. We observed that 1.4% of sequence from treatment-naïve patients had IN strand transfer inhibitor (INSTI) RAMs. These mutations confer resistance to raltegravir and elvitegravir. We also observed that 10.1% of the sequences have INSTI accessory RAMs. HIV-1 CRF02_AG was the predominant subtype (44.7%) in this study analyses. The occurrence of INSTI RAMs among the sequences at baseline needs to be monitored carefully.

Keywords: HIV-1, integrase, resistance, Cameroon, diversity, CRF02_AG

HIV/AIDS REMAINS a health concern that endangers the lives of millions of people worldwide. HIV-1 has three essential enzymes used for its replication, namely, protease, reverse transcriptase, and integrase (IN). HIV IN is a 32 kDa protein encoded at the 3'-end of the HIV pol gene.¹ IN consists of 288 amino acids, and can be divided into three structural and functional domains: a zinc-binding N-terminal domain, a catalytic core domain, and a DNA binding C-terminal domain.²

IN strand transfer inhibitors (INSTIs) are a class of antiretroviral drugs used to target the IN enzyme, to prevent viral complementary DNA integration into the host genome. IN resistance-associated mutations (RAMs) located at amino acid residues Y143C, Q148H, and N155H have shown to be the primary/major pathways that induce resistance to the INSTI raltegravir (RAL). Primary mutations that induce resistance against elvitegravir (EVG) include T66I, E92Q, Q148H/R/K, and N155H. RAL and EVG present broad cross-resistance.³ Dolutegravir (DTG) retains its activity against viruses that exhibited RAMs against RAL and EVG.⁴ However, combinations of mutations at position G140 and Q148 can result in significant drug resistance to DTG. Furthermore, DTG monotherapy in treated naïve patients has been reported to be associated with more frequent development of RAMs such as R263K, G118R, and S230.⁵

Naturally occurring polymorphisms (NOPs) have been identified at sites associated with secondary RAMs: L74V, M154I, I72V, and T125A.⁶ Studies have shown that difference across HIV-1 subtypes may play a role, by influencing NOPs in facilitating or compensating the development of major resistance pathway to antiretroviral drugs.⁷ Owing to the development of drug resistance across currently available drugs, the World Health Organization (WHO) has put forth the use of the INSTI DTG as part of suggested first-line therapy. Therapeutic advantage of DTG is its ability to maintain high potencies against mutant strains of HIV-1 that previously exhibited resistance to other combination antiretroviral therapy (cART) drugs.⁷ In Cameroon, there is a lack of data regarding the emergence of RAMs against INSTIs; particularly DTG as it is anticipated to be used in future treatment regimens.⁸ Therefore, it is of paramount importance to identify RAMs from samples sequenced before the initiation of INSTI therapy in Cameroon.

In this study, we evaluated only IN sequences from INSTIs-naïve HIV-infected individuals from Cameroon, dating between 1994 and 2009, obtained from the Los Alamos National Library (LANL) HIV-1 database (https://www.hiv.lanl.gov/components/sequence/HIV_search.com), accessed between June 10, 2019, and June 12, 2019. We

¹Division of Medical Virology, Department of Pathology, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

²South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa.


³Centre for the Study and Control of Communicable Diseases, Faculty of Medicine and Biomedical Sciences, University of Yaoundé I, Yaoundé, Cameroon

RESEARCH ARTICLE

Open Access

Interaction analysis of statistically enriched mutations identified in Cameroon recombinant subtype CRF02_AG that can influence the development of Dolutegravir drug resistance mutations



Sello Given Mikasi¹, Darren Isaacs², Rumbidzai Chitongo², George Mondide Ikomey³, Graeme Brendon Jacobs^{1*} and Ruben Cloete^{2*} 

Abstract

Background: The Integrase (IN) strand transfer inhibitor (INSTI), Dolutegravir (DTG), has been given the green light to form part of first-line combination antiretroviral therapy (cART) by the World Health Organization (WHO). DTG containing regimens have shown a high genetic barrier against HIV-1 isolates carrying specific resistance mutations when compared with other class of regimens.

Methods: We evaluated the HIV-1 CRF02_AG IN gene sequences from Cameroon for the presence of resistance-associated mutations (RAMs) against INSTIs and naturally occurring polymorphisms (NOPs), using study sequences ($n = 20$) and ($n = 287$) sequences data derived from HIV Los Alamos National Laboratory database. The possible impact of NOPs on protein structure caused by HIV-1 CRF02_AG variations was addressed within the context of a 3D model of the HIV-1 IN complex and interaction analysis was performed using PyMol to validate DTG binding to the Wild type and seven mutant structures.

Results: We observed 12.8% (37/287) sequences to contain RAMs, with only 1.0% (3/287) of the sequences having major INSTI RAMs: T66A, Q148H, R263K and N155H. Of these, 11.8% (34/287) of the sequences contained five different IN accessory mutations; namely Q95K, T97A, G149A, E157Q and D232N. NOPs occurred at a frequency of 66% on the central core domain (CCD) position, 44% on the C-terminal domain (CTD) position and 35% of the N-terminal domain (NTD) position. The interaction analysis revealed that DTG bound to DNA, 2MG ions and DDE motif residues for T66A, T97A, Q148H, N155H and R263K comparable to the WT structure. Except for accessory mutant structure E157Q, only one MG contact was made with DTG, while DTG had no MG ion contacts and no DDE motif residue contacts for structure D232N.

(Continued on next page)

* Correspondence: ruben@sanbiac.za

¹Graeme Brendon Jacobs is deceased.

²South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Robert Sobukwe Rd, Bellville, P.O. Box X17, Cape Town 7535, South Africa

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons