DIVERSITY IN APOBEC3 AND CCR5 HOST GENES AND HIV-1 IN A SOUTH AFRICAN POPULATION

By

Nontokozo D. Matume (11575283)

A thesis submitted in fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Microbiology

To

The Department of Microbiology, School of Mathematical and Natural Sciences, University of Venda

South Africa

Promoter: Prof Pascal O. Bessong (University Of Venda)

Co-Promoters: Prof Marie-Louise Hammarskjöld and Prof David Rekosh (University Of Virginia)

September 2018

DECLARATION

I, **NONTOKOZO D. MATUME**, hereby declare that this thesis for the award of Doctor of Philosophy in Microbiology at the University of Venda, hereby submitted by me, has not been submitted for a degree at this or any other University, and that it is my work in design and execution and that all the reference materials contained therein have been duly acknowledged.

.

STUDENT: Nontokozo D. Matume (11575283)

Signature..... Date.....

ACKNOWLEDGEMENTS

Glory be to God Almighty for all He has done, it is by his grace, endless love and divine purpose that this work came to completion.

Promoters

Courage, inspiration and guidance are essential for a production of quality and well-trained students. To Prof P.O. Bessong, thank you so much for being all of these, I'm so grateful that you had such great faith in me and gave me an opportunity to learn and acquire skills in the field of research. I would not have asked for a better supervisor, you are simply the best.

To my Co-promoters; Prof M-L Hammarskjöld and Prof D. Rekosh and the Ham-Rek lab at the University of Virginia, thank you so much for helping me acquire skills and knowledge in Next Generation Sequencing technology. This work would have not come to pass if not for your ideas and contribution to make it a success.

Clinicians and study participants

My sincere gratitude goes to the participants for consenting to participate in the study and the clinicians for helping us acquire samples and clinical data.

Family

I am grateful to the best family in the whole world with special thanks to my brothers; Xolani Fakude, Mzwakhe Maile, Sister; Lindokuhle Mkhabela, Aunts; Sokolile Mkhabela, Thulisile Mkhabela and Gugu Bulunga, they are awesome, amazing and a blessing in my life. Thank you so much for your support throughout. And to my mother (Priscilla Mkhabela), I would not trade you for anything. You are the best source of inspiration and the best mother ever. Thank you for giving us the chance to a better education.

Colleagues and friends

I would also like to thank the HIV/AIDS & Global Health research programme team, with special thanks to Dr Mavhandu, Dr Gachara, Dr Masebe, Elizabeth Etta, Mukhetwa Munzhedzi, Tambe Lisa, Bixa Ogola and Tshifhiwa Tambani for their support, I have learnt so much from all of you and thank you for being patient with me.

My heart felt gratitude goes to Dr Denis Tebit, Jing Huang and Laurie Gray for their efforts and time.

I would like to thank all my friends, Nonhlanhla Mashabane, Zinhle Dlamini, Philasande Mahlangu for their support and encouragements;

Special thanks to my friend Renay Ngobeni, she has not only been a support system but a part of my life throughout my studies.

I would also express my gratitude to Pastor and Mrs Mashabane and Pastor and Mrs Tshikororo for their prayers and support through out my studies.

Funders

I would also like to thank the following funding bodies for their support throughout the study;

1. Award # 95707 S&F- Innovation Doctoral Scholarships, National Research foundation

2. Award # D43 TW006578, Fogarty International Center/ NIH.

3. Award # RCDI 57009, South African Medical Research Council

4. Award # SMNS/16/MBY/28, University of Venda Higher Degree's Committee

DEDICATION

This thesis is dedicated to my mother:

Priscilla Mkhabela

ABSTRACT OF THE STUDY

Introduction

Human Immunodeficiency Virus (HIV-1) continues to be a global public health concern, even though Antiretroviral drugs (ARV), especially Highly Active Antiretroviral Therapy (HAART) has significantly reduced morbidity and mortality due to AIDS globally in developed and developing countries. However, there is still a great need to explore every avenue for new therapeutic interventions due to the limitations and side effects of HAART.

Potential major breakthroughs for future therapeutic development were the discoveries more than 10 years ago of the role of HIV-1 co-receptors and anti-viral activities of host restriction factors such as APOBEC3G protein, which is a member of the DNA cytosine deaminase family.

Entry of HIV in to the host cell is through the attachment of the viral envelope glycoprotein to the CD4 receptor, and subsequent interaction, mainly with either CCR5 or CXCR4 co-receptors. Inhibitors, such as Maraviroc, which binds to CCR5 inhibiting entry of CCR5 utilizing viruses (R5 viruses), is currently reserved for salvage therapy in many countries including South Africa. In the course of HIV infection, CXCR4 utilizing viruses (X4 viruses) may emerge and outgrow R5 viruses, and potentially limit the effectiveness of Maraviroc.

Several host cell APOBEC3 genes (A3D, A3F, A3G and A3H) have been shown to restrict HIV, and the HIV viral infectivity factor (Vif) protein serves to antagonize the action of APOBEC3 proteins, promoting viral replication. The CCR5 co-receptor and the HIV Env V3 loop have also been documented as playing roles in HIV-1 disease progression. The interplay between host and viral genes still needs widespread attention, given that disease outcomes of HIV depend on many factors, including host cell genetics. Since the discovery of these genes and their role in HIV replication, many studies have been conducted that show their association with viral polymorphism. The polymorphisms found in host cell genes can have significant effects on viral replication, transmission and fitness and can also contribute to the overall diversity in HIV-1 populations. It is hypothesized that there are significant polymorphisms in HIV-1 and cellular genes that may differ among different populations. Population-based studies have tried to establish a relationship between host factors such as APOBEC3 and CCR5 polymorphism and the rate of disease progression, but most studies have focused on Caucasian populations. In

contrast, little information is available for the effects of variation in these genes in African populations such as South Africa, where the HIV epidemic has expanded at an alarming rate. Although several population studies have focused on African Americans, these do not give us a complete picture of the potential variation in Africans, though the studies can be a good guide on which to base additional studies. A more comprehensive analysis involving different African populations will likely provide a better understanding of the mechanisms underlying host-pathogen interactions, especially in view of the fact that African Americans are primarily infected with HIV subtype B, which is rarely seen in Africa.

Methodology

This study characterized the genetic variability of the APOBEC3 D, F, G and H genes as well as the HIV-1 *vif*, in an ethnically diverse HIV-1 infected South African cohort using Next Generation Sequencing (NGS). In addition, polymorphism in CCR5 was analyzed in conjunction with an analysis of the V3 loop sequences in HIV-1 from this cohort. Genomic DNA was extracted from peripheral blood mononuclear cells (PBMCs) of 192 HIV-1 infected drug-experienced individuals who presented for routine care at the HIV/AIDS Prevention Group Wellness Clinic (HAPG) in Bela-Bela, Donald Fraser Hope Clinic (DFHC) in Vhufhuli and in local clinics in the Vhembe district of Limpopo Province, South Africa. Next generation sequencing custom based (Tn5 tagmentation and amplicon based) protocols to prepare libraries for host and HIV-1 genes were developed and validated with commercially available library preparation kits. The Tn5 tagmentation methods were used for longer DNA fragments and the custom amplicon based methods were used mainly for the shorter DNA fragments.

To determine the variability of the APOBEC3 and CCR5 host genes, gene-specific primers were designed to amplify complete 12.16 kb A3D, 13.31 kb A3F, 10.74 kb A3G, 6.8 kb A3H and 1.3 kb CCR5 genes targeting the regions containing the exons. Libraries for the resulting amplicons were prepared using Tn5 transposase tagmentation methods and sequenced on an NGS Illumina MiSeq platforms generating millions of reads with good read coverage for variant calling. Single nucleotide polymorphisms (SNPs) and indels were determined, verified in dbSNPs and compared to SNPs in other populations reported in the 1000 Genome Phase III and HapMap. A Chi-square goodness-of-fit was used to verify if whether observed genotype frequencies were in agreement with the Hardy-Weinberg Equilibrium. Haplotypes and Linkage disequilibrium were inferred to determine SNP association.

The HIV-1 *vif* and *env* V3 loop genes were also sequenced to determine their degree of variability of these genes and to infer co-receptor usage in the South African population. Gene-specific primers were designed to amplify the 579 bp Vif region and 440 bp containing the 105 bp V3 loop. Sequencing libraries from the resulting amplicons were prepared using either the Tn5 transposase or custom-based library preparation methods and sequenced on either an Illumina MiSeq or a MiniSeq platform generating millions of reads with good read coverage for variant calling. Phylogenetic analysis was done to determine the relatedness of the sequences. Major and minor variants were determined for HIV-1 and *env* V3 loop quasispecies was analysed for co-receptor usage; in an effort to draw inferences for the subsequent utility of Maraviroc as salvage therapy in South Africa.

Results and Discussion

Next generation library preparation; Tn5 tagmentation based and custom amplicon based protocols to sequence host and HIV genes were successfully developed and used to sequence and characterize variability in host cell APOBEC3D, F, G H, CCR5 and the HIV-1 *vif gene* and the V3 loop region of the *env* gene.

The HIV-1 *env* V3 loop sequences generated (and quasispecies analyzed) were used to infer co-receptor usage in treatment-experienced individuals; in an effort to draw inferences for the subsequent utility of Maraviroc as salvage therapy in South Africa. Quality V3 loop sequences were obtained from 72 patients, with 5 years (range: 0-16) median duration on treatment. Subtypes A1, B and C viruses were identified at frequencies of 4% (3/72), 4% (3/72) and 92% (66/72) respectively. Fifty four percent (39/72) of patients were predicted to exclusively harbor R5 viral quasispecies; and 21% (15/72) to exclusively harbor X4 viral quasispecies. Twenty five percent of patients (18/72) were predicted to harbor a dual/mixture of R5X4 quasispecies. Of these 18 patients, about 28% (5/18) were predicted to harbor the R5+X4, a mixture with a majority R5 and minority X4 viruses, while about 72% (13/18) were predicted to harbor the R5X4+ a mixture with a majority X4 and minority R5 viruses. The proportion of all patients who harboured X4 viruses either exclusively or dual/mixture was 46% (33/72). Thirty-five percent (23/66) of the patients who were of HIV-1 subtype C were predicted to harbor X4 viruses (χ^2=3.58; p=0.058), and 57% of these (13/23) were predicted to harbor X4 viruses exclusively. $CD4^+$ cell count less than 350 cell/µl was associated with the presence of X4 viruses (χ^2=4.99; p=0.008). The effectiveness of Maraviroc as a component in salvage therapy may be compromised for a significant number of chronically infected patients harboring CXCR4 utilizing

viruses in the study cohort.

Although from the current study a subset of patients harboring CCR5 utilizing viruses may benefit from Maraviroc, characterizing and identifying if variation in CCR5 are located at Maraviroc binding sites was of importance to investigate. The following variants; P35P, S75S, Y89Y, A335V and Y339F and their varying frequencies were detected in the CCR5 gene. The A335V variant was detected at a higher frequency of 17.4% (29/167). The G265S variant is reported for the first time in this study at 0.6% (1/167) frequency. The SNPs detected were in strong LD (D'= 1, R^2 = 0.0) with minor deviation from the Hardy-Weinberg Equilibrium. These variants were not located at the binding motif of Maraviroc. The variants A335V and Y339F were detected at a higher frequency in this study than previously reported in South Africa.

Variability in APOBEC3 host cell genes was also characterized in our study cohort. The following APOBEC3 variants compared to the GRCh37 consensus sequence were detected: R97C, R248K and T316T in A3D; R48P, A78V, A108S, S118S, R143R, I87L, Q87L, V231I, E245E, S229S, Y307C and S327S in A3F; S60S, H186R, R256H, Q275E and G363R in A3G and N15Δ, G105R, K140E, K121D, E178D in A3H. Minor allele frequency variants (MAF<5%); L221L, T238I, C224Y and C320Y in A3D; I87L, P97L and S229S in A3F; R256H, A109A, F119F and L371L in A3G, which are frequent in the European population, were also detected. In addition, novel R6K, L221R and T238I variants in A3D and I117I in A3F were detected. Most of the SNPs were in strong LD with minor deviation from the Hardy-Weinberg Equilibrium. Four, six, four, and three haplotypes were identified for A3D, A3F, A3G, and A3H respectively. In general, polymorphism in A3D, 3F, 3G and 3H were higher in our South African cohort than previously reported among other African, European and Asian populations.

The APOBEC3 antagonist HIV-1 *vif* gene was also sequenced to determine the level of diversity in a South African population and also correlated with APOBEC3 variation. Functional Vif without frameshift mutation was observed in all samples except in 4 samples. The functional domain and motifs, such as Zn binding motifs, proline-rich domain, human casein kinase, and the N and C-terminal CBF interaction site were highly conserved. APOBEC binding motifs and the nuclear localization signal were less conserved in the South African HIV-1 Vif. APOBEC3 H variation strongly correlates with Vif variation. All the Vif sequences were subtype C, except one sample, which was identified as an A1/C recombinant. The *vif* gene in a South African population was under purifying selection, with the dS= 0.2581 and dN= 0.0684 and the dN/dS value = 0.265. There is a high genetic diversity in the South African *vif* gene, which may

influence the neutralization and restriction of APOBEC genes.

Conclusions

In conclusion, the protocols developed in this study can be applied to amplify and sequence any host and HIV-1 genes of interest allowing much deeper and more sensitive profiling of host gene and HIV-1 genetic diversity.

Our findings show that a highly significant number of chronically HIV-1 subtype C infected patients in Maraviroc-free treatment harbor CXCR4 utilizing viruses. The data is useful in the consideration of whether to include entry antagonists such as Maraviroc in alternative forms of treatment for patients failing second line treatment regimen in the study setting. The determination of co-receptor usage prior to initiation of therapy consisting of Maraviroc is suggested.

Variation in the CCR5 coding region were observed at higher frequencies compare to other studies conducted in South African populations at different locations. This data may suggest that different populations in South Africa have different SNP frequencies. All the polymorphisms identified in the study were not located at the Maraviroc binding motif, therefore the subset of patient infected by R5 viruses may benefit from this drug.

We have shown that significant APOBEC3 variation exists among an ethnically diverse population of South Africa by providing extensive data for 4 different A3 genes that are known to restrict HIV infection, but have only been sparsely studied in African populations. This study provides a baseline for future studies that would functionally characterize SNPs identified in this population, in order to understand the role of novel and/or low frequency variants observed. *Ex vivo* and *in vivo* studies will increase our understanding of how these variants might have cumulatively impacted the epidemic in Northern South Africa.

This study also shows that there is a high level of HIV-1 Vif diversity in the study area. This diversity may impact the expression and packaging of Vif proteins, and the infectivity of HIV. In addition, a significant correlation was observed between HIV-1 Vif variation and APOBEC3 H haplotypes.

SCIENTIFIC COMMUNICATIONS FROM THIS STUDY

Published article

- **Nontokozo D. Matume**, Denis M. Tebit, Laurie R. Gray, Marie-Louise Hammarskjold, David Rekosh, Pascal O. Bessong. Next Generation Sequencing reveals a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced South African individuals. **Journal of Clinical Virology,** https://doi.org/10. 1016/ j. jcv. 2018. 02. 008.

Manuscript prepared for submission

- **Nontokozo D. Matume**, Denis M. Tebit, Laurie R. Gray, Stephen D. Turner, David Rekosh, Pascal O. Bessong, Marie-Louise Hammarskjöld. Characterization of APOBEC3 variation in South African population. (**Submitted to BMC Medical Genetics, 2018**).

Conference presentations

- **Nontokozo D. Matume**, Denis M. Tebit, Laurie R. Gray, David Rekosh, Pascal O. Bessong, Marie-Louise Hammarskjöld. Genetic variation within the coding region of CCR5 in chronically HIV-1 infected South African individuals. **(Abstract submitted)**. 2018 Annual Early Career Scientist Convention (ECSC). Cape Town, 17th & 18 October 2018

- **Nontokozo D. Matume**, Denis M. Tebit, Laurie R. Gray, Stephen D. Turner, David Rekosh, Pascal O. Bessong, Marie-Louise Hammarskjöld. Genetic Characterization of HIV-1 *vif* and Correlation with APOBEC3 Variation in Chronically Infected South African Treatment Experienced Patients. (**Oral presentation**). The 1st Pan African International Research Congress on Knowledge Generation and Dissemination. Grand Royal Swiss Hotel, Kisumu, Kenya. 17-25th June 2018.

- **Nontokozo D. Matume**, Denis M. Tebit, Pascal O. Bessong. Next Generation

Sequencing reveals a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced South African individuals. **(Oral presentation)**. 19th International Conference on AIDS and STIs in Africa (ICASA 2017). Abidjan, Côte D'Ivoire. 4- 9 December 2017.

- **Nontokozo D. Matume**, Denis M. Tebit, David Rekosh, Marie-Louise Hammarskjöld, Pascal O. Bessong. Characterization of APOBEC3 variation in a South African population. (**Poster presentation**). EMBL conference Mammalian Genetics and Genomics: From molecular mechanism to translation applications. Heidelberg, Germany. 23- 28 October 2017.

- **N.D Matume**, D Tebit, L.R Gray, M Hammarskjold, D Rekosh, P Bessong. Next Generation Sequencing reveals a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced South African individuals (**Oral presentation**). International Symposium on Global Health Research in Africa-FIC/NIH D43 Framework. 2 Ten Hotel, Sibasa, South Africa. 14-16 March 2017.

- **N.D Matume**, L Gray, D Tebit, P Jackson, D Rekosh, M Hammarskjöld, P Bessong. Characterization of APOBEC3G variation in Limpopo province, South Africa. (**Oral presentation**) 2nd WSU-Univen International Research Conference. The Ranch, Polokwane South Africa. 5-7 October 2016.

- **Nontokozo D. Matume**, Laurie R. Gray, Patrick E.H Jackson, Dennis M. Tebit, Pascal O. Bessong, David Rekosh, Marie-Louise Hammarskjöld. Characterization of Vif sequences and APOBEC3 genes in Limpopo Province of South Africa **(Poster presentation)** Infectious Disease & Biodefense Research Day, University of Virginia, Charlottesville, USA. 18 April 2016.

- **Nontokozo D. Matume**, Laurie R. Gray, Patrick E.H Jackson, Dennis M. Tebit, Pascal O. Bessong, David Rekosh, Marie-Louise Hammarskjöld. Analysis of potential Co-Evolution of host cell genes and HIV-1 in genetically diverse South African populations **(Poster presentation)**. Microbiology, Immunology and Cancer Biology Retreat, University of Virginia, Charlottesville, USA. 15 May 2015.

TABLE OF CONTENTS

DECLARATION ..... **i**

ACKNOWLEDGEMENTS ..... **ii**

DEDICATION ..... **iv**

ABSTRACT OF THE STUDY ..... **v**

SCIENTIFIC COMMUNICATIONS FROM THIS STUDY ..... **x**

LIST OF ABBREVIATIONS AND ACRONYMS ..... **xvii**

LIST OF FIGURES ..... **xxi**

LIST OF TABLES ..... **xxiii**

CHAPTER ONE **1**

GENERAL INTRODUCTION AND LITERATURE REVIEW **1**

1.1 INTRODUCTION **1**

1.2 LITERATURE REVIEW **2**

1.2.1 Epidemiology of HIV 2

1.3 TREATMENT OF HIV **5**

1.4 THE BIOLOGY OF HIV-1 AND ITS LIFE CYCLE **6**

1.4.1 Binding 7

1.4.2 Replication and transcription 7

1.4.3 Assembly and release 9

1.5 INTERACTION OF HIV CHEMOKINE CO-RECEPTORS WITH THE V3 LOOP OF HIV-1 GP120 ..... **12**

1.6 HOST PROTEINS INVOLVED IN HIV RESTRICTION **13**

1.7 INTERACTION OF APOBEC3 AND HIV-1 VIRAL INFECTIVITY FACTOR (*VIF*) **15**

1.8 NEXT GENERATION SEQUENCING TECHNOLOGIES **18**

1.9 STUDY RATIONALE **19**

1.10 OVERALL OBJECTIVE **22**

1.11 SPECIFIC OBJECTIVES **22**

1.12 HYPOTHESES **22**

1.13 THESIS OUTLINE **23**

1.14 REFERENCES 24

CHAPTER TWO

DEVELOPMENT OF LIBRARY PREPARATION PROTOCOLS FOR SEQUENCING HOST AND HIV-1 GENES

ABSTRACT 35
2.1 INTRODUCTION 36
2.2 Objective 37
2.2.1 Specific objectives 37
2.3 MATERIALS AND METHODS 38
2.3.1 Ethical considerations 38
2.3.2 Study population and DNA extraction 38
2.3.3 Study approach and strategy 38
2.3.3.1 Approach 1: Library preparation using the homemade Tn5 transposase 40
2.3.3.2 Approach 2: Library preparation using short custom amplicon based methods 52
2.3.3.3 Sequencing by synthesis 58
2.4 RESULTS 60
2.4.1 Quality score of basecalling 60
2.4.1 Read coverage 62
2.4.3 Read length distribution 63
2.5 DISCUSSION AND CONCLUSION 64
2.6 REFERENCES 65

CHAPTER THREE

NEXT GENERATION SEQUENCING REVEALS A HIGH FREQUENCY OF CXCR4 UTILIZING VIRUSES IN HIV-1 CHRONICALLY INFECTED DRUG EXPERIENCED SOUTH AFRICAN INDIVIDUALS

ABSTRACT 67
3.1 INTRODUCTION 69
3.2 Hypothesis 70
3.3 Objective 70
3.3.1 Specific objectives 70
3.4. MATERIALS AND METHODS 71
3.4.1 Study population and sample collection 71
3.4.2 DNA extraction and Polymerase Chain Reaction (PCR) amplification of V3 loop 71
3.4.3 Library preparation and MiniSeq sequencing 71
3.4.4 De-multiplexing, sequence quality control evaluation and mapping 71
3.4.5 Intra-patient quasispecies and co-receptor prediction 73
3.4.6 V3 loop amino acid sequence analysis 73

3.5. RESULTS **74**
3.5.1 Patients' characteristics 74
3.5.2 Frequency and distribution of predicted X4 and R5 viruses **76**
3.5.3 V3 loop amino acid sequence diversity 77
3.6 DISCUSSION AND CONCLUSIONS **81**
3.7 REFERENCES **85**

CHAPTER FOUR

GENETIC VARIATION WITHIN THE CODING REGION OF CCR5 IN CHRONICALLY HIV-1 INFECTED SOUTH AFRICAN INDIVIDUALS, THE MAJORITY OF WHOM WERE DRUG-EXPERIENCED

ABSTRACT **92**
4.1 INTRODUCTION **93**
4.2 Hypothesis **94**
4.3 Objective **94**
4.3.1 Specific objectives 94
4.4 MATERIALS AND METHODS **95**
4.4.1 Study population and sample collection 95
4.4.2 Primer design, DNA extraction and Polymerase Chain Reaction (PCR) amplification of CCR5 gene 95
4.4.3 Library preparation and MiniSeq sequencing 95
4.4.4 De-multiplexing and sequence quality control evaluation 95
4.4.5 Single nucleotide polymorphisms (SNPs) and indels detection and verification 96
4.4.6 Hardy-Weinberg Equilibrium (HWE) 97
4.4.7 Haplotypes and association mapping / Linkage disequilibrium inference 97
4.5 RESULTS **98**
4.5.1 PCR products and sequence analysis 98
4.5.2 Single Nucleotide Polymorphisms (SNPs) identification, Hardy Weinberg Equilibrium and Linkage Disequilibrium associations 98
4.6 DISCUSSION AND CONCLUSION **103**
4.7 REFERENCES **105**

CHAPTER FIVE

CHARACTERIZATION OF APOBEC3 VARIATIONS IN A SOUTH AFRICAN POPULATION

ABSTRACT **109**
5.1 INTRODUCTION **110**
5.2 Hypothesis **112**
5.3 Objective **112**
5.3.1 Specific objectives 112
5.4 MATERIALS AND METHODS **113**

5.4.1 Study population and DNA extraction....113
5.4.2 Primer design, Polymerase Chain Reaction (PCR) to amplify A3D, A3F, A3G and A3H genes....113
5.4.3 Fragmentation, tagmentation and addition of Illumina indices....113
5.4.4 Library normalization, pooling and sequencing....113
5.4.5 Demultiplexing and sequence quality control evaluation....114
5.4.6 Sequence filtering, trimming, mapping and variant calling....114
5.4.7 Statistical analysis....114
5.5 RESULTS....116
5.5.1 Single nucleotide polymorphisms (SNPs), detection of indels and verification....116
5.5.2 Allele frequencies and their comparison with other populations....127
5.6 DISCUSSION AND CONCLUSION....138
5.7 REFERENCES....141

CHAPTER SIX

GENETIC CHARACTERIZATION OF HIV-1 *VIF* AND CORRELATION WITH APOBEC3 VARIATION IN CHRONICALLY INFECTED SOUTH AFRICAN TREATMENT EXPERIENCED PATIENTS

ABSTRACT....147
6.1 INTRODUCTION....148
6.2 Hypothesis....149
6.3 Objective....149
6.3.1 Specific objectives....149
6.4 MATERIALS AND METHODS....150
6.4.1 Study population and sample collection....150
6.4.2 Primer design and Polymerase Chain Reaction (PCR) amplification of HIV-1 *vif* gene....150
6.4.4 Library preparations for HIV-1 Vif and MiSeq sequencing....150
6.5 Sequence analysis....150
6.5.1 Demultiplexing and sequence quality control evaluation....150
6.5.2 Sequence filtering, trimming and mapping....150
6.5.3 Phylogenetic analysis....151
6.5.4 Detection of selective pressure....151
6.5.5 HIV-1 vif amino acid analysis....152
6.5.6 Correlation of APOBEC3 and Vif gene variation....152
6.6 RESULTS....153
6.6.1 Phylogenetic analysis....153
6.6.2 Detection of selective pressure....155

6.6.3 HIV-1 Vif amino acid analysis....156
6.6.4 Correlation of APOBEC3 and variation Vif....159
6.7 DISCUSSION AND CONCLUSION....161
6.8 REFERENCE....163
CHAPTER SEVEN
GENERAL CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS
7.1 General conclusions....168
7.2 Study limitations....168
7.3 Recommendations....169
APPENDICES....162
Appendix I....170
Appendix II....178

LIST OF ABBREVIATIONS AND ACRONYMS

ABC	Abacarvir
AIDS	Acquired immunodeficiency syndrome
APOBEC3	Apoliprotein B messenger RNA (mRNA)-editing enzyme catalytic polypeptide like-3
ARV	Antiretroviral
CBFβ	Core-binding factor β
CCR5	Cysteine-cysteine chemokine receptor 5
CCR5Δ32	Cysteine-cysteine chemokine receptor 5 delta 32
CD4	Cluster of differentiation 4
CDC	Centre for disease control
CDS	Coding DNA sequence
cDNA	Complementary deoxyribonucleic acid
CE	Capilary electrophoresis
CRFs	Circulating recombinant forms
CXCR4	CXC receptor 4
Cul5	Cullin 5
dbSNPs	Single-nucleotide polymorphisms database
DFHC	Donald Fraser Hospital Hope Clinic
DNA	Deoxiribonucleic Acid

Env	Envelope
EloB/C	Elongin B or C
E3	Ubuquitin ligase
FDC	Fixed-dose combination
FTC	Emtricitabine
Gp120	Glycoprotein120
Gp41	Glycoprotein 41
HAART	Highly active antiretroviral therapy
HAPG	HIV/AIDS Prevention Group Wellness clinic
HEK	Human embryonic kidney
HIV	Human immunodeficiency virus
HWE	Hardy-Weinberg Equilibrium
IN	Integrase
jpHMM	jumping profile Hidden Markov Model
kDa	Kilodalton
LD	Linkage disequilibrium
LDhap	Linkage disequilibrium haplotype
LDlink	Linkage disequilibrium linkage
LTR	Long termal repeat
MAF<5%	Minor allele frequency less than 5 percent
MHC	Major Histocompatibility Complex
MT-4	Human T-cell leukaemia cells

mRNA	Messenger ribonucleic acid
MVC	Maraviroc
NARTI	Nucleoside analogue reverse transcriptase inhibitors
NCBI	National center for biotechnology information
NEB	New England Biolabs
NGS	Next Generation Sequencing
NRTI	Nucleoside reverse transcription inhibitors
NTD	N-terminal domain
NNRTI	Non-nucleoside reverse transcriptase inhibitors
NVP	Nevirapine
p24	Capsid protein
PBMC	Peripheral blood mononuclear cells
PCR	Polymerase chain reaction
PIC	Pre integration complex
PR	Protease
PSSM	Position-specific scoring matrix
QC	Quality control
RFLP	Restriction fragment length polymophism
RNA	Ribonucleic acid
RNA-Seq	Ribonucleic acid sequencing
RnaseH	Ribonuclease H
RT	Reverse transcriptase

Rs	Reference SNP cluster ID
R5	CCR5 utilizing viruses
SBS	Sequencing by Synthesis
SIVcpz	Chimpanzee simian immunodeficiency virus
SQV	Saquinavir
SNPs	Single-nucleotide polymorphisms
TAE	Tris-acetate-EDTA
TAPS-DMF	(N-Tris[hydroxymethyl]methyl-3-aminopropanesulfonic-acid) dimethylformamide
TRIM5α	Tripartite motif-containing protein 5
Tn5	Transposase
UNAIDS	Joint United Nations Programme on HIV/AIDS
USA	United States of America
UV	Ultraviolet
µl	Microlitre
µM	Micromolar
V	Voltage
Vif	Viral infectivity factor
WHO	World HealthOrganization
X4	CXCR4 utilizing viruses
Zn	Zinc

LIST OF FIGURES

Figures **Page Numbers**

Figure 1.1: Global prevalence of HIV/AIDS (Unaids 2017)……**3**

Figure 1.2: National prevalence of HIV/AIDS……**4**

Figure 1.3: The illustration of the export of mRNA transcripts out of the nucleus to the cytoplasm……**9**

Figure 1.4: The figure illustrates the main steps in the HIV-1 replication cycle……..**11**

Figure 1.5: Illustration of the HIV-1 attachment and entry into a host cell……..**13**

Figure 1.6: Mechanism of action of the APOBEC3 and the HIV-1 *vif* protein……**17**

Figure 1.7: South African map showing geographic distribution of ethnic populations ……....**20**

Figure 2.1: An overview the methodology……**39**

Figure 2.2: CCR5 gene map and primer location……...**41**

Figure 2.3: APOBEC3 gene map and primer location……...**43**

Figure 2.4: Schematic representation of library preparation using Nextera XT kit/custom Tn5 tagmentation protocol, with amplicon1 from A3D, A3F and A3F……..**49**

Figure 2.5: Bionalalyzer 2100 expert with a High Sensitive DNA assay kit (Agilent Genomics) profile verifying the size and concentration of libraries……...**51**

Figure 2.6: HIV-1 genome map and location of HIV-1 Vif (A) and V3 loop (B). ……**52**

Figure 2.7: Schematic representation of the overall Illumina SBS sequencing workflow……**59**

Figure 2.8: Quality control check of a sample of libraries prepared using Tn5 transposase (A), custom based (B) and Nextera (C)……**61**
Figure 2.9: Reads coverage from a sample of libraries prepared from Tn5 transposade (A), custom based (B) and Nextera (C)……**62**

Figure 2.10: Read length distribution from a sample of libraries prepared from Tn5 transposase (A), custom based (B) and Nextera (C)……**63**

Figure 3.1: Workflow showing the bioinformatics processes for viral subtype and co-receptor determination……**72**

Figure 3.2 Variation within the V3 region of Subtype C R5 and X4 viruses……...**79**

Figure 3.3: Comparison of (A) amino acid length and (B) amino acid net charge of HIV-1 subtype C isolates capable of using CCR5 and CXCR4……...**80**

Figure 4.1: Graphic illustration of transmebrane organization of human CCR5 showing the location of the mutation highlited in red identified in the study……**102**

Figure 6.1: Phylogeny of 86 South African Vif consensus sequences……**155**

Figure 6.2: Cumulative dN/dS graph with indels and stop codon of 86 HIV-1 Vif consensus sequences……**156**

Figure 6.3: South African HIV-1 Vif protein sequence entropy……..**157**

Figure 6.4: South African Vif protein sequence alignment summary……**158**

Figure 6.5: Correlation between the global distribution of HIV-1 Vif alleles and human A3H haplotypes……**160**

LIST OF TABLES

Tables **Page number**

Table 1.1: Next generation sequencing platforms and their descriptions……..……..**19**

Table 2.1: List of APOBEC3 primers designed; primer name, sequence and product size are indicated……**45**

Table 2.2 Thermal cycling conditions……..**48**

Table 2.3: List of HIV-1 Vif and V3 loop primers designed; primer name, sequence and product size are indicated……**54**

Table 2.4: PCR amplification protocol for HIV-1 Vif AND V3 loop……....**56**

Table 3.1: Characteristics of individuals harboring R5 or X4 monotropic and those that harbor dual/mixed R5+X4 and R5X4+ viruses……**75**

Table 3.2: Subtype confirmation with Geno2pheno co-receptor determination tool, the jumping profile Hidden Markov Model (jpHMM) and REGA subtyping tools……..**76**

Table 3.3: Frequency of predicted R5, X4, R5+X4 and R5X4+ viruses in the study subjects…**77**

Table 4.1: Human CCR5 nonsynonymous and synonymous changes, genotype frequencies and Hardy Weinberg Equilibrium calculations from the study population……..**99**

Table 4.2: Comparison of CCR5 allele frequencies between a South African population and total populations from East Asian (EAS), European (EUR), African (AFR), Ad Mixed American (AMR) and South Asian (SAS)……**100**

Table 4.3: Comparison of CCR5 allele frequencies between a South African population and other……..**101**

Table 5.1: APOBEC 3D, 3F, 3G and 3H nonsynonymous and synonymous changes, genotype frequencies and Hardy Weinberg Equilibrium calculations from the study population……**118**

Table 5.2: Haplotype formation and the frequencies with respect to the A3D, A3F, A3G and A3H……..**124**

Table 5.3: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population (this study) and total populations from East Asian (EAS), European (EUR), African (AFR), Ad Mixed American (AMR) and South Asian (SAS). ……**128**

Table 5.4: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population and other Africa population…….**133**

CHAPTER ONE

GENERAL INTRODUCTION AND LITERATURE REVIEW

1.1 INTRODUCTION

Human immunodeficiency virus (HIV) is a Lentivirus, belonging to the Retroviridae family, and causes Acquired Immunodeficiency Syndrome (AIDS), a condition in humans that results in the progressive failure of the immune system, allowing life-threatening opportunistic infections and cancers to develop. HIV can be transmitted through sexual contacts, contaminated blood products and from mother to child.

It was discovered in 1983 after the first cases of AIDS were recognized between 1980 and early 1981 (Centers for Disease Control 1981). Several initial patients showed symptoms of *Pneumocystis carinii* pneumonia (Gottlieb *et al.* 1981), a rare opportunistic infection that was known to be present in people with severely compromised immune systems. Soon thereafter, additional patients developed a previously rare skin cancer called Kaposi's sarcoma (Friedman-Kien 1981). At the end of 2016, there was an estimate of 36.7 million people living with HIV worldwide (UNAIDS 2017). The two types of HIV (HIV1 and HIV-2) are now known to have originated from non-human primates in sub-Saharan Africa where in the late 19th or early 20th century they were transmitted to humans (Korber *et al.* 2000). HIV-1 appears to have originated in southern Cameroon through the evolution of SIV (cpz) (Gao *et al.* 1999), a Simian Immunodeficiency Virus (SIV) that infects wild chimpanzees. The closest relative of HIV-2 is SIV from Sooty Mangabey monkeys (Keele *et al.* 2006). Distinct from many viruses, HIV exhibits very high genetic variability.

There is currently no cure for HIV infection. The main treatment for HIV is Antiretroviral therapy (ART), which aims at keeping the amount of HIV in the body at undetectable levels. Primarily, treatment is with combination ARV therapy (cART) referred to as Highly Active Antiretroviral Therapy (HAART). It is composed of a combination of several classes of antiviral drugs such as the nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), and protease inhibitors. These treatments have been highly beneficial to many patients since their introduction in 1996, when the protease/RT inhibitor-based HAART first became available (Palella *et al.* 1998). One major problem in HIV treatment is that the virus has a prolific and highly error-prone replication process that can lead to the development of drug resistance. In addition, some severe side effects are associated with treatment. The shortcomings of these drugs thus

continue to open new vistas in the development of alternative therapeutic approaches to treating HIV-1- infected individuals.

Potential major breakthroughs for future therapeutic development were the discoveries of the role of co-receptors in HIV-1 in 1996 and anti-viral activities of the host restriction factors such as APOBEC3G protein, a member of the DNA cytosine deaminase family in 2008. Since the discovery of these genes and their role in HIV replication, many studies have been conducted that show their association with virus polymorphism.

Sequencing of DNA and its applications has played a significant role in our understanding of biology and contributed in many fields such as drug discovery and development. The latest sequencing technology referred to as Next Generation Sequencing (NGS), due to its advantages such as low cost and large-scale high-throughput capacity, it has be applied in studying many microbes and made the sequencing of whole genome easier. The Human Genome Project and the re-sequencing of selected regions of the human genome in disease association studies have contributed to a refined understanding of the molecular basis of many diseases. Although there still need for Human Exome sequencing of genes such as the APOBEC in limited recourses settings.

1.2 LITERATURE REVIEW

1.2.1 Epidemiology of HIV

By the end of 2016, an estimated 36.7 million people were living with HIV throughout the world. From the total of people living with HIV worldwide, 34.5 million were adults and 2.1 were children under the age of 15 years. In 2016, an estimated 1.8 million people were newly infected with HIV (figure 1.1), a decline from 2.1 million new infections in 2015. Adult new infection declined by 11% and 16% for the general population between 2010 and 2016, whereas there was only an 8% decline between 2010 and 2015. New infections amongst children globally have halved from 300,000 in 2010 to 160,000 in 2016, with 47% prevalence. In 2016, 1 million people died of AIDS- related illnesses. (UNAIDS, 2017).

The vast majority of people living with HIV are located in low and middle-income countries, and sub-Saharan Africa remains the region most heavily affected. In 2016, about 69.4% of all people living with HIV resided in sub-Saharan Africa, a region with only 12% of the global population. In 2016, adults and children living with HIV in this region were estimated to

comprise 19.5 million, whereas those newly infected with HIV were 790,000. An estimated 420,000 adults and children died of AIDS in 2016 (UNAIDS 2017).

In North Africa and the Middle East, approximately 230,000 people were living with HIV, and an estimated 18,000 people became newly infected in 2016, while an estimated 11,000 adults and children died of AIDS. As for Asia and the Pacific, 5.1 million people were living with HIV, with 270,000 newly infected in 2016, and the disease had claimed 170,000 lives. There were 2.1 million people in Latin America and Caribbean living with HIV in 2016, with 190,000 new infections and 45,400 lives lost (UNAIDS 2017).

In Eastern Europe and Central Asia, an estimated 1.6 million people are living with HIV and 190,000 were newly infected in 2016. HIV/AIDS had claimed 40,000 lives by 2016 in the region. As for Central Europe, Western Europe and North America, there were 73,000 new cases of HIV in 2016, bringing the number of people living with HIV/AIDS to 2.1 million. An estimated 18,000 people in this region died of AIDS in 2016 (UNAIDS 2017).

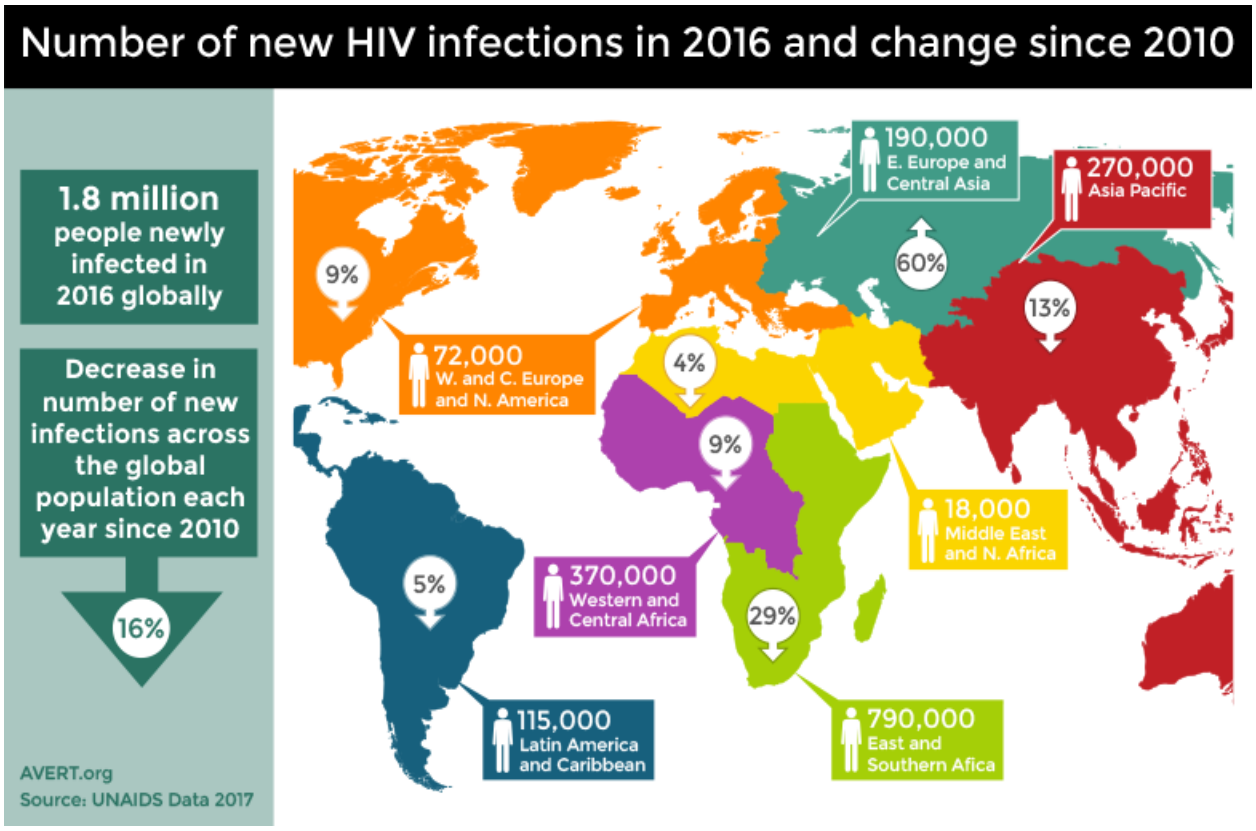


Figure 1.1: Global prevalence of HIV/AIDS (UNAIDS, 2017).

The epidemic continues to be severe in South Africa, where there are more people living with HIV than in any other country in the world, with an estimated 7.1 million people living with HIV (figure 1.2). In 2016, there were 270,000 new HIV infections and 110,000 AIDS-related deaths. This country also has the largest antiretroviral programs, with 56% adults and 55% children on antiretroviral treatments (UNAIDS 2017). There was a drop in the 2011 HIV prevalence in KwaZulu-Natal with an estimate of 37.4%, indicating a decline by 2.1% from 39.5%. In contrast, the Mpumalanga Province showed an increase in estimated HIV prevalence of 2.0%, from 34.7% in 2009 to 36.7% in 2011. There was also an increase in HIV prevalence in Free State from 30.6% in 2010 to 32.5% in 2011, and the North West from 29.6% in 2010 to 30.2% in 2011. Limpopo showed an increase from 21.4% in 2009 to 22.1% in 2011 (South African National Department of Helath 2010) (National Antenatal Sentinel, 2012).

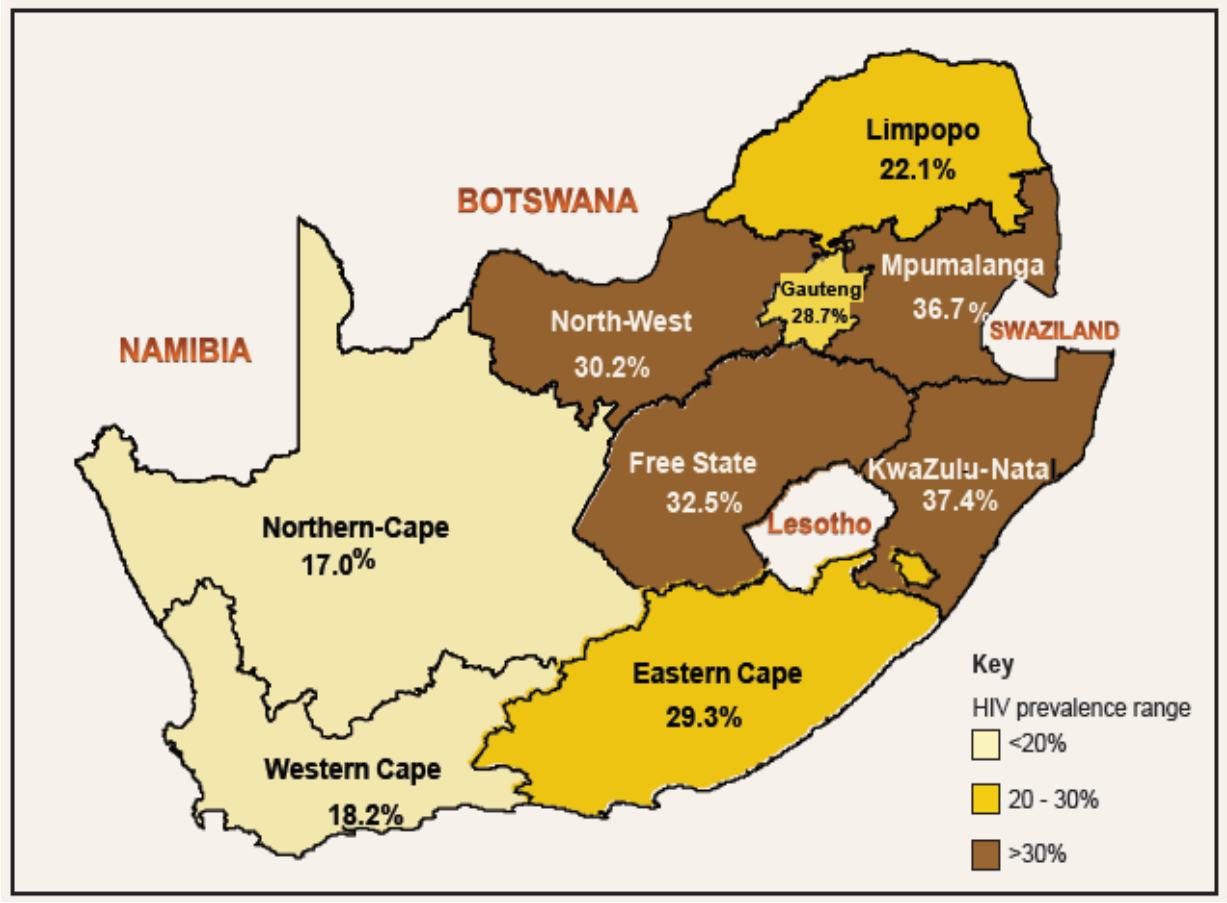


Figure 1.2: National prevalence of HIV/AIDS (South African National Department of Helath 2010).

1.3 TREATMENT OF HIV

There is currently no cure for HIV/AIDS and the main treatment for HIV is antiretroviral treatment, which aims to provide maximal and durable suppression of viral copies in the body, restore immune function, reduce HIV-related infections and improve quality of life. The goals of HIV treatment can be achieved when adherence is highly practiced.

There are currently five classes of ART drugs approved, which target different stages of the virus's life cycle. Treatment methods primarily target the reverse transcription step with reverse transcriptase inhibitors, the protease cleavage step with protease inhibitors, the viral DNA integration step with integrase inhibitors, viral attachment with attachment inhibitors, and the co-receptor binding step with co-receptor inhibitors or the fusion step with fusion inhibitors.

Current cART options are a combination consisting of two or more drugs belonging to at least two types or classes of antiretroviral agents. The most common drug combination in treatment initiation consists of two NRTIs combined with either an NNRTI or a boosted protease inhibitor. Ritonavir in small doses is the most commonly used booster. An example of the common antiretroviral combination is the two NRTIs zidovudine (AZT) and lamivudine (3TC), combined with the NNRTI efavirenz (EFV). The second line consists of two of either NRTIs or NNRTIs and boosted PI such as Abacavir (ABC), Lamavudine (TDF) and Lopinavir (LPV/r). Third line also referred t as salvage therapy is given when the first and second line has failed and drug resistance testing conducted. Integrase such as Raltegravir (RAL) or entry inhibitors such as Maraviroc (MVC) are recommended.

The decline in new infections and AIDS deaths can be attributed largely to the scale-up of antiretroviral therapy programs, especially in developing countries. There has been a dramatic increase in access to life-saving ART in the past several years, particularly in resource limited countries and substantial additional efforts were invested to achieve the goal of 15 million people on ART in 2015 (UNAIDS 2013)A major milestone was achieved in 2016 where half (53%) of the people living with HIV had access to ART. By the end of 2016, 19.5 million people living with HIV were receiving ART from 17 million in June 2016 and 7.5 million in 2010. If the scale up of treatment continues this way, it is estimated the world will reach the global target of 30 million people on treatment by 2020 (Unaids 2017). Significant progress has also been made in the prevention of mother to child

transmission of HIV. In 2016, 76% of all pregnant women living with HIV had access to treatment.

In sub-Saharan Africa, about 61% of adults and 51% of children needing treatment were on treatment by the end of 2016 (Unaids 2017). Of the sub-Saharan countries, South Africa is a country with the largest antiretroviral treatment programme globally. In 2016, more than 3.7 million people were receiving ART, which equates to 56% people living with HIV in the country. In 2012, it was only 31.2% of people living with HIV were on ART (SANAC 2011; Unaids 2017).

In spite of this remarkable progress, drug resistance to two or three antiretroviral drug classes is a major problem in the long-term management of HIV-1 infected patients. When patients have few or no available standard treatment options, they have to seek salvage therapy. One of the drugs currently used for salvage therapy is Maraviroc (MVC), the first licensed antiretroviral drug from the class of co-receptor antagonists, which antagonize C-C Chemokine co-receptor type 5 (CCR5).

1.4 THE BIOLOGY OF HIV-1 AND ITS LIFE CYCLE

Human immunodeficiency virus type 1 (HIV-1) is a complex retrovirus. The genome is about 9kb and it contains nine genes and encodes 15 proteins. The genes are classified into structural, catalytic, regulatory and accessory genes. Outside a human cell, a mature HIV virion exists as a roughly spherical particle with a diameter between 100-150 nanometers. The central core consists of two copies of positive sense single-stranded RNA. This RNA is bound to nucleocapsid proteins (p9 and p7) and is enclosed by a conical capsid composed of 1300 copies of the viral capsid protein p24 and a matrix protein p17. Enzymes needed for the virion maturation and replication is also packaged in the virion (protease (PR), reverse transcriptase (RT), and integrase (IN) (Chan *et al.* 1997).

The HIV capsid is surrounded by a protein matrixanda bilayer lipid membrane, derived from the host cell through “viral” budding from regions that contain the viral Envelope proteins. The envelope proteins consist of a surface glycoprotein (*gp*120) and a transmembrane (fusion) protein (*gp*41). The viral envelope proteins are essential for attachment and fusion with a target susceptible cell (Chan *et al.* 1997).

The nine viral genes of HIV are *gag*, *pol*, *env*, *tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu*. The *gag* and *env* genes encode the structural proteins. The six genes *tat*, *rev*, *nef*, *vif*, *vpr*, and *vpu* are regulatory and accessory genes for proteins that control the ability of HIV to replicate and cause disease (Chan *et al.* 1997). The *pol* gene encodes the essential viral enzymes (PR, RT and IN). All the genes have specific functions that control the ability of the virus to infect cells and produce new virus.

Once HIV is in the body, it targets and infects a variety of immune cells such as $CD4^+$ T-cells, monocytes/macrophages, dendritic cells and microglial cells. The virus takes over these cells and turns them into factories that produce thousands of copies of the virus. HIV goes through several steps, which are also drug targets, to establish itself in the human body, shown in figure 1.4. The steps are as follows:

1.4.1 Binding

Human immunodeficiency virus infects $CD4^+$ cells by binding of the *env* glycoprotein (gp120) on its surface, initially to CD4. This leads to a conformational change that allows binding to chemokine receptors (CCR5 or CXCR4) on the target cell. Attachment to the cell is followed by fusion of the viral envelope with the cell membrane (through gp41) and release of the viral capsid into the cell (Arthos *et al.* 2008). At this stage, entry and fusion inhibitors can block the attachment and entry of HIV in to the cell.

1.4.2 Replication and transcription

After the viral capsid enters the cell into the cytoplasm in the absence of monkey TRIM5α, reverse transcriptase (RT), which is an RNA-dependent DNA polymerase, transcribes single-stranded RNA into single-stranded DNA. It also functions as DNA-dependent DNA polymerase that synthesizes a second strand of DNA complementary to the reverse transcribed single-stranded cDNA after degrading the original mRNA with its RNaseH activity. The end result of reverse transcription is a double-stranded viral DNA containing a Long Terminal Repeat (LTR) at each end. A successful reverse transcription and transport of viral genome to the nucleus may be prevented by the presence of TRIM5α restriction factor in the cytoplasm during the release and uncoating of the viral genome (Stremlau *et al*. 2004). Additionally, the NRTIs, NNRTIs as well as host restriction factors such as APOBEC3 genes and SAMHD1 are capable of restricting the replication HIV-1 during reverse

transcription. In contrary HIV-1 has also evolved mechanism to block the action of the host genes through its encoded *vif* and *vpx* genes respectively allowing the resulting DNA to move into the nucleus as a component of the pre-integration complex (PIC) that then integrates into the host cell genome (Sherman *et al.* 2002). Integration of DNA is essential for a productive infection and Integrase inhibitors can block the integration of viral genome into host chromosome.

Once the provirus is integrated into the host chromosome, the virus is expressed using the host cell transcription, RNA processing and translation machinery. The integrated DNA provirus is transcribed into full-length RNA that interacts with the cellular RNA processing machinery like any cellular pre-mRNA. Some of the RNA is spliced (either completely or incompletely) before export to the cytoplasm and translation by the host protein synthesis machinery (Figure 1.3). A fraction of the RNA gets exported to the cytoplasm in an unspliced form that serves as the mRNA template for viral protein production (Gag and GagPol proteins) and is also packaged into new virus particles (figure 1.4).

The completely spliced ~2kb mRNA is exported out of the nucleus to the cytoplasm for the translation of *nef*, *tat* and *rev*. The *tat* and *rev* factors are absolutely required for efficient viral gene expression and function at the transcriptional and post-transcriptional levels in infected cells. As the newly produced *rev* protein accumulates in the nucleus, it binds to an element present in the ~4kb incompletely spliced and the ~9kb unspliced mRNAs known as the Rev Response Element (RRE). This allows export of these mRNAs from the nucleus to the cytoplasm, where they will be translated. The partially spliced mRNA encodes *env*, *vif*, *vpu* and *vpr* (figure 1.3).

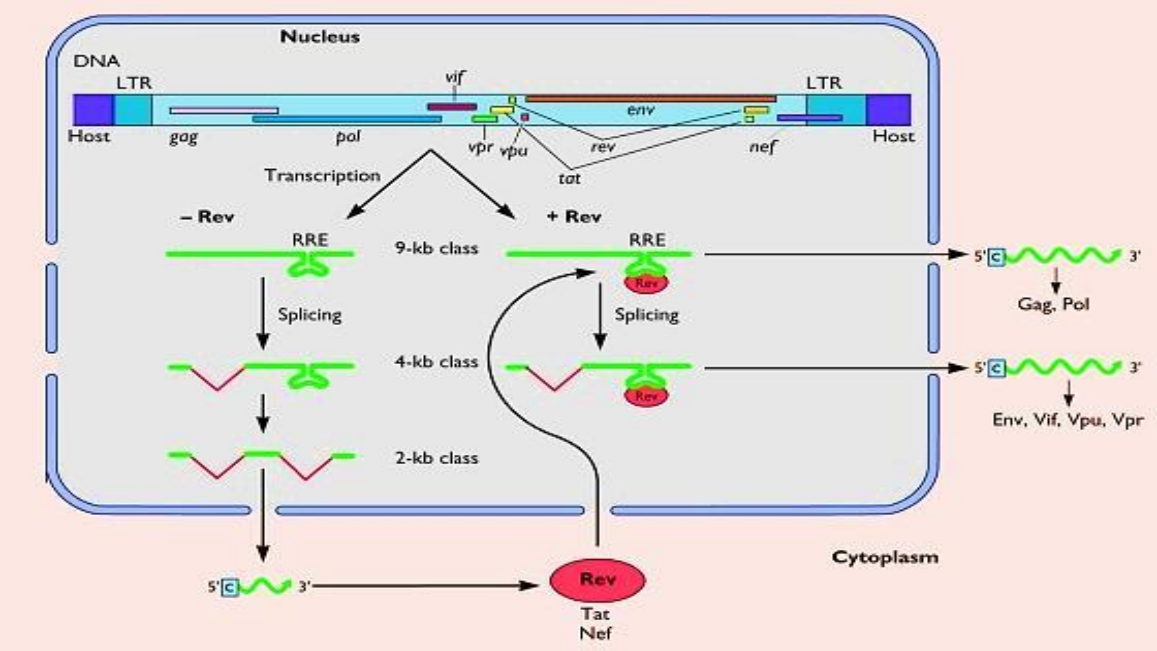


Figure 1.3: The illustration of the export of mRNA transcripts out of the nucleus to the cytoplasm (Cullen 2003).

1.4.3 Assembly and release

The final step of the viral cycle is the assembly of the new HIV-1 virions, which takes place at the plasma membrane of the host cell. The Env polyprotein Gp 160 traffics through the endoplasmic reticulum and is transported to the Golgi complex, where it is cleaved by a host protease called furin and processed into two HIV envelope glycoproteins (Gp 41 and Gp 120), which are then inserted into the plasma membrane of the host cell. The Gp 41 protein is a transmembrane protein that anchors the Gp 120 surface glycoprotein to the membrane of the infected cell. The assembly of newly synthesized HIV-1 virions is mediated by the HIV-1 Gag polyprotein, which directs the recruitment of GagPol and other components required for the assembly and release of new viral particles (Freed 1998). After trafficking to the membrane, budding takes place in regions that also contain gp120 and gp41. Maturation occurs either in the forming bud or in the immature virion after it buds from the host cell. During maturation, the HIV protease cleaves the Gag and GagPol polyproteins into individual HIV structural proteins and enzymes (Douglas *et al.* 1997). The mature virus is then able to infect another cell. Protease inhibitors can block this stage and Tetherin a host restriction factor is capable of restricting the release of HIV-1 particles in the absence of HIV-1 Vpu viral protein (figure 1.4).

Upon transmission to a new host, HIV infects $CD4^+$ cells of the human immune system, such as CD4 helper T-cells, $CCR5^+CD4^+$ effector memory T-cells, macrophages and dendritic cells. Depletion of such cells during HIV infection leads to low levels of $CD4^+$ T-cells through mechanisms such as direct viral killing of infected cells, increased rates of apoptosis in uninfected cells and killing of infected $CD4^+$ cells by $CD8^+$ cytotoxic lymphocytes that recognize infected cells. When the $CD4^+$ T-cell number declines below a critical level, cell-mediated immunity is severely compromised. In the absence of antiviral drug treatment, the body becomes progressively more susceptible to opportunistic infections, ultimately resulting in AIDS (Grossman *et al.* 2006).

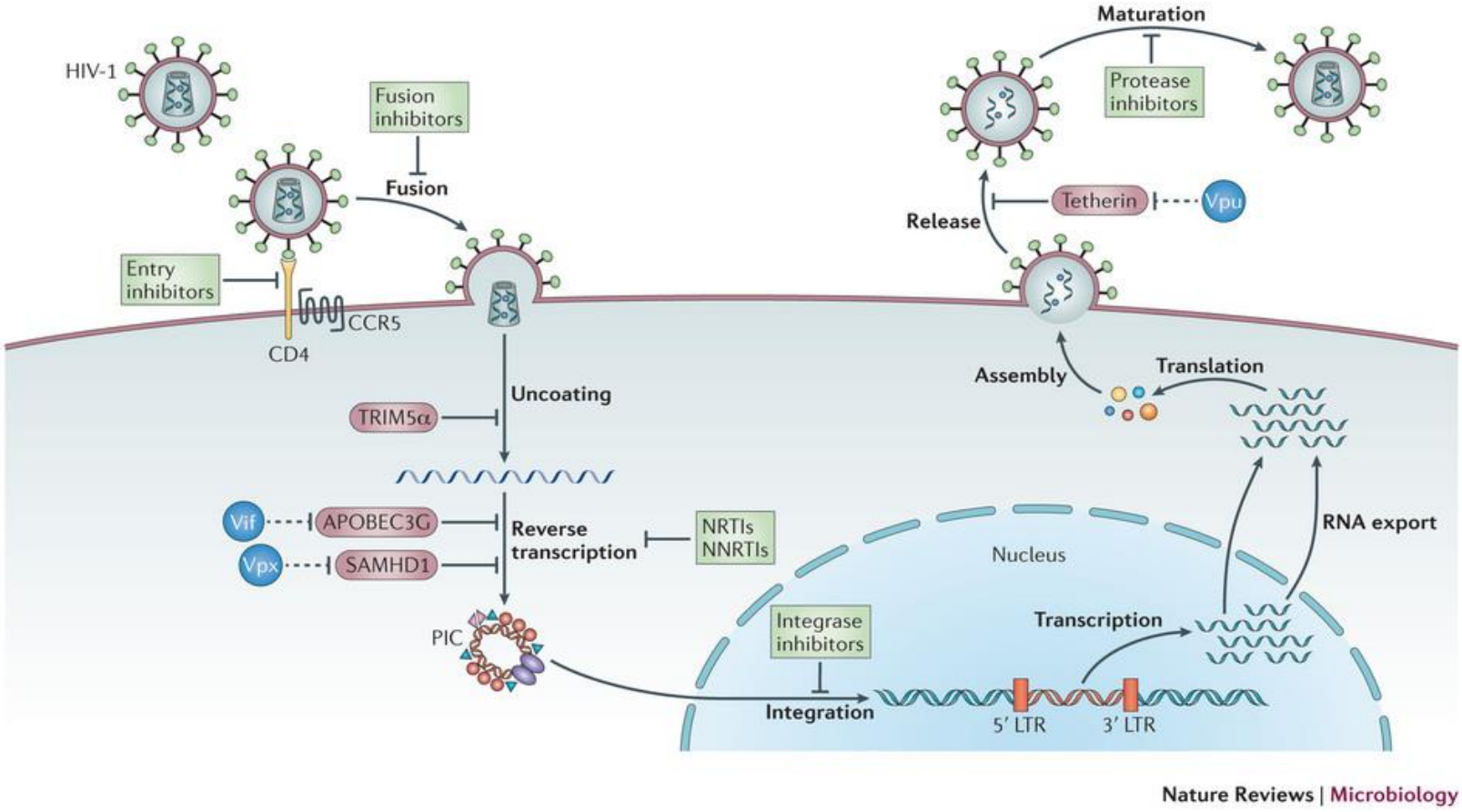


Figure 1.4: The figure illustrates the main steps in the HIV-1 replication cycle. The major families of antiretroviral drugs (green), and the step of the life cycle that they block, are indicated. Also shown are the key HIV restriction factors and their corresponding viral antagonists (*vif*, *vpx* and *vpu* in blue); CCR5, CC-chemokine receptor 5; LTR, long terminal repeat; NRTIs, nucleoside reverse transcriptase inhibitors; NNRTIs, non-nucleoside reverse transcriptase inhibitors (Barré-Sinoussi *et al.* 2013).

1.5 INTERACTION OF HIV CHEMOKINE CO-RECEPTORS WITH THE V3 LOOP OF HIV-1 GP120

Entry of HIV into a host cell does not only require CD4, but also a chemokine co-receptor. The virus can use CCR5, CXCR4 or both co-receptors to gain entry into the host cell. Viruses that use CCR5 as co-receptor are referred to as R5, those that use CXCR4 as X4 and those use both co-receptors as dual R5X4 viruses.

Characterization of HIV-1 co-receptor usage is clinically significant and of high importance because of future therapeutic use of HIV-1 co-receptors as targets. Minimal changes in the V3 region amino acid are sufficient to switch co-receptor from CCR5 to CXCR4 (Pastore *et al.* 2006). The majority of initial HIV-1 infections involve the use of CCR5, while CXCR4 use often predominates at late stages of infection (Connor *et al.* 1997; Scarlatti *et al.* 1997; Berger *et al.* 1998). This has been strongly associated with low $CD4^+$ cell counts and occurs in approximately 50% of subtype B infections (Tersmette *et al.* 1989; Richman and Bozzette 1994; Connor *et al.* 1997; Hatse *et al.* 2003). At the late stages of infection, a co-receptor switch from R5 to X4 has also been reported in subtypes A, C, D, CRF01_AE, and CRF02_AG infections (Björndal *et al.* 1997; van Rij *et al.* 2002; Esbjörnsson *et al.* 2010).

The third variable domain (V3) of the HIV-1 envelope typically spans 33 to 35 residues that are bounded by two invariant cysteines that form a disulfide bond to create a loop (Poon *et al.* 2007). The V3 loop appears to be the major genotypic determinant of HIV-1 co-receptor usage (Sander *et al.* 2007). Interaction of the CD4 and gp120 exposes the V3 loop that predominantly mediates the molecular recognition of the chemokine receptor (Cormier and Dragic 2002; Suphaphiphat *et al.* 2003). Upon V3 loop/co-receptor binding, a series of rearrangements in the envelope occurs that leads to fusion of the host virus cell membrane (figure 1.5). Entry inhibitors such as Maraviroc bind to CCR5 and inhibit the gp120-CCR5 interaction through an allosteric mechanism (Garcia-Perez *et al.* 2011). The United States Food and Drug Administration first approved Maraviroc as salvage therapy in 2007. Currently, it is also recommended as a component in first-line antiretroviral therapy (ART) in the United States and other developed countries (Woollard and Kanmogne 2015).

Shortly after the role of CCR5 as co-receptor was discovered, many studies were conducted that led to the discovery of the CCR5Δ32 mutant and its association with protection to HIV-1 infection in individuals homozygous for this allele (Picton *et al.* 2010). This discovery provided the first genetic evidence of natural protection to HIV-1 infection and prompted further studies investigating the effects of polymorphism in CCR5 and how it influences the outcome of HIV-1 exposure and infection. To date, several mutations and single-nucleotide polymorphisms (SNPs) in this gene have been discovered and found to be important genetic factors capable of influencing susceptibility to HIV-1 infection or affecting the rate of disease progression (Picton *et al.* 2010).

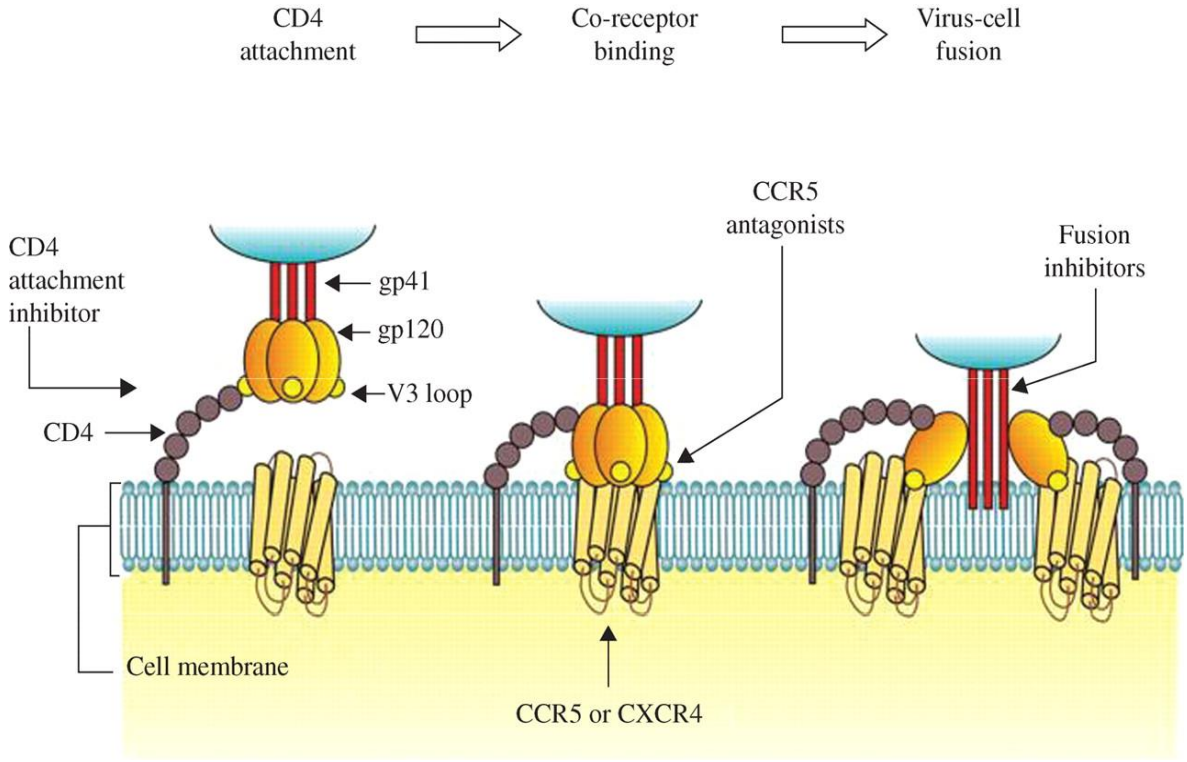


Figure 1.5: Illustration of the HIV-1 attachment and entry into a host cell (Vandekerckhove *et al.* 2009).

1.6 HOST PROTEINS INVOLVED IN HIV RESTRICTION

Retroviruses such as HIV-1 requires the host cell and its machinery to achieve efficient replication (Bushman *et al.* 2009). Mammalian cells on the other hand also express a number of diverse, dominantly acting proteins that are widely expressed and function in a

cell autonomous manner to suppress viral replication. These proteins are known as restriction factors or intrinsic resistance factors, which are a part of the innate immune responses, providing first line of defense against infections (Bieniasz 2003).

Host restriction factors play an important role in viral replication, and are defined by the following characteristics: First, these proteins must be able to directly and dominantly cause a significant decrease in viral infectivity (in the absence of counteracting viral proteins). Second, if a restriction gene is a potential threat to viral replication, then viruses should show signs of having evolved a potent counter-restriction mechanism. Third, restriction genes also often show signatures of rapid evolution because of longstanding interactions with infecting viruses. Fourth, restriction genes should be part of an innate immune response (Sheehy *et al.* 2002; Harris 2008; Malim and Emerman 2008; Malim and Bieniasz 2012).

A number of different host restriction proteins important in HIV infection have been discovered using different approaches, leading to the discovery of numerous AIDS restriction genes that affect susceptibility of viral infection (Kaur and Mehra 2009). Genome-wide knockdown and proteomic studies have suggested that up to 10% of human proteins either directly or indirectly impact HIV replication (Brass *et al.* 2008; König *et al.* 2008; Zhou *et al.* 2008; Yeung *et al.* 2009; Jäger *et al.* 2012). The known and widely studied restriction factors include the apoliprotein B messenger RNA (mRNA)-editing enzyme catalytic polypeptide like-3 (APOBEC3) family of proteins, tetherin/bone marrow stromal cell antigen 2 (BST2/CD317 (thereafter called tetherin), and tripartite-motif-containg 5α (TRIM5α) (Stremlau *et al.* 2004; Nisole *et al.* 2005; Neil *et al.* 2008; Van Damme *et al.* 2008; Laguette *et al.* 2011).

The APOBEC gene family was the first restriction factors to be studied and comprise Cytidine deaminases capable of editing DNA and or RNA sequences. Although the APOBEC genes belong to larger superfamily deaminases, APOBECs are restricted to vertebrates (Conticello *et al.* 2003). The family of APOBEC comprises eleven members in humans with distinct functions and locations. The activation-induced deaminases (AID) and APOBEC 1 genes are located on chromosome 12, the APOBEC 2 gene is located in chromosome 6, the seven of the APOBEC3 genes are located on chromosome 22 and the APOBEC4 gene is located in chromosome 1 (Espinosa *et al.* 1994; Liao *et al.* 1999; Muto *et*

al. 2000; Jarmuz *et al.* 2002). Members of this family are distinguished by the presence of one or two catalytic domains containing a Zinc-binding deaminase motif, characterized by the conserved amino acid sequence H-X-E-X $_{(23-28)}$-P-C-X $_{(2-4)}$-C in which the His and Cys residues coordinate Zn^{2+} and the Glu residue is involved in the proton shuttle during the deamination reaction (Jarmuz *et al.* 2002; Wedekind *et al.* 2003).

One of the most studied HIV restriction factors of the APOBEC3 family is APOBEG3G. In the absence of the viral *vif gene*, this protein was shown to convert a cell line with a permissive phenotype into a non-permissive phenotype (Sheehy *et al.* 2002; van Rij *et al.* 2002). In the absence of *vif*, APOBEC3G is efficiently packaged into viral particles, causing restriction during reverse transcription. Soon after the initial discovery and functional characterization of APOBEC3G, attention turned towards its six highly related family members. It was later reported that only four (APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H) package into virus-like particles and inhibit viral replication when stably expressed in human T-cell lines (Hultquist *et al.* 2011). Endogenous APOBEC3D and APOBEC3F combine to inflict the 5'-GA-to-AA mutation pattern observed in the non-permissive T-cell line CEM2n (Refsland *et al.* 2012). Seven different human haplotypes of APOBEC3H proteins have been reported. Cell-based studies indicate that only subsets of these haplotypes are stable at the protein level and capable of HIV restriction (Dang *et al.* 2008; OhAinle *et al.* 2008; Harari *et al.* 2009; Wang *et al.* 2011).

1.7 INTERACTION OF APOBEC3 AND HIV-1 VIRAL INFECTIVITY FACTOR (*VIF*)

Several human APOBEC3 proteins can enter assembling viral particles, during viral particle production (figure 1.6). When these particles then infect a target cell, the APOBEC3 protein deaminates viral cDNA, changing cytosine to uracil during reverse transcription. This causes G-to-A hypermutations and genome degradation, which results in viral inhibition (Zhang *et al.* 2003). HIV-1 has evolved a mechanism to counteract this restriction through its encoded Vif protein.

The Vif protein is a 23kd protein essential for replication of HIV-1 in cells that express APOBEC3, which includes monocytes, macrophages, primary CD4+ T lymphocytes, and some leukemic T-cell lines (Fisher *et al.* 1987; Gabuzda *et al.* 1992; Madani and Kabat 2000; Sheehy *et al.* 2002). Studies have shown that during counter restriction, Vif binds to APOBEC3 and forms an E3 ligase complex by also binding to cellular proteins including

Cullin 5, Elongin B and C. This complex induces polyubiquitination and directs the degradation of APOBEC3 by the 26S proteasesome (Sheehy *et al.* 2002; Conticello *et al.* 2003; Yu *et al.* 2003; Mehle *et al.* 2004)

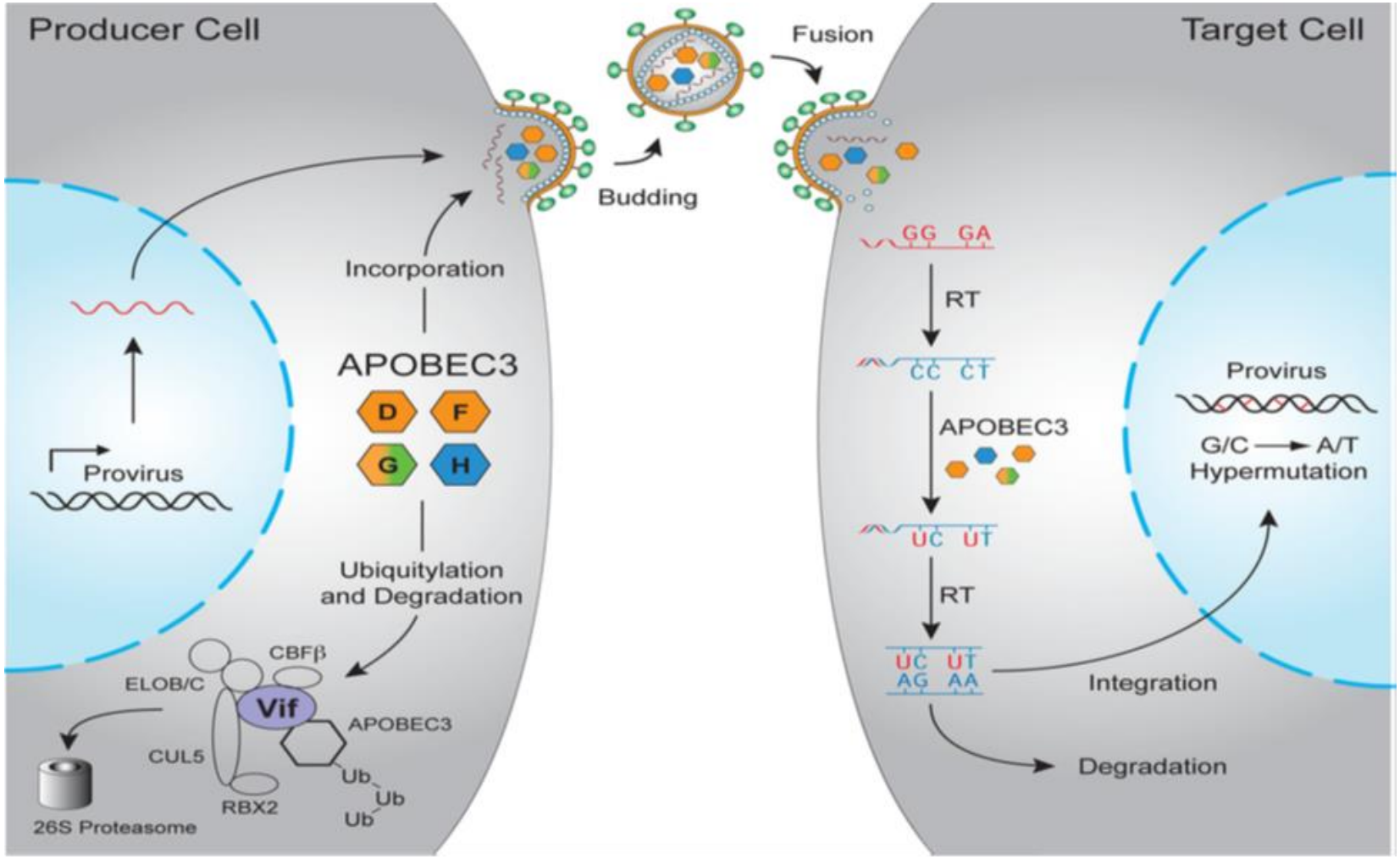


Figure 1.6: Mechanism of action of the APOBEC3 and the HIV-1 Vif protein (Hultquist *et al.* 2011).

1.8 NEXT GENERATION SEQUENCING TECHNOLOGIES

DNA sequencing, which is a technique used to determine the order of nucleotides sequence is DNA, has been around since the mid 1970s (Griffin and Griffin 1993). The first sequencing technologies were discovered and named by Allan M. Maxan and Walter Gilbert and the Sanger method by Frederik Sanger between 1977-1981 (Sanger *et al.* 1977; Men *et al.* 2008). These methods relied on different principle and technique of sequencing DNA. The Sanger sequencing method became popular and formed basis of many studies including the Human Genome Project. Its principle involves DNA synthesis using deoxy and fluorescence labeled dideoxyribonucleic acid that results in termination of synthesis at specific nucleotides. This method was improved and later performed using automated sequencing machines such as the ABI 370 in 1985, ABI 377 in 1995 and the ABI 3700 in 1999 by the Roche Company.

Sanger based sequencing has build a solid foundation of genomic sequencing that the next generation of sequencing technologies had to build upon through re-sequencing and comparative genomic studies. In addition to the Whole genome sequencing, this method has been used to sequence countless amplicons for applications such as verification of clones, searching for SNPs and forensic analysis (Goldberg *et al.* 2006; Fakruddin and Chowdhury 2012; Baudhuin *et al.* 2015; Gao *et al.* 2016; Chang *et al.* 2017). Over the past 10 years, they had been significant improvements in sequencing technology such as the introduction of Next Generation Sequencing. This technology made it possible for millions to billion of DNA molecules to be sequenced simultaneously reducing cost of sequencing from ~$10/Kb to ~$1/Kb (Blow 2008; Mardis 2008; Shendure and Ji 2008). Several application of NGS allowed sequencing of multiple strains of pathogenic bacteria/viruses to monitor drug resistance and pathogenesis, re-sequencing of selected regions to search for human variation in population, monitor the onset of drug resistance in HIV or HCV, and profiling tumors to guide cancer therapies (Nielsen *et al.* 2011; Paredes 2012; Capobianchi *et al.* 2013; Johnson *et al.* 2014; Jour *et al.* 2014; G. and W. 2016).

The first NGS instrument was a 454GS Junior by Roche Applied Science in 2005 followed by a 545GS FLXT which relied on pyro-sequencing chemistry and phased out in 2016. Improvements on data output in these technologies led to the introduction of many instruments such as the Selexa 1G Analyzer HiSeq, MiSeq and MiniSeq by Illumina Inc, SOLiD and Ion Proton by Applied Biosystems and the PacBio Rs. Next generation

sequencing holds the promise of not making current experiments more feasible, allowing flexibility in experiment design afforded by quicker and more cost efficient sequencing. These new technologies as well as the current state-of the-art Sanger sequencing platform are summarized in table 1.1

Table 1.1: Next generation sequencing platforms and their description.

Platform	Chemistry	Read Length	Run Time	Gb/Run	Advantage	Disadvantage
454 GS Junior (Roche)	Pyro-sequencing	500	8 hrs.	0.04	Long Read Length	High error rate in homopolymer
454 GS FLX+ (Roche)	Pyro-sequencing	700	23 hrs.	0.7	Long Read Length	High error rate in homopolymer
HiSeq (Illumina)	Reversible Terminator	2*100	2 days (rapid mode)	120 (rapid mode)	High-throughput / cost	Short reads Long run time (normal mode)
SOLiD (Life)	Ligation	85	8 days	150	Low Error Rate	Short reads Long run time
Ion Proton (Life)	Proton Detection	200	2 hrs.	100	Short Run times	New*
PacBio RS	Real-time Sequencing	3000 (up to 15,000)	20 min	3	No PCR Longest Read Length	High Error Rate

1.9 STUDY RATIONALE

African populations are characterized by a high level of genetic diversity owing to a large number of variable genes and alleles (Cavalli-Sforza 1991; Jorde *et al.* 2000; Tishkoff and Williams 2002; Jakobsson *et al.* 2008). Patterns of genetic variation in the modern African population are influenced by a demographic history that includes changes in population size, admixture and locus-specific forces such as natural selection, recombination and mutation. To date, few genetic studies of structural variation of genes across ethnically diverse

populations have been conducted (Conrad and Hurles 2007). There are currently no studies within and between ethnically diverse African populations.

South Africa carries a higher burden of HIV infection than any other country in the world and embodies a rich collection of ethnic backgrounds, (figure 1.7). The major ethnic groups include the Bapedi, Basotho, Ndebele, Swati, Tsonga, Tswana, Xhosa, Venda and Zulu. The genetic substructure of these populations was assessed by studying the Y-chromosome and autosomal DNA of these populations (Lane *et al.* 2002). The Y-chromosome data demonstrated that people speaking these ethnic languages cluster into three specific groups (Tswana/Sotho, Nguni and Venda). Thus people belonging to these linguistic groups also show genetic differences. Because of their genetic variability, these populations are also likely to carry variable host cell genes (restriction and other factors) that may play a crucial role in viral replication.

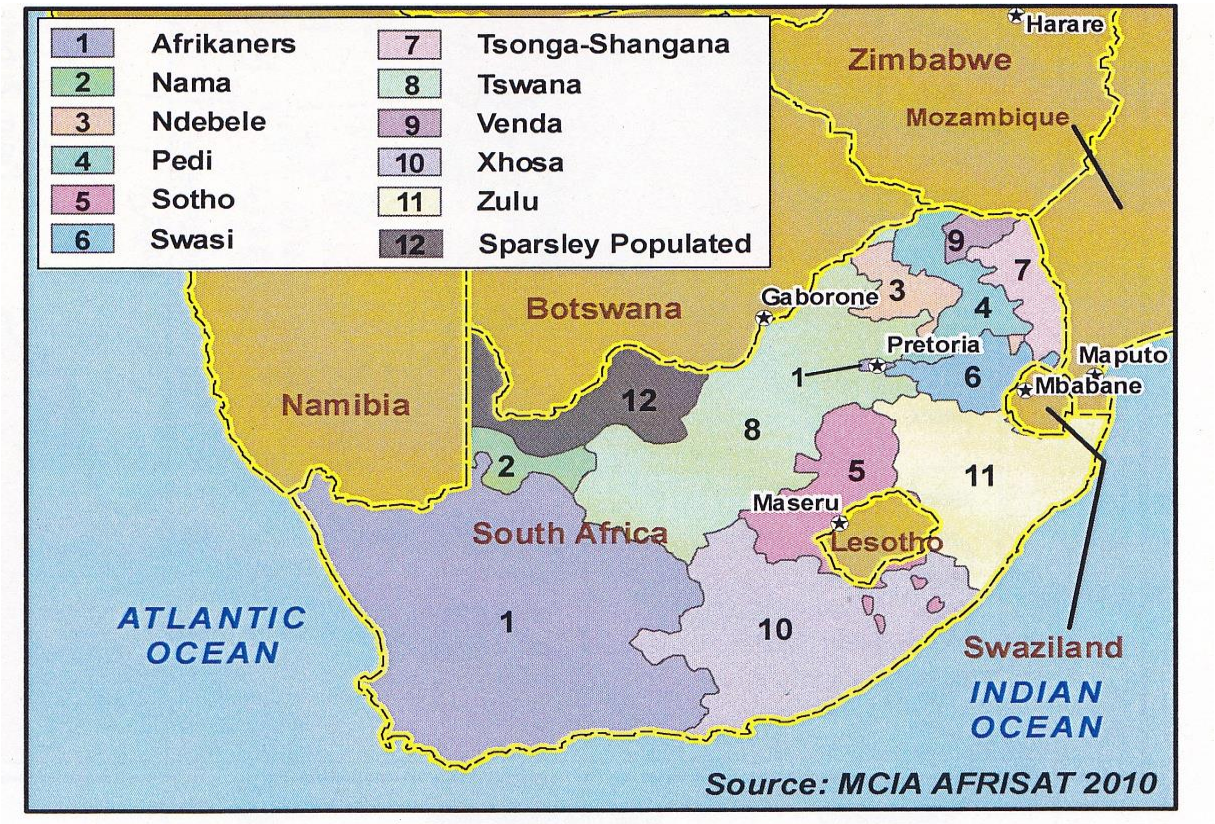


Figure 1.7: South African map showing geographic distribution of ethnic populations

As described above, host cell genes such as the C-C chemokine receptor type 5 (CCR5) play a major role in HIV-1 infection and disease progression. One study described how

polymorphisms within the open reading frame as well as the regulatory regions could influence the amount of CCR5 expressed on the cell surface, and hence individual susceptibility to HIV-1 infection (Picton *et al.* 2010). There are currently no studies describing CCR5 expression levels in ethnically diverse populations of South Africa and how expression levels may be influenced by variability. In addition, mutations in the viral envelope protein V3 loop, a major determinant for tropism, could affect syncytium formation, viral infectivity, neutralization, replication efficiency and host cell tropism.

A fundamental concept in the biology of viral restriction factors is that HIV-1 generally evades their potent inhibitory activities in human cells, thereby allowing viral replication to proceed efficiently (Malim and Bieniasz 2012). Many HIV restriction factors and their antagonist have been described. Of importance in this study is the Vif protein, which antagonizes apoliprotein APOBEC3 proteins, particularly APOBEC3D, APOBEC3F, APOBEC3G and APOBEC3H (Sheehy *et al.* 2002).

Since restriction and antagonizing factors interact with each other through direct protein-protein binding, both restriction factors and their viral counterpart proteins often show the signature of relatively rapid evolution. This evolution drives selection for beneficial mutations in restriction factor genes and their viral antagonists. As a consequence of positive selection, even phylogenetically similar species are likely to differ in terms of restriction factor functionality. Many APOBEC3 genes, including APOBEC3D, F, G and H have evolved under positive selection for millions of years in primates (Sawyer *et al.* 2004; OhAinle *et al.* 2008; Duggal *et al.* 2011; Duggal and Emerman 2012). Within humans, several APOBEC3 genes are known to possess common polymorphisms that render them defective, reduce antiviral activity, and increase sensitivity to HIV-1 *vif*. Furthermore, such polymorphisms are common in some populations, while not present in others.

Population-based studies have tried to establish a relationship between host factor polymorphism, expression or activity and the rate of disease progression (Pace *et al.* 2006; Vázquez-Pérez *et al.* 2009). Most studies have focused on Caucasian populations. In contrast, little information is available for the effects of variation in these genes in African populations, where the HIV epidemic has expanded at an alarming rate. Although several population studies have focused on African Americans, these do not give us a complete picture of the potential variation in Africans, though the studies can be a good guide from which to base additional studies. A more comprehensive analysis involving different African

ethnic groups will likely provide a better understanding of the mechanisms underlying host pathogen interactions, especially in view of the fact that African Americans are primarily infected with HIV subtype B, which is rarely seen in Africa.

1.10 OVERALL OBJECTIVE

The main objective of this study was to apply next generation sequencing (NGS) to characterize the variability of CCR5 and APOBEC3 genes and to determine how they influence HIV-1 evolution and genetic diversity in diverse populations of South Africa. This study was correlated with sequence analysis of the HIV-1 *env* V3 loop and *vif* regions.

1.11 SPECIFIC OBJECTIVES

Next Generation Sequencing (NGS) was used to characterize host cell DNA and HIV from ethnically diverse populations of South Africa with respect to:

1. HIV-1 gp120 V3 loop gene diversity.
2. Co-receptor usage (based on V3 loop sequences)
3. CCR5 polymorphism
4. APOBEC3 D, F, G and H polymorphism
5. HIV-1 *Vif* diversity

1.12 HYPOTHESES

It was hypothesized that:

1. HIV-1 subtype C chronically infected South Africans harbor a significant level of CXCR4 utilizing viruses.
2. There is limited variability in the CCR5 Maraviroc binding motif in HIV-1 chronically infected individuals.
3. There are variations in *apobec3* genes in South African populations that could significantly affect viral replication, transmission, fitness and evolution.
4. Variations in HIV-1 *vif* influences variation in *apobec3* at a genetic level.

1.13 THESIS OUTLINE

The next successive chapters describe specific topics that this study set out to investigate. **Chapter Two** describes next generation sequencing protocols to amplify and sequence CCR5 the APOBEC3 genes, HIV-1 *vif* and *env* V3 loop genes that were developed and validated. **Chapter Three** describes experiments investigating co-receptor usage and demonstrates that there is a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced South African individuals. **Chapter Four** describes experiments that characterize CCR5 gene variation and its predicted influence on Maraviroc binding. **Chapter Five** describes experiments that characterize variation in the APOBEC3 genes and its level of diversity in South African populations. **Chapter Six** explores the genetic characterization of HIV-1 Vif and correlation with APOBEC3 variation in chronically infected South African treatment experienced patients.

Each of the chapters consists of an abstract, introduction, materials and methods, results, discussion, conclusions and references. The general overview of the entire study is made where the main findings are highlighted in a context with recommendations and limitations in **Chapter Seven**.

1.14 REFERENCES

Arthos, J., Cicala, C., Martinelli, E., Macleod, K., Van Ryk, D., Wei, D., Xiao, Z., Veenstra, T.D., Conrad, T.P., Lempicki, R. a, McLaughlin, S., Pascuccio, M., Gopaul, R., McNally, J., Cruz, C.C., Censoplano, N., Chung, E., Reitano, K.N., Kottilil, S., Goode, D.J., Fauci, A.S. (2008) 'HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral T cells.', *Nature immunology*, 9(3), 301–309.

Barré-Sinoussi, F., Ross, A.L., Delfraissy, J.-F. (2013) 'Past, present and future: 30 years of HIV research', *Nature Reviews Microbiology*, 11(12), 877–883.

Baudhuin, L.M., Lagerstedt, S.A., Klee, E.W., Fadra, N., Oglesbee, D., Ferber, M.J. (2015) 'Confirming variants in next-generation sequencing panel testing by sanger sequencing', *Journal of Molecular Diagnostics*, 17(4), 456–461.

Berger, E.A., Doms, R.W., Fenyö, E.-M., Korber, B.T.M., Littman, D.R., Moore, J.P., Sattentau, Q.J., Schuitemaker, H., Sodroski, J., Weiss, R.A. (1998) 'A new classification for HIV-1', *Nature*, 391(6664), 240–240.

Bieniasz, P.D. (2003) 'Restriction factors: A defense against retroviral infection', *Trends in Microbiology*.

Björndal, a, Deng, H., Jansson, M., Fiore, J.R., Colognesi, C., Karlsson, a, Albert, J., Scarlatti, G., Littman, D.R., Fenyö, E.M. (1997) 'Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype.', *Journal of virology*, 71(10), 7478–7487.

Blow, N. (2008) 'DNA sequencing: Generation next-next', *Nature Methods*, 5(3), 267–274.

Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., Elledge, S.J. (2008) 'Identification of host proteins required for HIV infection through a functional genomic screen', *Science*, 319 (5865), 921–926.

Bushman, F.D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Konig, R., Bandyopadhyay, S., Ideker, T., Goff, S.P., Krogan, N.J., Frankel, A.D., Young, J.A., Chanda, S.K. (2009) 'Host cell factors in HIV replication: meta-analysis of genome-wide studies', *PLoS Pathog*, 5(5), e1000437.

Capobianchi, M.R., Giombini, E., Rozera, G. (2013) 'Next-generation sequencing technology in clinical virology', *Clinical Microbiology and Infection*, 19(1), 15–22.

Cavalli-Sforza, L.L. (1991) 'Genes, People and Languages', *Scientific American*, 265(5), 104–110.

Centers for Disease Control (1981) 'Pneumocystis Pneumonia - Los Angeles', *Morbidity and Mortality Weekly Report*, 30(6), 250–252.

Chan, D.C., Fass, D., Berger, J.M., Kim, P.S. (1997) 'Core structure of gp41 from the HIV envelope glycoprotein', *Cell*, 89(2), 263–273.

Chang, Y.S., Huang, H. Da, Yeh, K.T., Chang, J.G. (2017) 'Evaluation of whole exome sequencing by targeted gene sequencing and Sanger sequencing', *Clinica Chimica Acta*, 471, 222–232.

Connor, R.I., Sheridan, K.E., Ceradini, D., Choe, S., Landau, N.R. (1997) 'Change in Coreceptor Use Correlates with Disease Progression in HIV-1-Infected Individuals', *Journal of Experimental Medicine*, 185(4), 621–628.

Conrad, D.F., Hurles, M.E. (2007) 'The population genetics of structural variation', *Nature Genetics*, 39(7S), S30–S36.

Conticello, S.G., Harris, R.S., Neuberger, M.S. (2003) 'The Vif Protein of HIV Triggers Degradation of the Human Antiretroviral DNA Deaminase APOBEC3G', *Current Biology*, 13(22), 2009–2013.

Cormier, E.G., Dragic, T. (2002) 'The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor.', *Journal of virology*, 76(17), 8953–7.

Cullen, B.R. (2003) 'Nuclear mRNA export: Insights from virology', *Trends in Biochemical Sciences*.

Van Damme, N., Goff, D., Katsura, C., Jorgenson, R.L., Mitchell, R., Johnson, M.C., Stephens, E.B., Guatelli, J. (2008) 'The Interferon-Induced Protein BST-2 Restricts HIV-1 Release and Is Downregulated from the Cell Surface by the Viral Vpu Protein', *Cell Host and Microbe*, 3(4), 245–252.

Dang, Y., Lai, M.S., Wang, X., Han, Y., Lampen, R., Zheng, Y.H. (2008) 'Human cytidine deaminase APOBEC3H restricts HIV-1 replication', *Journal of Biological Chemistry*, 283(17), 11606–11614.

Douglas, N.W., Munro, G.H., Daniels, R.S. (1997) 'HIV/SIV glycoproteins: Structure-function relationships', *Journal of Molecular Biology*, 273(1), 122–149.

Duggal, N.K., Emerman, M. (2012) 'Evolutionary conflicts between viruses and restriction factors shape immunity', *Nature Reviews Immunology*.

Duggal, N.K., Malik, H.S., Emerman, M. (2011) 'The Breadth of Antiviral Activity of Apobec3DE in Chimpanzees Has Been Driven by Positive Selection', *Journal of Virology*, 85(21), 11361–11371.

Esbjörnsson, J., Månsson, F., Martínez-Arias, W., Vincic, E., Biague, A.J., da Silva, Z.J., Fenyö, E.M., Norrgren, H., Medstrand, P. (2010) 'Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease - indication of an evolving epidemic in West Africa', *Retrovirology*, 7.

Espinosa, R., Funahashi, T., Hadjiagapiou, C., Le Beau, M.M., Davidson, N.O. (1994) 'Assignment of the gene encoding the human apolipoprotein B mRNA editing enzyme (APOBEC1) to chromosome 12p13.1', *Genomics*, 24(2), 414–415.

Fakruddin, M., Chowdhury, A. (2012) 'Pyrosequencing-an alternative to traditional Sanger sequencing', *American Journal of Biochemistry and Biotechnology*.

Fisher, A.G., Ensoli, B., Ivanoff, L., Chamberlain, M., Petteway, S., Ratner, L., Llo, R.C., Wong-Staal, F. (1987) 'The sor Gene of HIV-1 is Required for Efficient Virus Transmission in Vitro', *Science*, 237(4817), 888–893.

Freed, E.O. (1998) 'HIV-1 Gag proteins: Diverse functions in the virus life cycle', *Virology*.

Friedman-Kien, A.E. (1981) 'Disseminated Kaposi's sarcoma syndrome in young homosexual men', *Journal of the American Academy of Dermatology*, 5(4), 468–471.

G., M., W., H. (2016) 'Analysis of drug resistance associated mutations in HCV and HIV using next generation sequencing', *Clinical Chemistry and Laboratory Medicine*, 54(5), eA3-eA4.

Gabuzda, D.H., Lawrence, K., Langhoff, E., Terwilliger, E., Dorfman, T., Haseltine, W.A., Sodroski, J. (1992) 'Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes.', *Journal of virology*, 66(11), 6489–95.

Gao, F., Balles, E., Robertson, D.L., Chen, Y., Rodenburg, C.M., Michael, S.F., Cummins, L.B., Arthur, L.O., Peeters, M., Shaw, G.M., Sharp, P.M., Hahn, B.H. (1999) 'Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes', *Nature*, 397(6718), 436–441.

Gao, J., Wu, H., Shi, X., Huo, Z., Zhang, J., Liang, Z. (2016) 'Comparison of next-generation sequencing, quantitative PCR, and sanger sequencing for mutation profiling of EGFR, KRAS, PIK3CA and BRAF in clinical lung tumors', *Clinical Laboratory*, 62(4), 689–696.

Garcia-Perez, J., Rueda, P., Alcami, J., Rognan, D., Arenzana-Seisdedos, F., Lagane, B., Kellenberger, E. (2011) 'Allosteric model of maraviroc binding to CC Chemokine Receptor 5 (CCR5)', *Journal of Biological Chemistry*, 286(38), 33409–33421.

Goldberg, S.M.D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., Li, K., Rogers, Y.-H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M., Venter, J.C. (2006) 'A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes', *Proceedings of the National Academy of Sciences*, 103(30), 11240–11245.

Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A., Saxon, A. (1981) 'Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency.', *The New England journal of medicine*, 305(24), 1425–31.

Griffin, H.G., Griffin, a M. (1993) 'DNA sequencing.', *Methods in molecular biology (Clifton, N.J.)*, 23, 1–8.

Grossman, Z., Meier-Schellersheim, M., Paul, W.E., Picker, L.J. (2006) 'Pathogenesis of HIV infection: What the virus spares is as important as what it destroys', *Nature Medicine*, 12(3), 289–295.

Harari, A., Ooms, M., Mulder, L.C.F., Simon, V. (2009) 'Polymorphisms and Splice Variants Influence the Antiretroviral Activity of Human APOBEC3H', *Journal of Virology*, 83(1), 295–303, available: http://jvi.asm.org/cgi/doi/10.1128/JVI.01665-08.

Harris, R.S. (2008) 'Enhancing immunity to HIV through APOBEC', *Nature Biotechnology*.

Hatse, S., Princen, K., Vermeire, K., Gerlach, L.O., Rosenkilde, M.M., Schwartz, T.W., Bridger, G., De Clercq, E., Schols, D. (2003) 'Mutations at the CXCR4 interaction sites for AMD3100 influence anti-CXCR4 antibody binding and HIV-1 entry', *FEBS Letters*, 546(2–3), 300–306.

Hultquist, J.F., Lengyel, J.A., Refsland, E.W., LaRue, R.S., Lackey, L., Brown, W.L., Harris, R.S. (2011) 'Human and Rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H Demonstrate a Conserved Capacity To Restrict Vif-Deficient HIV-1', *Journal of Virology*, 85(21), 11220–11234.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G.M., Roth, S.L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., Oǧdonoghue, A.J., Todorovic, A., Morris, J.H., Maltby, D.A., Alber, T., Cagney, G., Bushman, F.D., Young, J.A., Chanda, S.K., Sundquist, W.I., Kortemme, T., Hernandez, R.D., Craik, C.S., Burlingame, A., Sali, A., Frankel, A.D., Krogan, N.J. (2012) 'Global landscape of HIV-human protein complexes', *Nature*, 481(7381), 365–370.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C., Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., Van De Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A., Singleton, A.B. (2008) 'Genotype, haplotype and copy-number variation in worldwide human populations', *Nature*, 451(7181), 998–1003.

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., Navaratnam, N. (2002) 'An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22', *Genomics*, 79(3), 285–296.

Johnson, D.B., Dahlman, K.H., Knol, J., Gilbert, J., Puzanov, I., Means-Powell, J., Balko, J.M., Lovly, C.M., Murphy, B.A., Goff, L.W., Abramson, V.G., Crispens, M.A., Mayer, I.A., Berlin, J.D., Horn, L., Keedy, V.L., Reddy, N.M., Arteaga, C.L., Sosman, J.A., Pao, W. (2014) 'Enabling a Genetically Informed Approach to Cancer Medicine: A Retrospective Evaluation of the Impact of Comprehensive Tumor Profiling Using a Targeted Next-Generation Sequencing Panel', *The Oncologist*, 19(6), 616–622.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T., Batzer, M.A. (2000) 'The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data', *The American Journal of Human Genetics*, 66(3), 979–988.

Jour, G., Scarborough, J.D., Jones, R.L., Loggers, E., Pollack, S.M., Pritchard, C.C., L. Hoch, B. (2014) 'Molecular profiling of soft tissue sarcomas using next-generation sequencing: A pilot study toward precision therapeutics', *Human Pathology*, 45(8), 1563–1571.

Kaur, G., Mehra, N. (2009) 'Genetic determinants of HIV-1 infection and progression to AIDS: Susceptibility to HIV infection', *Tissue Antigens*.

Keele, B.F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E.M., Bienvenue, Y., Delaporte, E., Brookfield, J.F.Y., Sharp, P.M., Shaw, G.M., Peeters, M., Hahn, B.H. (2006) 'Chimpanzee reservoirs of pandemic and nonpandemic HIV-1', *Science*, 313(5786), 523–526.

König, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M.C., Irelan, J.T., Chiang, C. yuan, Tu, B.P., De Jesus, P.D., Lilley, C.E., Seidel, S., Opaluch, A.M., Caldwell, J.S., Weitzman, M.D., Kuhen, K.L., Bandyopadhyay, S., Ideker, T., Orth, A.P., Miraglia, L.J., Bushman, F.D., Young, J.A., Chanda, S.K. (2008) 'Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication', *Cell*, 135(1), 49–60.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T. (2000) 'Timing the ancestor of the HIV-1 pandemic strains', *Science*, 288(5472), 1789–1796.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., Yatim, A., Emiliani, S., Schwartz, O., Benkirane, M. (2011) 'SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx', *Nature*, 474(7353), 654–657.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M.E., Jonker, E., Freeman, C., Young, L., Morar, B., Toffie, L. (2002) 'Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies', *American Journal of*

Physical Anthropology, 119(2), 175–185.

Liao, W., Hong, S.H., Chan, B.H.J., Rudolph, F.B., Clark, S.C., Chan, L. (1999) 'APOBEC-2, a cardiac- and skeletal muscle-specific member of the cytidine deaminase supergene family', *Biochemical and Biophysical Research Communications*, 260(2), 398–404.

Madani, N., Kabat, D. (2000) 'Cellular and Viral Specificities of Human Immunodeficiency Virus Type 1 Vif Protein', *J. Virol.*, 74(13), 5982–5987.

Malim, M.H., Bieniasz, P.D. (2012) 'HIV restriction factors and mechanisms of evasion', *Cold Spring Harbor Perspectives in Medicine*, 2(5).

Malim, M.H., Emerman, M. (2008) 'HIV-1 Accessory Proteins-Ensuring Viral Survival in a Hostile Environment', *Cell Host and Microbe*.

Mardis, E.R. (2008) 'The impact of next-generation sequencing technology on genetics', *Trends in Genetics*.

Mehle, A., Goncalves, J., Santa-Marta, M., McPike, M., Gabuzda, D. (2004) 'Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation', *Genes and Development*, 18(23), 2861–2866.

Men, A.E., Wilson, P., Siemering, K., Forrest, S. (2008) 'Sanger DNA Sequencing', in *Next Generation Genome Sequencing: Towards Personalized Medicine*, 1–11.

Muto, T., Muramatsu, M., Taniwaki, M., Kinoshita, K., Honjo, T. (2000) 'Isolation, tissue distribution, and chromosomal localization of the human activation-induced cytidine deaminase (AID) gene', *Genomics*, 68(1), 85–88.

Neil, S.J.D., Zang, T., Bieniasz, P.D. (2008) 'Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu', *Nature*, 451(7177), 425–430.

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S. (2011) 'Genotype and SNP calling from next-generation sequencing data', *Nature Reviews Genetics*.

Nisole, S., Stoye, J.P., Saïb, A. (2005) 'TRIM family proteins: Retroviral restriction and antiviral defence', *Nature Reviews Microbiology*.

OhAinle, M., Kerns, J.A., Li, M.M.H., Malik, H.S., Emerman, M. (2008) 'Antiretroelement

Activity of APOBEC3H Was Lost Twice in Recent Human Evolution', *Cell Host and Microbe*, 4(3), 249–259.

Pace, C., Keller, J., Nolan, D., James, I., Gaudieri, S., Moore, C., Mallal, S. (2006) 'Population level analysis of human immunodeficiency virus type 1 hypermutation and its relationship with APOBEC3G and vif genetic variation.', *Journal of virology*, 80(18), 9259–69.

Palella, F.J., Delaney, K.M., Moorman, A.C., Loveless, M.O., Fuhrer, J., Satten, G.A., Aschman, D.J., Holmberg, S.D. (1998) 'Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators.', *The New England journal of medicine*, 338(13), 853–60.

Paredes, R. (2012) 'Importance of minor variants and their detection (ultra-deep sequencing) in the management of HIV infection', *Journal of the International AIDS Society*, 15(Suppl 4), 18309.

Pastore, C., Nedellec, R., Ramos, A., Pontow, S., Ratner, L., Mosier, D.E. (2006) 'Human Immunodeficiency Virus Type 1 Coreceptor Switching: V1/V2 Gain-of-Fitness Mutations Compensate for V3 Loss-of-Fitness Mutations', *Journal of Virology*, 80(2), 750–758.

Picton, A.C.P., Paximadis, M., Tiemessen, C.T. (2010) 'Genetic variation within the gene encoding the HIV-1 CCR5 coreceptor in two South African populations.', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 10(4), 487–94.

Poon, A.F.Y., Lewis, F.I., Kosakovsky Pond, S.L., Frost, S.D.W. (2007) 'An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope', *PLoS Computational Biology*, 3(11), 2279–2290.

Refsland, E.W., Hultquist, J.F., Harris, R.S. (2012) 'Endogenous origins of HIV-1 G-to-A hypermutation and restriction in the nonpermissive T cell line CEM2n', *PLoS Pathogens*, 8(7), 39.

Richman, D.D., Bozzette, S.A. (1994) 'The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression', *J Infect Dis*, 169(5), 968–974.

van Rij, R.P., Visser, J.A., van Praag, R.M.E., Rientsma, R., Prins, J.M., Lange, J.M.A., Schuitemaker, H. (2002) 'Both R5 and X4 Human Immunodeficiency Virus Type 1 Variants Persist during Prolonged Therapy with Five Antiretroviral Drugs', *J. Virol.*, 76(6), 3054–3058.

SANAC (2011) 'National Strategic Plan on HIV, STIs and TB 2012-2016.', *Sanac 2011*, 84.

Sander, O., Sing, T., Sommer, I., Low, A.J., Cheung, P.K., Harrigan, P.R., Lengauer, T., Domingues, F.S. (2007) 'Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage', *PLoS Computational Biology*, 3(3), 0555–0564.

Sanger, F., Nicklen, S., Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467, available: http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463.

Sawyer, S.L., Emerman, M., Malik, H.S. (2004) 'Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G', *PLoS Biology*, 2(9).

Scarlatti, G., Tresoldi, E., Björndal, Å., Fredriksson, R., Colognesi, C., Kui Deng, H., Malnati, M.S., Plebani, A., Siccardi, A.G., Littman, D.R., Maria Fenyö, E., Lusso, P. (1997) 'In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression', *Nature Medicine*, 3(11), 1259–1265.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H. (2002) 'Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein', *Nature*, 418(6898), 646–650.

Shendure, J., Ji, H. (2008) 'Next-generation DNA sequencing', *Nat Biotechnol*, 26(10), 1135–1145.

Sherman, M.P., De Noronha, C.M.C., Williams, S. a, Greene, W.C. (2002) 'Insights into the biology of HIV-1 viral protein R.', *DNA and cell biology*, 21(9), 679–88.

South African National Department of Helath (2010) 'The 2011 National Antenatal Sentinel HIV & Syphilis Prevalence Survey in South Africa', *Health (San Francisco)*, 1–72.

Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., Sodroski, J. (2004) 'The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys.', *Nature*, 427(6977), 848–853.

Suphaphiphat, P., Thitithanyanont, A., Paca-Uccaralertkun, S., Essex, M., Lee, T.-H. (2003) 'Effect of amino acid substitution of the V3 and bridging sheet residues in human immunodeficiency virus type 1 subtype C gp120 on CCR5 utilization.', *Journal of virology*, 77(6), 3832–7.

Tersmette, M., Gruters, R.A., de Wolf, F., de Goede, R.E., Lange, J.M., Schellekens, P.T., Goudsmit, J., Huisman, H.G., Miedema, F. (1989) 'Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates.', *Journal of virology*, 63(5), 2118–25.

Tishkoff, S.A., Williams, S.M. (2002) 'Genetic analysis of African populations: human evolution and complex disease', *Nat Rev Genet*, 3(8), 611–621.

Unaids (2017) UNAIDS Fact Sheet - Latest Statistics on the Status of the AIDS Epidemic [online], *ending the aids epidermics*.

UNAIDS (2013) *UNAIDS 2011-2015 Strategy: Getting to Zero* [online], 2011-2015 Strategy.

Vandekerckhove, L., Verhofstede, C., Vogelaers, D. (2009) 'Maraviroc: Perspectives for use in antiretroviral-naive HIV-1-infected patients', *Journal of Antimicrobial Chemotherapy*, 63(6), 1087–1096.

Vázquez-Pérez, J.A., Ormsby, C.E., Hernández-Juan, R., Torres, K.J., Reyes-Terán, G. (2009) 'APOBEC3G mRNA expression in exposed seronegative and early stage HIV infected individuals decreases with removal of exposure and with disease progression', *Retrovirology*, 6.

Wang, X., Abudu, A., Son, S., Dang, Y., Venta, P.J., Zheng, Y.-H. (2011) 'Analysis of human APOBEC3H haplotypes and anti-human immunodeficiency virus type 1 activity.', *Journal of virology*, 85(7), 3142–52.

Wedekind, J.E., Dance, G.S.C., Sowden, M.P., Smith, H.C. (2003) 'Messenger RNA editing in mammals: New members of the APOBEC family seeking roles in the family business', *Trends in Genetics*.

Woollard, S.M., Kanmogne, G.D. (2015) 'Maraviroc: a review of its use in HIV infection and beyond.', *Drug design, development and therapy*, 9, 5447–68.

Yeung, M.L., Houzet, L., Yedavalli, V.S.R.K., Jeang, K.T. (2009) 'A genome-wide short hairpin RNA screening of Jurkat T-cells for human proteins contributing to productive HIV-1 replication', *Journal of Biological Chemistry*, 284(29), 19463–19473.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., Yu, X.F. (2003) 'Induction of APOBEC3G Ubiquitination and Degradation by an HIV-1 Vif-Cul5-SCF Complex', *Science*, 302(5647), 1056–1060.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., Gao, L. (2003) 'The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA', *Nature*, 424(6944), 94–98.

Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J., Espeseth, A.S. (2008) 'Genome-scale RNAi screen for host factors required for HIV replication.', *Cell Host & Microbe*, 4(5), 495–504.

CHAPTER TWO

DEVELOPMENT OF LIBRARY PREPARATION PROTOCOLS FOR SEQUENCING HOST AND HIV-1 GENES

ABSTRACT

Background: Next Generation Sequencing (NGS) has made possible the sequencing of large amounts of DNA in a single machine run, but the DNA library preparation is expensive and time consuming. One of the most common NGS platforms is Illumina, which uses Tagmentation-based library preparation procedure mediated by a hyperactive variant of the transposase (Tn5). This procedure represents major advances in the preparation of sequencing libraries. Unfortunately, it is proprietary to Illumina and remains prohibitively expensive. The cost of these commercially available reagents limits its use in small or medium sized projects in many settings. Therefore the objective of the study was to generate less expensive procedures using high quality homemade tagmentation-ready Tn5 transposase and or custom amplicon based methods for library preparation.

Methods: Two approaches were applied to prepare libraries; Direct sequencing of short amplicons and Tn5 transposase tagmentation of large amplicons. Libraries prepared using both these methods were sequenced in a MiniSeq or MiSeq instrument. In some cases libraries were also prepared with an Illumina Nextera XT kit as control for comparison.

Results: The run profile of libraries prepared using the homemade Tn5 and custom amplicon-base methods was analyzed to document the number of cycles, data yield, error rate and quality score. Generally a MiSeq run for 2x 300bp paired end reads using a V3 reagent kit and a MiniSeq paired end reads 2x150- bp reads option with the MiniSeq High Output reagents (300 cycles) generated about 23.2 to 15 Gb data, with 44-50 million reads passing filter, 1200-1400 kmm2 cluster passing filter giving a QC30 of >70%.

Conclusion: Custom based Tn5 tagmentation and amplicon based library preparation protocols to sequence host cell and HIV-1 genes were developed and validated to lower costs in existing and novel massively parallel, sequencing-based applications, which will allow us to harness the full potential of NGS in all types of projects.

Key words: Next generation sequencing, Tn5 transposase, custom amplicon based, library preparation protocols

2.1 INTRODUCTION

Next Generation Sequencing (NGS), also referred to as deep sequencing, high-throughput sequencing or massively parallel DNA sequencing is a non-Sanger-based DNA sequencing technology, where a million to billion of DNA strands can be sequenced in parallel yielding substantially more throughput than automated Sanger sequencing.

For the past decades, limitations of sequencing methods have been limiting researchers in their ability to explore the human genome. The incredible increase in sequencing technology (Metzker 2010) has revolutionized DNA and RNA sequencing (RNA-seq) applications with improved throughput. This has lowered the cost per base by more than a 100,000-fold (Lander 2011). For example, expression profiles from hundreds of different single cells can be generated in a single lane of an Illumina HiSeq 2000 (Hashimshony *et al.* 2012; Islam *et al.* 2014; Jaitin *et al.* 2014). While acknowledging the drastic decrease in sequencing costs, there are other limiting factors such as library preparation costs and labor intensiveness and this represents a severe constraint in many next generation sequencing–based projects, especially now that we can sequence thousands of DNA molecules simultaneously (Eberwine *et al.* 2014; Head *et al.* 2014; Sandberg 2014).

One of the main steps in NGS applications is library preparation, which employs fragmentation of long DNA fragments and adding sequencing adaptors. Different NGS platforms rely on unique chemistries and methods of preparing libraries. Illumina is the most common platform, which uses a DNA tagmentation procedure that represents a major advance in the preparation of sequencing libraries using a hyperactive variant of the transposase (Tn5). This enzyme fragments and simultaneously ligates sequencing adaptors on both ends of DNA in a 5 minutes reaction. (Adey *et al.* 2010). In current commercial solutions, the Illumina Nextera DNA kit is used because it is rapid, straightforward and requires low input DNA. Unfortunately, this technology is proprietary to Illumina and remains prohibitively expensive for most labs. Although the data quality and reproducibility are high when using Illumina Nextera kits, the cost of these commercial reagents often limits its use in small or medium sized projects. Therefore, lowered reagent costs in sequence library generation are needed to harness the full potential of current sequencing technology and to make it more affordable for laboratories in resource limited settings. The production of homemade Tn5 transposase and custom amplicon based library preparation significantly cuts down costs and allow sequencing genes of interest in small and medium size projects.

2.2 Objective

The main objective was to develop cost-effective library preparation protocols to sequence selected host and HIV-1 genes.

2.2.1 Specific objectives

2.2.1.1 To develop library preparation protocols for amplifying and sequencing host cell genes.
2.2.1.2 To develop library preparation protocols for amplifying and sequencing HIV-1 genes.
2.2.1.3 To validate the developed library preparation protocols using Illumina Nextera XT kits.

2.3 MATERIALS AND METHODS

2.3.1 Ethical considerations

The study protocol was approved by the Research Ethics Committee of the University of Venda (SMNS/16/MBY/28) and the University of Virginia Institutional Review Board (IRB-HSR #16815). Signed informed consent was obtained from study subjects prior to blood draw and collection of demographic and clinical data. Personal identifiers were stripped prior to sample processing and data analysis.

2.3.2 Study population and DNA extraction

The study population was comprised of HIV positive individuals who presented for routine care at the Bela-Bela HIV/AIDS Prevention Group Wellness clinic (HAPG), Donald Fraser Hospital Hope Clinic (DFHC) in Vhufhuli and Polokwane Hospital in the Limpopo province of Northern South Africa. These individuals were recruited from July 2013 to December 2015. DNA was extracted from peripheral blood mononuclear cells (PBMC) of 192 HIV infected individuals, using QIAamp DNA blood mini kit according to the manufacturer's instructions (Qiagen) and the DNA concentration ranged from 1-780ng/ul.

2.3.3 Study approach and strategy

Two approaches were used to develop library preparation protocols to sequence host and HIV-1 genes. The approaches are as follows; (1) Tn5 tagmentation library preparation using large amplicons and (2) Custom amplicon based library preparation using short amplicons. Libraries prepared using these approaches were sequenced in parallel with libraries prepared using an Illumina Nextera XT kit for validation. These approaches were based on the length of the amplicons considered for sequencing and the Illumina platform used. Two Illumina platforms (MiSeq and MiniSeq) were used in this study. An Illumina MiSeq can sequence a maximum of 2x 300 bp forward and reverse reads, while an Illumina MiniSeq can sequence a maximum of 2x 150bp forward and reverse reads. Primers were designed based on each approach. Figure 2.1 shows an overview of the workflow.

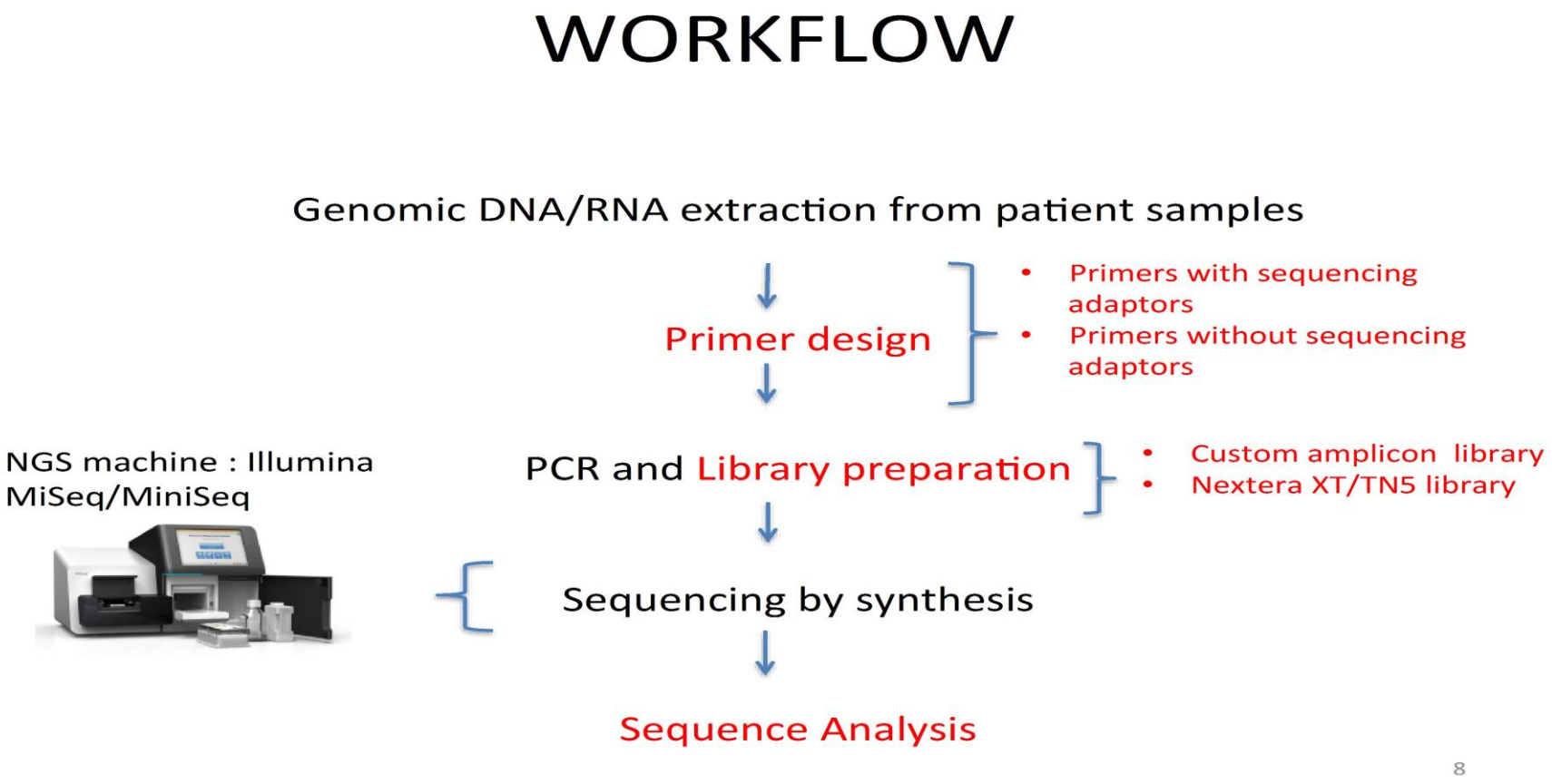


Figure 2.1: An overview the NGS methodology.

2.3.3.1 Approach 1: Library preparation using the homemade Tn5 transposase

This approach was applied mainly for preparing libraries from host cell genes because of their long amplicon size. Amplicons from five host genes were synthesized in this approach and are as follows; CCR5, APOBEC3D (A3D), APOBEC3F (A3F), APOBEC3G (A3D) and APOBEC3H (A3H). Only coding regions were targeted for sequencing. These genes are located at different chromosome shown in figure 2.2 and 2.3. Gene specific primers designed for this approach were without overhang/ adapter sequences because sequencing adapters were simultaneously added during the tagmentation reaction. This is different from custom amplicon based methods where adapter sequences were added during primer design.

2.3.3.1.1 CCR5 gene location in human chromosome and primer map

The CCR5 gene is located in human chromosome 3 from 46,372,670 to 46,373,961 forward strands. This gene, consist of one coding exon. Only two sets of primers for first and nested PCR were designed to amplify the coding region of this gene as amplicon 1, shown in figure 2.2.

Figure 2.2: CCR5 gene map and location of primers.

2.3.3.1.2 APOBEC3 genes map in human chromosome and primer location

The overlapping APOBEC3 genes are located in human chromosome 22. A3D is from 39,014,362 to 39,032,316 long with 7 exons. The overlapping A3F to A3D starts from 39, 040,604 to 045,192, with 7 exons. The A3G starts from 39,040,961 to 39,087,421, with 8 exons and A3H starts from 39,097,224 to 39,097,343, with 5 exons. Primers to amplify these genes were designed to target only the coding regions in different sets of amplicons shown in figure 2.3.

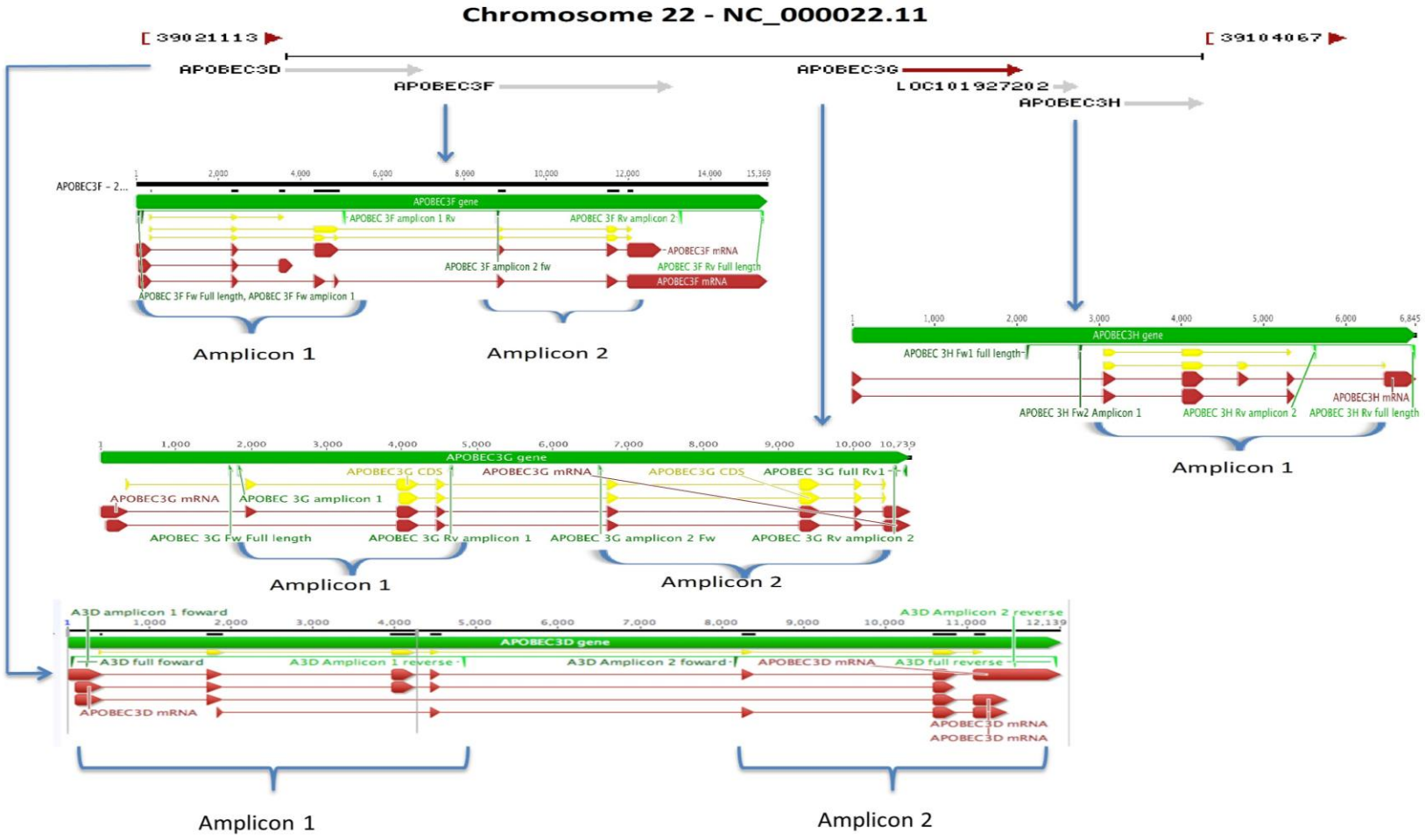


Figure 2.3: APOBEC3 gene map showing the location of amplified regions. Only the exons were amplified for A3D, F and G (Amplicons 1 and 2); and only 1 amplicon for H.

2.3.3.1.3 Primer design for host genes

Primers to amplify the four APOBEC3 genes (A3D, A3F, A3G, and A3H) and CCR5 were designed using Primer3 2.3.4 tool within Geneious® 8.1.5 software (Biometters Ltd). For each APOBEC3 gene except for A3H, three sets of primers were designed: The first set amplified a complete gene fragment in a first external polymerase chain reaction (PCR), while the last 2 sets of primers amplified two fragments of each gene in a nested PCR using the first round product as the template. The first set of APOBEC3H primers amplified amplicon 1, which incorporates all the exons. Because this gene is the most variable, several sets of nested primers targeting exons with the most variation were designed. Two sets of primers, first round and nested primers were designed to amplify CCR5 as amplicon 1. The first long fragments comprised of both exons and introns while exons were mostly targeted in the nested PCR. These primers were mapped in Ensembl genome browser to transcripts of the following genes; ENSG00000160791, ENSG00000243811, ENSG00000128394, ENSG00000239713 and ENSG00000100298 for CCR5, A3D, A3F, A3G and A3H respectively. The designed primers were tested on DNA from HEK 293T cells for specificity and only amplified expected fragments. The primer names, sequences and product sizes are shown in table 2.1.

Table 2.1: List of APOBEC3 primers designed; primer name, sequence and product size are indicated.

Name	Sequence (5'-3')	Product size
A3D (12.1 kb)		
A3D Forward primer	AGGAAGCCTCGCTCTCTCA	12,069 bp
A3D Forward primer	CAGGCAGGGTCTTGATCTGT	
A3D Amplicon 1F	AAAAAGAGGGAGACTGGGACAAGCGTATCTAAGA	4,300 bp
A3D Amplicon 1R	GAGTGTGGGTGAGGGGGTGTAACCATGAG	
A3D Amplicon 2F	AGCTAGGAGAGGTCACCCTG	3,188 bp
A3D Amplicon 2F	CAGGAGGCTAGAAGAGACAGACCATGAGGC	
A3F (13.31 kb)		
A3F 1st round f	ACCAGAAAGAGGGTGAGAGACTGAGGAAGATAAAG	13,142 bp
A3F 1st round rv	AGCCATTTATTGCAGAAGCTATGGATAAAGCTGGT	
A3F Amplicon 1 f	ACCAGAAAGAGGGTGAGAGACTGAGGAAGATAAAG	4,918 bp
A3F Amplicon 1 rv	GGGTGAGGGGTGTAACCATG	
A3F Amplicon 2 f	TTCAGAAACCCGATGGAGGC	4,478 bp
A3F Amplicon 2 rv	AGCCATTTATTGCAGAAGCTATGGATAAAGCTGGT	
A3G (10.74 kb)		
A3G 1st round f	TGTTAACCAGAGGCTGCTCTTCCCAGG	11,852 bp
A3G 1st round rv	TCCCTGGGACTCAGCTCC	
A3G Amplicon 1 f	ATTTGTCCCCAGCTCTGTGG	3,231 bp

A3G Amplicon 1 rv	AGAGGACCTGGTCTGGAACA	
A3G Amplicon 2 f	CAAGGGAGGAAGCGTGGAG	3,908 bp
A3G Amplicon 2 rv	TGCATTGCTTTGCTGGTGTC	
A3H (6.8 kb)		
APOBEC3 H forward primer full length	TCTGTTGCACAGAAACACGATGG	3522bp
APOBEC3 H reverse primer full length	CAACTGACATGCCCCAGGG	
APOBEC3 H forward primer Exon2 (A3HfE2)	TCTGTTGCACAGAAACACGATGG	452bp
APOBEC3 H Reverse primer Exon 2(A3HrE2)	TTCCCGAAGTAGTGACTGAGC	
APOBEC3 H forward primer Exon 3 &4(A3HfE3/4	GCCACGCACTAGAAAGTTCAC	934bp
APOBEC3 H Reverse primer Exon 3&4(A3HrE3/4)	ACAGTGCCTCACCTTTATCC	

2.3.3.1.4 Polymerase chain reaction (PCR) to amplify CCR5, A3D, A3F, A3G and A3H genes

Genomic DNA was extracted and the Takara (LA) PCR Kit Ver. 2.1 for long DNA fragments amplification (Clontech) was used to amplify a 12.16 kb A3D, 13.31 kb A3F, 10.74 kb A3G, 6.8 kb A3H and a 6.32kb CCR5 region from each sample in a 1st round 30 cycle PCR. The 1st round PCR products of each complete gene were used as templates in a nested PCR reaction to generate shorter fragments of these genes in a 30 cycle nested PCR reaction. Both the 1st and nested PCRs were in a 20µl reaction consisting of 1X PCR Mg^{2+} plus buffer, 400µM dNTPs, 0.2µM each primer (table 2.1) and 1.25 units of Takara LA Taq high fidelity

polymerase with the following cycling conditions: Initial denaturation at 94°C for 1 min, 30 cycles of denaturation at 98°C for 10sec, annealing between 53°C and 68°C for 15 min and extension at 72°C for 10 min.

2.3.3.1.5 Purification, normalization and pooling/multiplexing all targeted genes per sample

The amplicons were purified using AMpure XP beads (Beckman coulter) and quantified using a Qubit 3 with the dsDNA HS kit (Invitrogen). Equimolar concentrations of the shorter amplicons were pooled first to have all targeted exons per gene, followed by pooling all the genes per sample. The library pool per sample were normalized to 10ng (5µl of 2ng/µl) using a 10mM Tris elution buffer.

2.3.3.1.6 Fragmentation and tagmentation using the Tn5 transposase

The Tn5 transposase enzyme employs similar principles as the Illumina Nextera Kit, which fragments DNA and simultaneously adds adapter sequences (figure 2.4). The Tn5 transposase enzyme in this study was produced and characterized in the Ham-Rek lab, University of Virginia using a published protocol (Picelli *et al.* 2014). This enzyme was used to fragment about 1-10ng DNA of the library pool in a reaction mixture of 4µl tagmentation buffer (5X TAPS-DMF), 1-5µl Tn5 transposase (1X-5X) and 10ng DNA, with an addition of nuclease free water to add up to a final volume of 20µl. The reaction was placed on a thermal cycler at 55°C for 6 min. Zymo-column or 0.2% SDS was used to remove transposase from dsDNA after the incubation to avoid interference with PCR reaction. Zymo-columns were used when DNA input was >1ng and when input DNA is ≤ 1ng 5µl 0.2% SDS was added into the mixture and incubated at room temperature for 5 minutes. To add unique Illumina dual-index barcodes; Index 1 (i7) and index 2 (i5), a 50µl reaction was prepared with 25µl Kappa HiFi HotStart mix, 5µl of each index, i7 and i5, 5µl of tagmentation reaction and 10µl of water. The reaction was mixed and placed in a thermal cycler for a 12 cycle PCR program (table 2.2)

Table 2.2 Thermal cycling conditions

Steps	Temperature	Time
Initial denaturation	72ºC	3 min
Denaturation	95ºC	30 sec
Denaturation	95ºC	10 sec
Annealing	55ºC	10 sec
Extension	72ºC	30 sec
Incubation	72ºC	5 min
Storage	4ºC	Infinity

12 Cycles

2.3.3.1.7 Fragmentation and tagmentation using the Nextera XT kit as control

The Illumina Nextera XT kit was used to fragment 1ng input DNA of the library pool from the above protocol and tag with the Illumina sequencing adapters at both ends of amplicons at 55ºC for 5 min. Using the full complement of Nextera XT indices, up to 96 samples were pooled for each sequencing run.

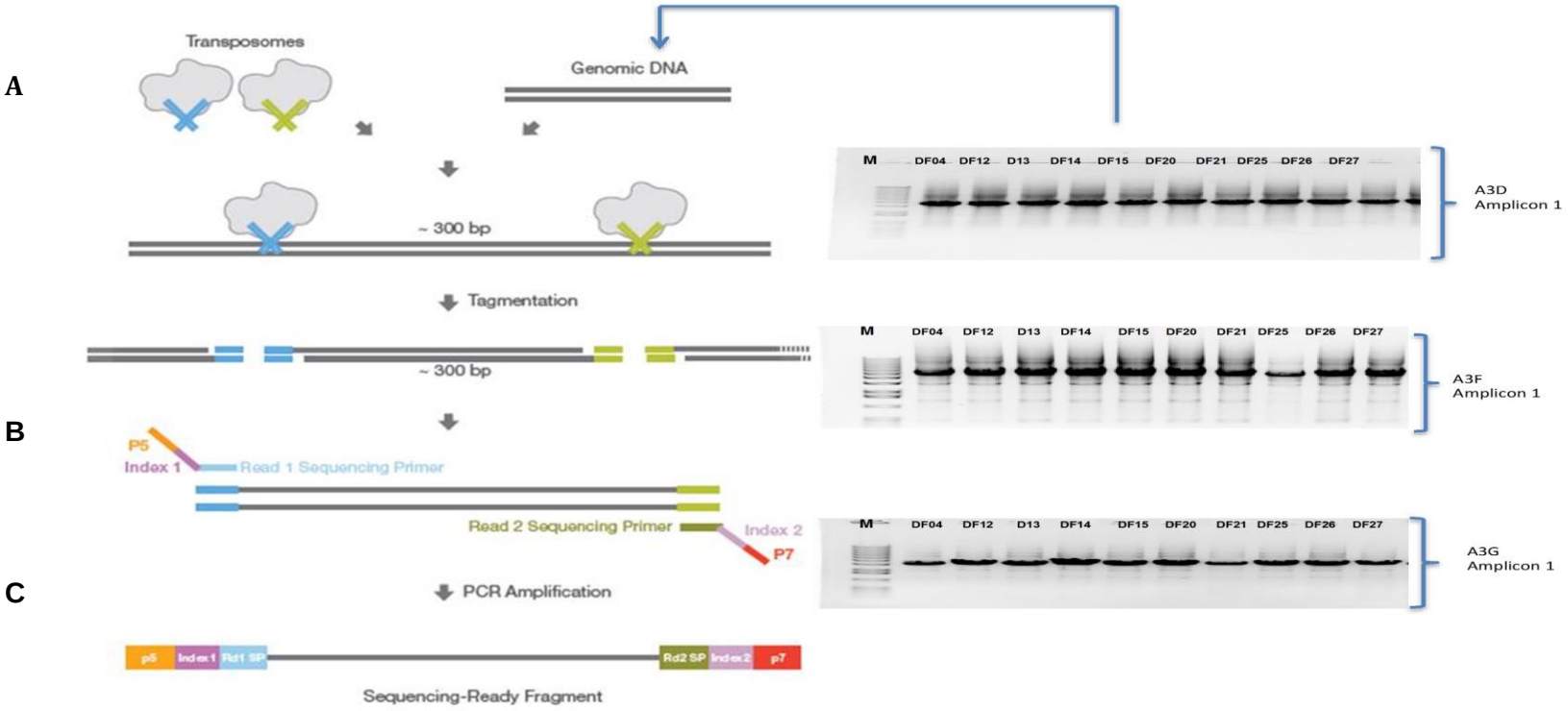


Figure 2.4: Schematic representation of library preparation using a Nextera XT kit/custom Tn5 tagmentation protocol, with amplicon1 from A3D, A3F and A3F. Image (A) are the amplicons (genomic DNA) subjected to fragmentation and tagmentation with either the Nextera XT kit or the custom Tn5, (B) shows the tegmentation product undergoing limited PCR to add adaptors; and (C) is the end product of the limited cycle PCR with the added sequence adaptors and Indices.

2.3.3.1.8 Library normalization, pooling and sequencing

The libraries prepared using both the Nextera kit and the Tn5 transposase were purified as above with the AMPure beads. They were then size verified using a bionalalyzer 2100 Expert with a High Sensitive DNA assay kit (Agilent Genomics) profile shown in figure 2.5, quantified using Qubit 3 with the dsDNA HS kit (Invitrogen) and normalized to 4nM each.

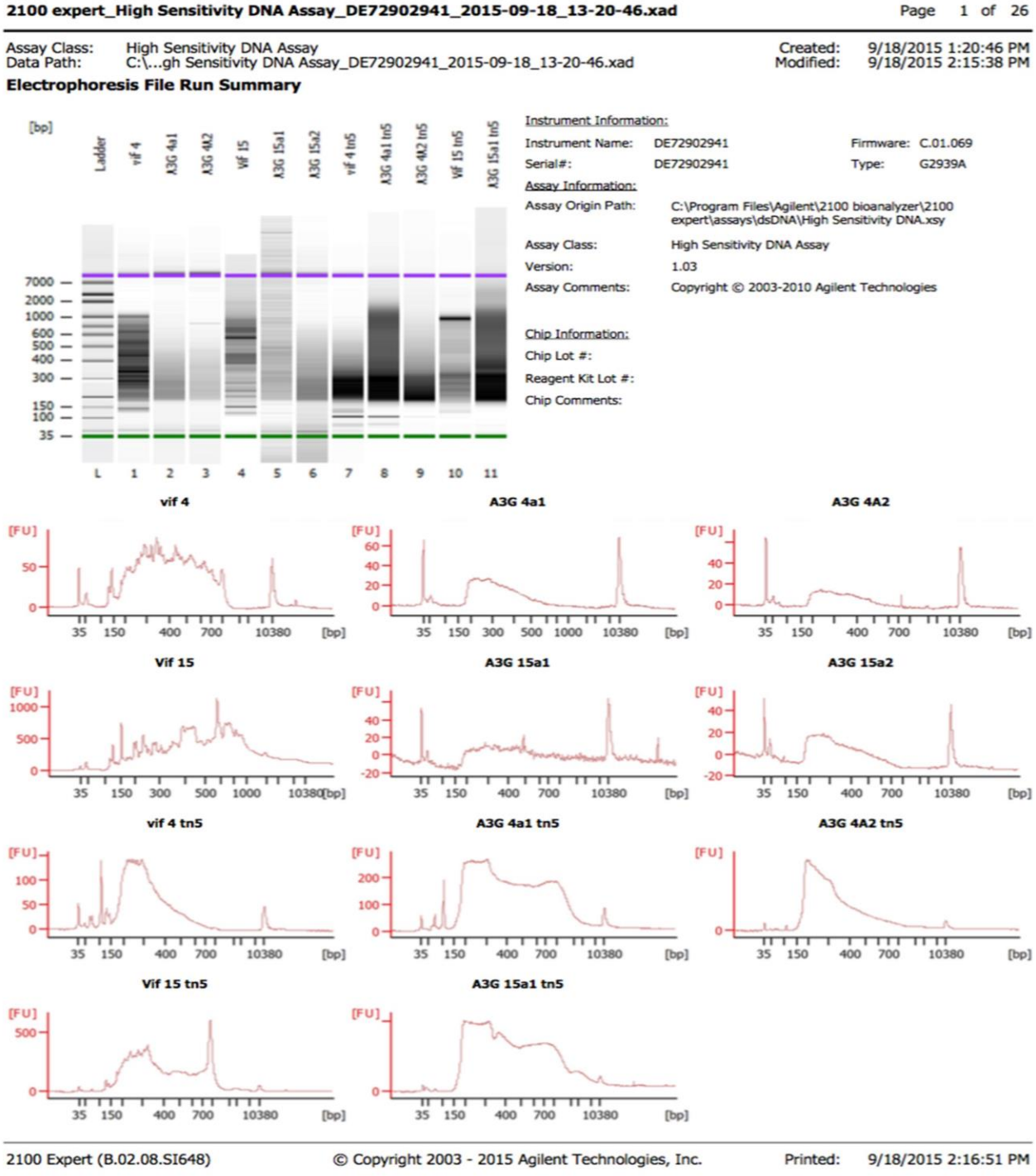


Figure 2.5: Size and concentration profiles of prepared libraries using a Bioanalalyzer 2100 Expert with a High Sensitive DNA assay kit (Agilent Genomics).

The normalized barcoded libraries per sample were then pooled, and denatured into single strands. For good cluster generation, about 1.8pM of the library spiked with 25-30% PhiX (Illumina) was loaded into the sequencing cartridge (Illumina MiSeq Reagent v3 kit). Biological sample sheets were created in Basespace by giving each sample the appropriate indice and setting up a sequencing run for the MiSeq.

2.3.3.2 Approach 2: Library preparation using short custom amplicon based methods

This approach was mostly utilized for preparing libraries for the HIV-1 genes because of their short length. In this approach, regions of interest were amplified as overlapping amplicons of less than 250bp thus eliminating the need for fragmentation. Therefore, nested primers were modified by adding adaptor sequences at the 5' end of the forward and 3' of the reverse primer.

2.3.3.2.1 Location of HIV-1 genes of interest and primer design

The HIV-1 *vif* gene is located between position 5041 to 5619bp in the HIV HXB2 reference sequence (figure 2.6 Box A), while the V3 loop is part of gp120 within the *env* region (figure 2.6 Box B).

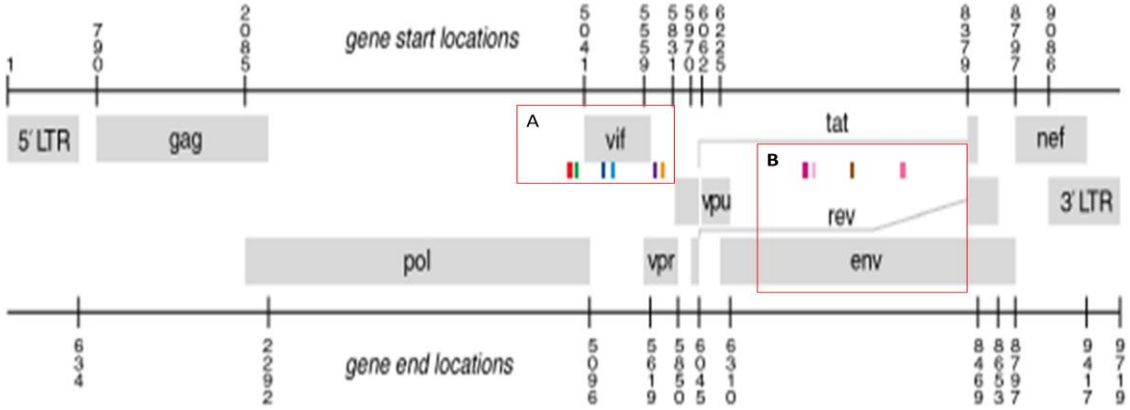


Figure 2.6: HIV-1 genome map showing primer location. Text box A indicates *Vif* gene and primer location with the red and orange highlights indicating outer primers and the rest indicating nested primers. Text box B indicates the location of the 105bp V3 loop within the envelope gene with the dark pink highlights indicating outer primers and the light pink and brown indicating the nested primers.

2.3.3.2.2 Primer design for HIV-1 genes

Gene-specific primers between 18-21bp forward and reverse were designed using Primer3 2.3.4 tool within Geneious® 8.1.5 software and blasted in the HIV-1 sequence database using the conventional quick align tool. The first set of primers targeted the complete genes in the first round and short overlapping amplicons were created in the second round in nested PCR. Transposase adapter sequences of about 33bp designed by Illumina were added to all forward and reverse nested primer. The primers for both genes were mapped to subtype C Vif and Env V3 loop reference sequences using QuickALIGN V2 tool in HIV database (figure 2.6) and tested on a subtype C based plasmid (IndieC) for specificity. They only amplified expected fragments. The primer names, sequences and product sizes are shown in table 2.2.

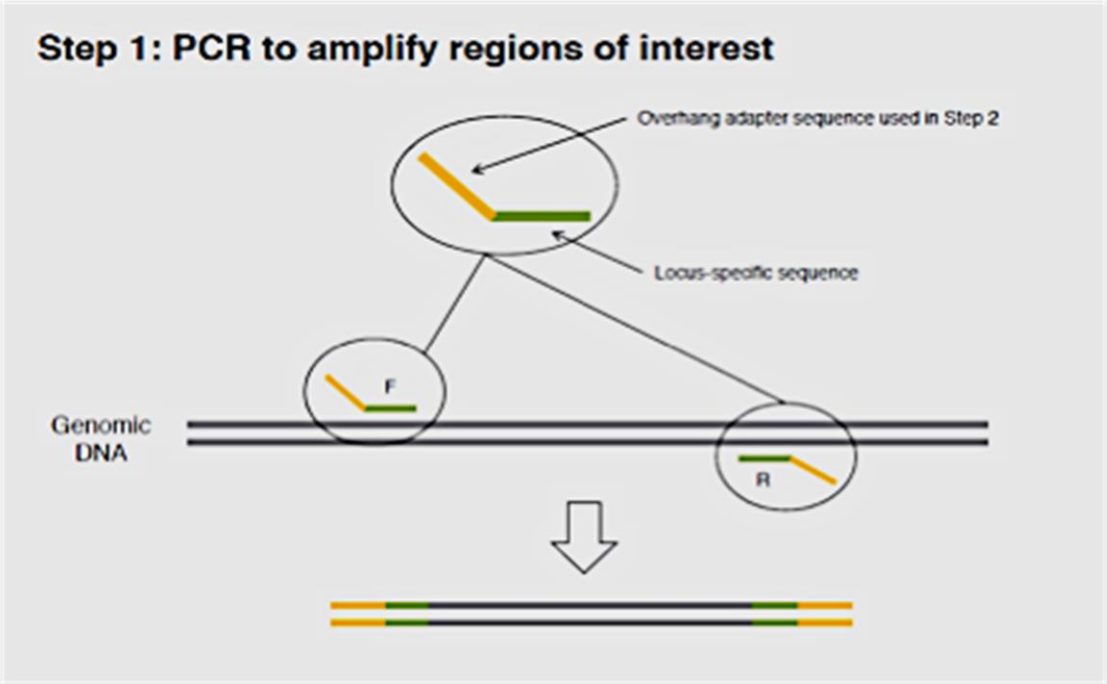
Table 2.3: List of HIV-1 Vif and V3 loop primers designed; primer name, sequence and product size are indicated.

Primer name	Sequence (5'-3')	Expected product size
HIV-1 VIF (519 bp)		
Vif outer forward primer	GTTTATTACAGAGACAGCAG	826 bp
Vif outer reverse primer	ACTCCTGTCCAAGTATCCCC	
Vif full length forward (VifFF)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACACT GGAAAGGTGAAGGGGCAG	708 bp
Vif full length reverse (VifFR)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AGGAAAGTGTCTGACAGCTTC	
Vif Amplicon 1 foward (VifA1F)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACACT GGAAAGGTGAAGGGGCAG	328 bp
Vif Amplicon 1 reverse (VifA1R)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG CTATGGAGACTCCATGACCC	
Vif Amplicon 2 foward (VifA2F)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGA GTTCAGAAGTACACATCCC	430bp
Vif Amplicon 2 reverse (VifA2R)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG AGGAAAGTGTCTGACAGCTTC	
HIV-1 V3 loop (105 bp)		
Envelope V3 1st forward (Env-1F)	CAGCACAGTACAATGCACACATGGAAT	876bp
Envelope V3 1st reverse (Env-1R)	TGACGCCGCGCCCATAGTGCTTCCTGCTG	
Envelope V3 2nd forward (EnVCF2)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGA TAATGATTAGATCTGAAAA	440bp
Envelope V3 2nd reverse (EnvV4R)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG TGATGTATTGCAATAGAAAAA	

#Note

The red and green coloured oligonucleotides are the Illumina transposase adapter sequences.

2.3.3.2.3 First Polymerase Chain Reaction steps (PCR) to amplify regions of interest

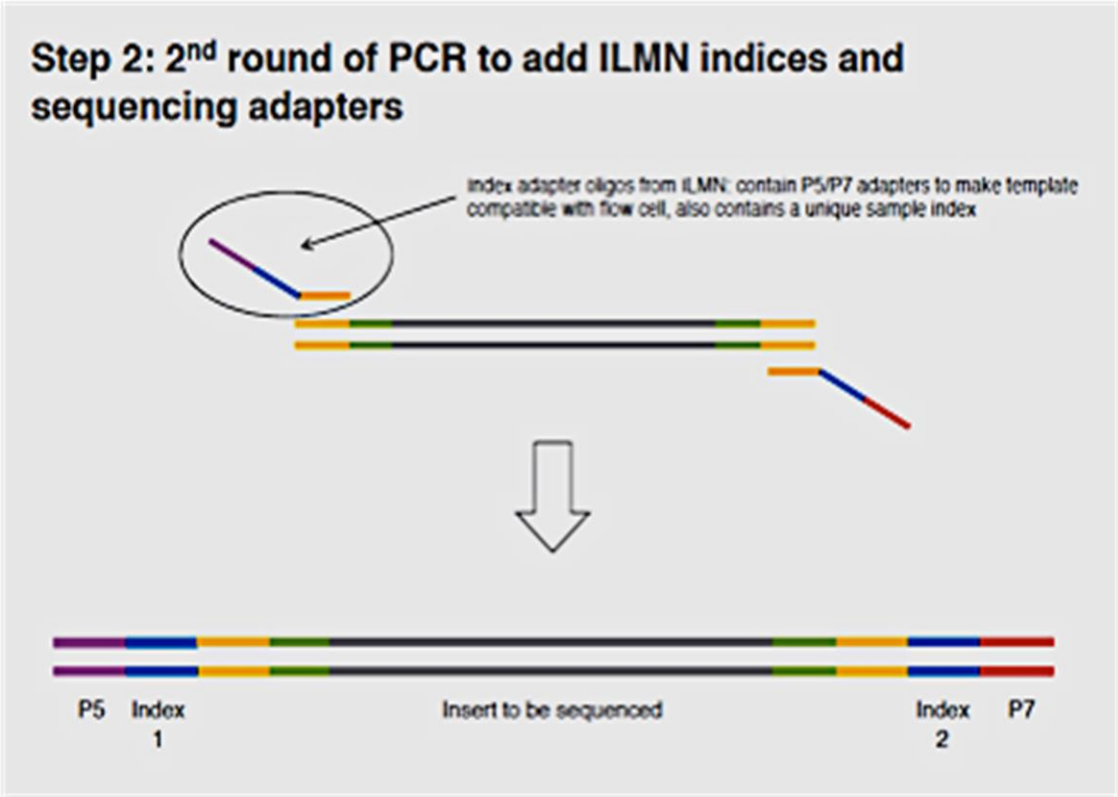


The HIV-1 *vif* and C2-V3 region of the envelope gp120 gene containing the V3 loop were amplified using in a nested PCR using standard PCR reagents and the primers shown in table 2.3 to generate two overlapping amplicons for *vif* and one amplicon for V3 loop. The first round (in a 20µl volume) and nested (in a 40µl volume) reaction protocols were the same for both Vif and V3 loop amplification (table 2.4). The following cycling conditions were used for both first round and nested PCR; Initial denaturation at 94°C for 2 min, followed by a 39 cycles of denaturation at 94°C for 20sec, annealing at 53°C for 45 sec, extension at 72°C for 1 min, final extension at 72°C for 5 min and storage at 4°C.

Table 2.4: PCR amplification protocol for HIV-1 *vif* and V3 loop

Reagent	**Stock concentration**	**Final concentration in 20µl**	**Volume µl /rxn**
RNase/DNA free water	-	-	13.7
10X buffer (supplied)	-	-	2.0
MgS04 (supplied)	50mM	2mM	0.8
dNTPs	10mM	0.2mM	0.4
Taq (Platinum HF)		0.025 units/ul	0.1
Forward primer	20 µM	0.2 µM	1
Reverse primer	20 µM	0.2 µM	1
Total			19
DNA template			1

2.3.3.2.4 Second-round PCR to add Illumina indices and sequencing adaptors



The PCR products were purified using AMpure beads and used as template in a second-round PCR to add Illumina indices and sequencing adaptors, which hybridize to adaptors on the flow cell. To add unique Illumina dual-index barcodes Index 1 (i7) and index 2 (i5), a 50µl reaction was prepared with 25µl Kapa HiFi HotStart mix, 5µl of each index, i7 and i5, 5µl of PCR reaction and 10µl of water. The reaction was mixed and placed in a thermal cycler for a 12 cycle PCR program (table 2.2)

2.3.3.2.5 Library purification, size verification, normalization and pooling

The libraries were cleaned up using AMpure beads, and size-verified using 1% agarose gels prepared by adding 1.5g of agarose into 150ml 1X TAE buffer and microwaved for 2

minutes. Before the gel-solidified 10ul of ethidium bromide was added, the solution was poured into a gel-casting tray and 3ul of a molecular weight ladder mixed with a loading dye was loaded into one well. Five microliters of the DNA samples was loaded into the rest of the wells**.** The libraries were quantified using Qubit 3 with the ds DNA HS kit (Invitrogen), normalized to 4nM and pooled. The libraries were denatured as above and included in the same run as above. For good cluster generation, about 1.8pM of library spiked with 25-30% PhiX (Illumina) was loaded into the sequencing cartridge. Biological sample sheets were created in Basespace by giving each sample the appropriate index and setting up a sequencing run for either a MiSeq or MiniSeq.

2.3.3.3 Sequencing by synthesis

Sequencing by synthesis (SBS) also referred to as bridge amplification or reversible terminator method, is an Illumina sequencing technology where four fluorescently labeled nucleotides are used to sequence the tens of millions of cluster on the flow cell surface in parallel as shown in figure 2.7.

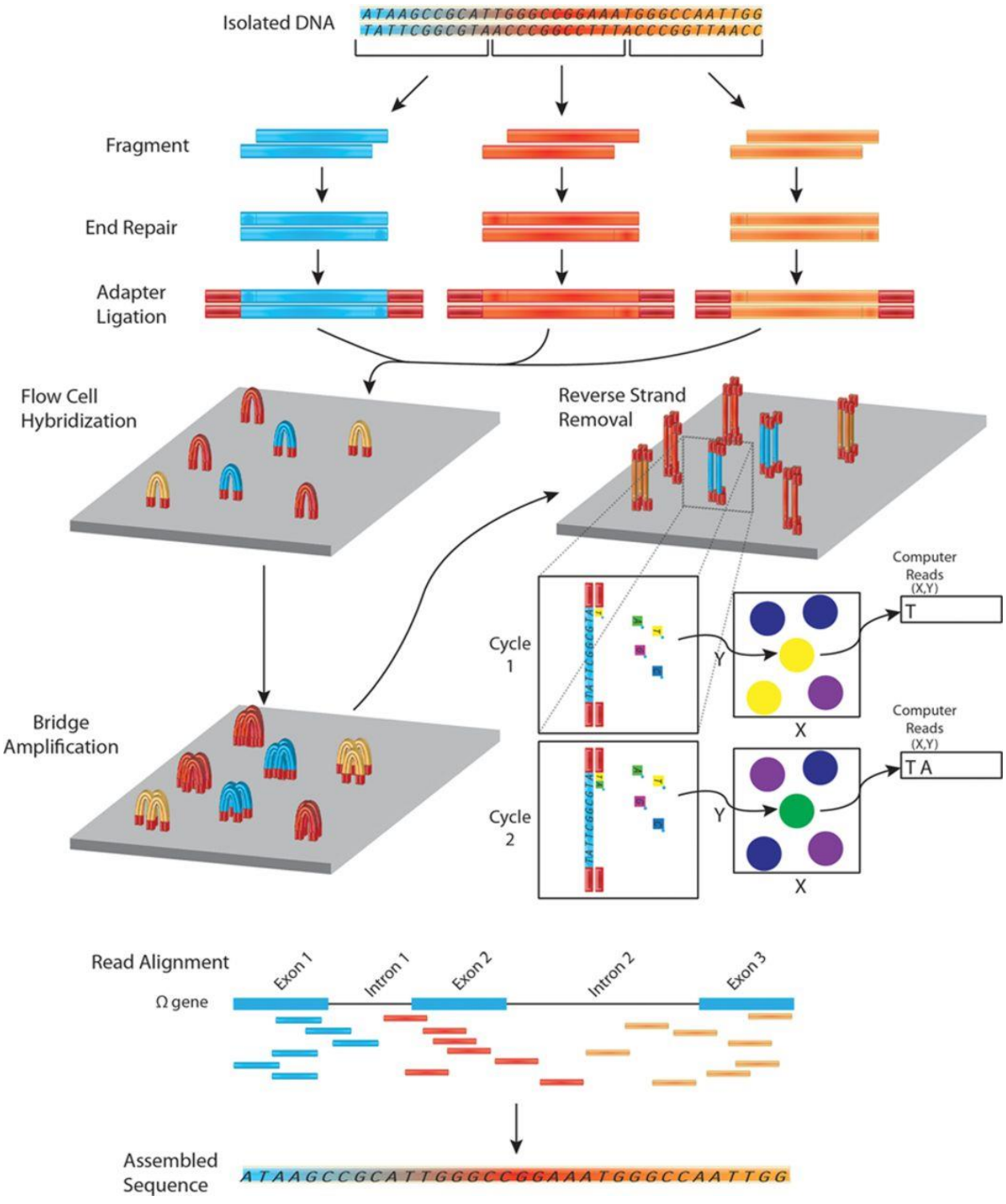


Figure 2.7: Schematic representation of the overall Illumina SBS sequencing workflow (Jared et al., 2013).

2.4 RESULTS

Generally a MiSeq run for 2x 300bp paired end reads using a V3 reagent kit is expected to generates about 23.2 to 15 Gb data, with 44-50 million reads passing filter, 1200-1400 kmm2 cluster passing filter giving a QC30 of >70%. The profile of our sequencing run was analyzed to document the number of cycles, data yield, error rate and quality score. The run was successful and generated about 10.7 Gb of data, error rate of 1.4%, 516/600 cycles and QC30 of 78.9%.

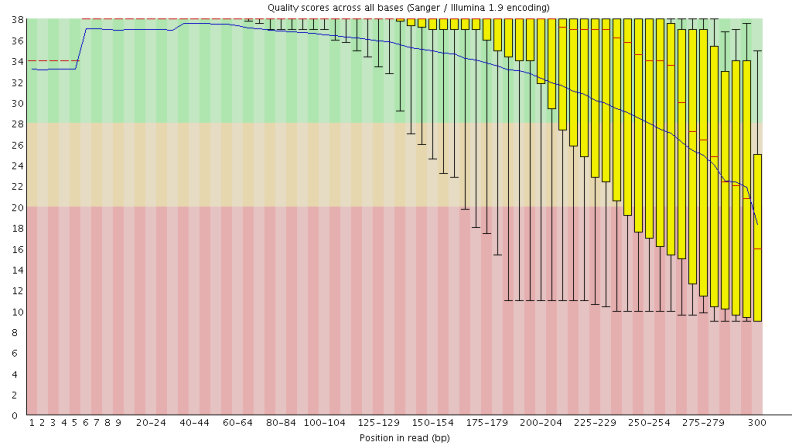
These constitute the raw data of the sequencing experiment. After each run the instrument would have automatically demultiplexed the samples so that each FASTQ files represents an individual experiment. Several runs were conducted in this study, but the data reported below was from a randomly selected run, which had libraries prepared using the Tn5, custom amplicon based as well the Nextera XT used as control to validate the developed protocols. Before analyzing the data (FASTQ), raw sequence reads were checked for quality. Some of the key quality parameters were:

1. Quality score of base calling,
2. Number of reads (coverage)
3. Reads length distribution. In addition, reads were also checked for possible sequencing adapter contamination, especially if using small sequencing target.

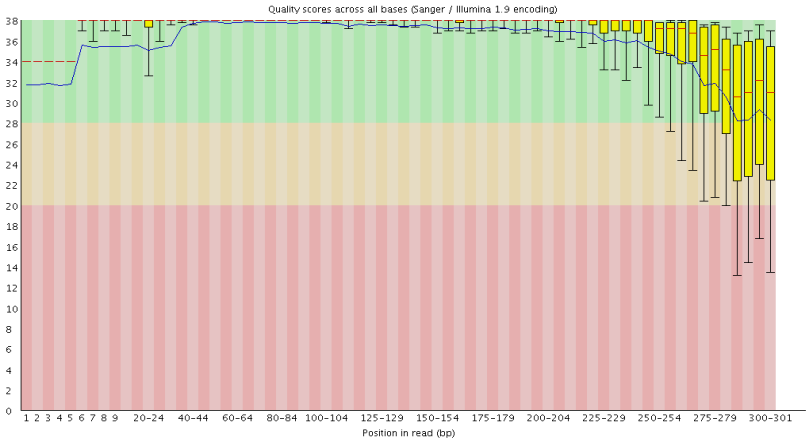
2.4.1 Quality score of basecalling

A quality score (Q-Score) is a prediction of the probability of an error in base calling. The percentage of bases > Q30 is avarage across the entire run and is equivalent to the probability of an incoparating base call 1 in 1000 times (the probability of the correct base call is 99.9%. A Q30 quality score is considered a benchmark for quality in NGS and was achieved in our sequencing run. Quality of reads per sample were also accessed using the FastQC and read 1 (forward reads) across all the methods used to prepare libraries had good quality compared to read 2 (reverse reads). Read from the Nextera, Tn5 and custom based methods were compared. Reads from Nextera and custom based methods had good quality compare to Tn5 method, where the quality dropped after 150 nucleotides (figure 2.8)

A Tn5



B Custom amplicon-based



C Nextera

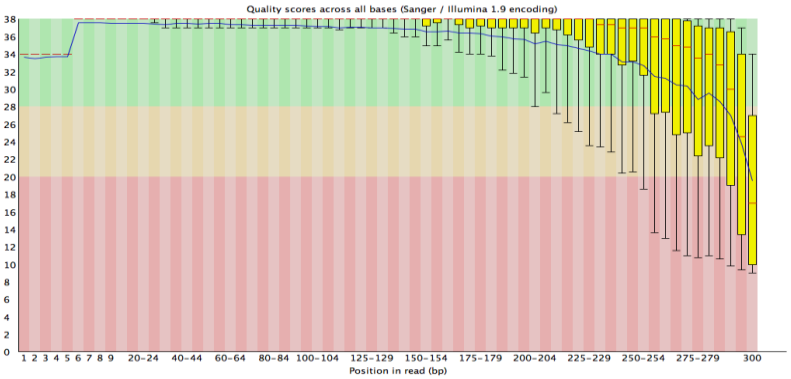
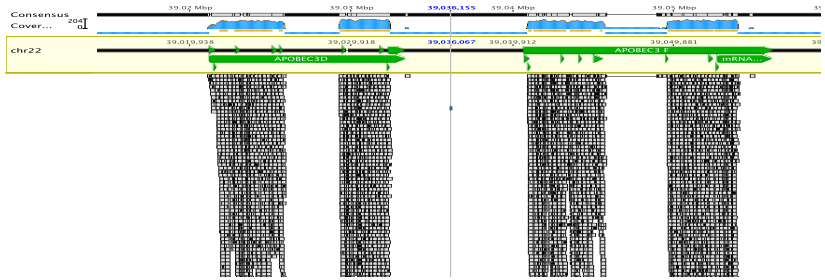


Figure 2.8: Read quality evaluastion across the 3 methods; Tn5 transposase (A), custom amplicon based (B) and Nextera (C). It has been shown that the quality of reads often drops after the 150bp. Libraries prepared with custom based protocols produced high quality reads compared to the Tn5 and Nextera methods.

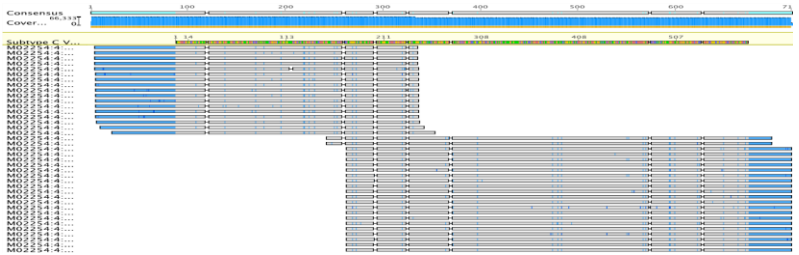
2.4.1 Read coverage

Sequence coverage describes the avarage number of reads that align to, or cover known reference bases. The NGS coverage level often detemines whether variant discovery can be made with a certain degree of confidence at a particular base position. Coverage from 10x to 30x depth is recommended for detecting mutations and SNPs. Good read coverage was achieved with all the methods used to prepare libraries (figure 2.9).

A Tn5



B Custom amplicon-based



C Nextera

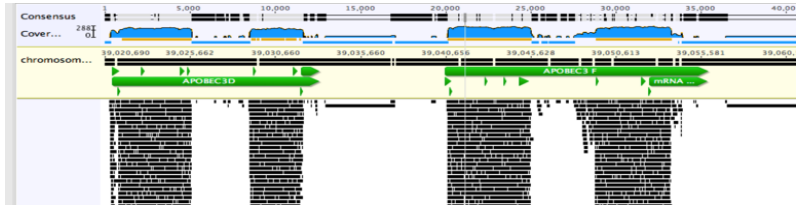


Figure 2.9: Read coverage from libraries prepared from Tn5 transposase (A), custom amplicon based (B) and Nextera (C). Image A and C are showing APOBEC3D and F reads mapped to chromosome 22 with coverage indicated in blue and reads in black. I mage B shows HIV-1 Vif reads mapped overlapping to subtype C Vif reference sequence with average indicated in blue at the top and reads in gray. All the methods had good read coverage.

2.4.3 Read length distribution

A maximum of 300pb read length was achieved with the Nextera and the Tn5 methods, whereas with the custom based methods 350 and 470 bp read lengths were achieved (figure 2.10).

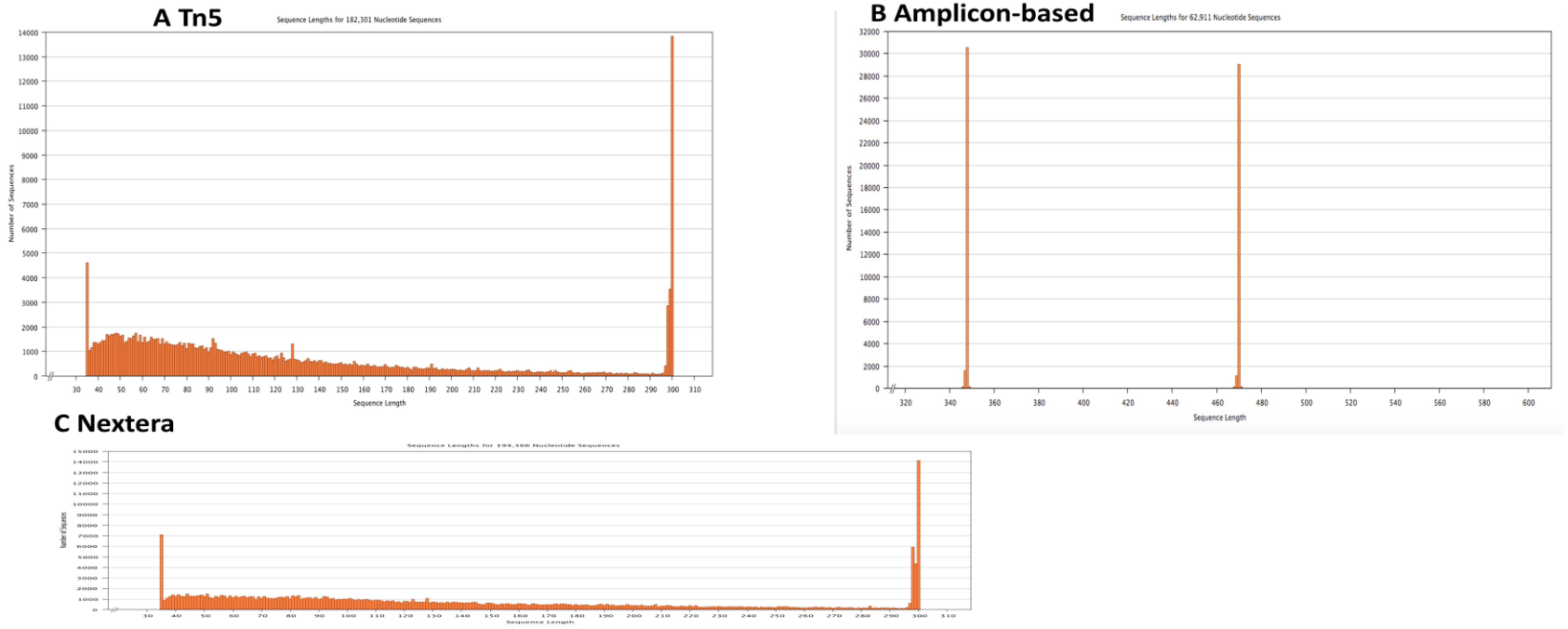


Figure 2.10: Read length distribution from a sample of libraries prepared from homemade Tn5 transposase (A), custom amplicon based (B) and Nextera (C). Both the Tn5 and Nextera libraries produced a minimum of 35 to maximum 300bp reads with a high number of maximum reads. The amplicon based methods had maximum of 328 and 430bp as expected from the two overlapping fragment amplified.

2.5 DISCUSSION AND CONCLUSION

The recent development and advancement of next generation sequencing (NGS) technologies have made it possible to rapidly analyze genomic DNA and viral genomes from many individuals. In this study library preparation protocols to amplify and sequence host cell and HIV-1 genes were developed and validated.

Efficient preparation of high-quality sequencing libraries that will represent the biological sample is key in research when NGS is applied. The availability of custom Tn5 transposase protein (Caruccio *et al.* 2009; Picelli *et al.* 2014; Hennig *et al.* 2018) allows innovation in protocol design and application. This study describes development of several derivative protocols for transposase catalyzed fragmentation and sequencing of the complete coding exon content of host cell gene libraries derived using this method. We have demonstrated reliability and high quality of libraries prepared by this method by comparing with libraries prepared using commercially available kits. This protocol was demonstrated to be suitable for various experiments, and that it could be used to sequence multiplexed or a pool of up to five host genes. We were also able to demonstrate that host cell genes and HIV-1 genes can be pooled and sequenced simultaneously producing high quality reads.

In this study we also demonstrate an alternative amplicon based library preparation protocol which uses a two step PCR process where the region of interest is amplified with a pair of adapter tailed primers in a primary PCR reaction followed by addition of sample specific dual indices and flow cell adapter in a limited number of PCR cycles. This protocol is primer dependent and was applied mostly to prepare libraries for the HIV-1 genes due to their shorter length. We have demonstrated that this approach can also be used to prepare libraries for longer regions by designing primers to amplify overlapping regions of the gene of interest. This approach presented best quality reads when compared to the Tn5 and commercial Nextera kit.

In conclusion, the protocols developed in this study can be applied to amplify and sequence any host and HIV-1 genes of interest allowing much deeper and more sensitive profiling of host genes and HIV-1 genetic diversity at a lower cost.

2.6 REFERENCES

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., Shendure, J. (2010) 'Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition', *Genome Biology*, 11(12).

Caruccio, N., Grunenwald, H., Syed, F., Biotechnologies, E. (2009) 'Nextera TM technology for NGS DNA library preparation: simultaneous fragmentation and tagging by in vitro transposition', *Nature Methods*, 16(October), 3.

Eberwine, J., Sul, J.Y., Bartfai, T., Kim, J. (2014) 'The promise of single-cell sequencing', *Nature Methods*.

Hashimshony, T., Wagner, F., Sher, N., Yanai, I. (2012) 'CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification', *Cell Reports*, 2(3), 666–673.

Head, S.R., Kiyomi Komori, H., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhanian, P. (2014) 'Library construction for next-generation sequencing: Overviews and challenges', *BioTechniques*, 56(2), 61–77.

Hennig, B.P., Velten, L., Racke, I., Tu, C.S., Thoms, M., Rybin, V., Besir, H., Remans, K., Steinmetz, L.M. (2018) 'Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol.', *G3 (Bethesda, Md.)*, 8(1), 79–89.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S. (2014) 'Quantitative single-cell RNA-seq with unique molecular identifiers', *Nature Methods*, 11(2), 163–166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., Amit, I. (2014) 'Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types', *Science*, 343(6172), 776–779.

Lander, E.S. (2011) 'Initial impact of the sequencing of the human genome', *Nature*.

Metzker, M.L. (2010) 'Sequencing technologies the next generation', *Nature Reviews Genetics*.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R. (2014)

'Full-length RNA-seq from single cells using Smart-seq2', *Nature Protocols*, 9(1), 171–181.

Sandberg, R. (2014) 'Entering the era of single-cell transcriptomics in biology and medicine', *Nature Methods*.

CHAPTER THREE

Next Generation Sequencing Reveals a High Frequency of CXCR4 Utilizing Viruses in HIV-1 Chronically Infected Drug Experienced South African Individuals

---

Nontokozo D. Matume, Denis M. Tebit, Laurie R. Gray, Marie-Louise Hammarskjold, David Rekosh, Pascal O. Bessong. Next Generation Sequencing reveals a high frequency of CXCR4 utilizing viruses in HIV-1 chronically infected drug experienced South African individuals. **Journal of Clinical Virology,** https://doi.org/10. 1016/ j. jcv. 2018. 02. 008. **(Appendix I)**

ABSTRACT

Background: Entry inhibitors, such as Maraviroc, bind to the CCR5 co-receptor inhibiting entry of CCR5 utilizing viruses (R5 viruses). In the course of HIV infection, CXCR4 utilizing viruses (X4 viruses) may emerge and outgrow R5 viruses, and potentially limit the effectiveness of Maraviroc. The use of Maraviroc in South Africa is reserved for salvage therapy.

Objective: In this study, we examined the predicted frequency of R5 and X4 viruses based on HIV Env V3 loop sequences, using next generation sequencing, in patients under treatment to draw inferences for the utility of Maraviroc in a South African population.

Methods: Proviral DNA was isolated from peripheral blood mononuclear cells (PBMC) of 72 chronically HIV infected patients on antiretroviral treatment. HIV V3 loop gene was amplified and sequenced on an Illumina MiniSeq platform. Viral subtypes were determined using the jumping profile Hidden Markov Model (jpHMM) genotyping tool. De Novo consensus sequences were derived for the majority and minority populations for each patient using Geneious® software version 8.1.5. HIV-1 tropism was inferred using PSSMsinsi, Geno2pheno and Phenoseq-C web-based tools.

Results: Quality V3 loop sequences were obtained from 72 patients, with 5 years (range: 0-16) median duration on treatment. Subtypes A1, B and C viruses were identified at frequencies of 4% (3/72), 4% (3/72) and 92% (66/72) respectively. Fifty four percent (39/72) of patients were predicted to exclusively harbor R5 viral quasispecies; and 21% (15/72) to exclusively harbor X4 viral quasispecies. Twenty five percent of patients (18/72) were predicted to harbor a dual/mixture of R5X4 quasispecies. Of these 18 patients, about 28% (5/18) were predicted to harbor the R5+X4, a mixture with a majority R5 and minority X4 viruses, while about 72% (13/18) were predicted to harbor the R5X4+ a mixture with a majority X4 and minority R5 viruses. The proportion of all patients who harbored X4 viruses either exclusively or dual/mixture was 46% (33/72). Thirty-five percent (23/66) of the patients who were of HIV-1 subtype C were predicted to harbor X4 viruses (χ^2=3.58; p=0.058), and 57% of these (13/23) were predicted to harbor X4 viruses exclusively. $CD4^+$ cell count less than 350 cell/µl was associated with the presence of X4 viruses (χ^2=4.99; p=0.008).

Conclusion: The effectiveness of Maraviroc as a component in salvage therapy may be compromised for a significant number of chronically infected patients harboring CXCR4

utilizing viruses.

Keywords: Next generation sequencing; Chronic HIV infection; Co-receptor usage; Maraviroc; South Africa

3.1 INTRODUCTION

The entry of human immunodeficiency virus type 1 (HIV-1) into target cells is mediated by the virus' envelope glycoproteins. Specifically, gp120 interacts with the CD4 receptor and co-receptors, either CCR5 or CXCR4 (Doms and Peiper 1997). Viruses with the ability to use CCR5 are classified as R5 viruses, those that use CXCR4 are classified as X4 viruses, while those that use both co-receptors are referred to as dual tropic (R5X4 viruses) (Berger *et al.* 1998) R5X4 viruses are further classified as dual-R (R5+X4, a mixture of R5 and X4 viruses with R5 existing as majority and X4 as minority) or dual-X (R5X4+ a mixture of R5 and X4 viruses with X4 existing as majority and R5 as minority) (Yi *et al.* 1999; Huang *et al.* 2007; Irlbeck *et al.* 2008).

Initial HIV-1 infections involve the use of CCR5, while the use of CXCR4 often occurs at late stages of infection. This has been strongly associated with low $CD4^+$ cell counts and eventually occurs in approximately 50% of subtype B infections (Tersmette *et al.* 1989; Richman and Bozzette 1994; Connor *et al.* 1997; Hatse *et al.* 2003). In addition to subtype B, the co-receptor switch from R5 to X4 has also been reported in subtypes A, C, D, CRF01_AE, and CRF02_AG infections (Björndal *et al.* 1997; van Rij *et al.* 2002; Esbjörnsson *et al.* 2010)

HIV-1 subtype C accounts for over half of HIV-1 infections worldwide and it is the predominant variant circulating in Southern Africa. Studies have demonstrated an overwhelming predominance of R5 viruses at various stages of HIV-1 subtype C infection, although X4 viruses have also been reported (Björndal *et al.* 1997; Ping *et al.* 1999; Cecilia *et al.* 2000; Engelbrecht *et al.* 2001; Papathanasopoulos *et al.* 2002; Cilliers *et al.* 2003; Choge *et al.* 2006). In vitro studies have also indicated that about 30% of HIV-1 subtype C viruses in individuals with advanced disease proficiently use CXCR4 (Connell *et al.* 2008; Raymond *et al.* 2010).

Maraviroc is an entry inhibitor that binds to CCR5 and inhibits gp120-CCR5 interaction through an allosteric mechanism (Garcia-Perez *et al.* 2011). The United States Food and Drug Administration first approved Maraviroc in 2007 for salvage therapy. Currently, it is also recommended as a component in first-line antiretroviral therapy (ART) in the United States and other developed countries (Krauskopf, 2007; Reuters, 2007; Woollard, 2015), and is also being considered in HIV prophylaxis (Gulick *et al.* 2008, 2017; Gilliam *et al.* 2010; R. *et*

al. 2016).

In settings such as South Africa, genotypic drug resistance testing is recommended following treatment failure with the first and second line regimens. Additionally, adherence and drug interactions are reviewed to guide drug change in drug class. In the case where a third line treatment is required, the decision is reserved for an infectious disease specialist, who may recommend Maraviroc as salvage therapy (Meintjes *et al.* 2014). However, testing for sensitivity to Maraviroc is not routinely done. Phenotypic and genotypic approaches are used in deciphering HIV co-receptor usage. Monogram Trofile, is a commercial phenotypic based assay that requires cell culture procedures and is time and labor intensive (Schuitemaker *et al.* 2010). On the contrary, genotypic based algorithms for co-receptor prediction are less time consuming and require only use of web-based algorithms such as Subtype C position-specific scoring matrix (PSSMsinsi), Geno2Pheno and Phenoseq-C (Riemenschneider *et al.* 2016).

3.2 Hypothesis

HIV-1 subtype C chronically infected South Africans harbor a significant level of CXCR4 utilizing viruses.

3.3 Objective

The objective of this study was to use next generation sequencing to generate HIV-1 V3 loop sequences from treatment-experienced individuals; and analyze the quasispecies for prediction of co-receptor usage; in an effort to draw inferences for the subsequent utility of Maraviroc as salvage therapy in South Africa.

3.3.1 Specific objectives

3.3.1.1 To sequence HIV-1 V3 loop using NGS

3.3.1.2 To analyze the quasispecies for predicted co-receptor usage; in an effort to draw inferences for the subsequent utility of Maraviroc as salvage therapy in South Africa

3.3.1.3 To determine V3 loop amino acid sequence diversity

3.4. MATERIALS AND METHODS

3.4.1 Study population and sample collection

Characteristics of the study population and sample collection procedures are as described in chapter two, section 2.3.2.

3.4.2 DNA extraction and Polymerase Chain Reaction (PCR) amplification of V3 loop

Protocols to achieve specific objective 3.2.1.1 are fully described in chapter two, section 2.3.3.2.3. Briefly, DNA was extracted from PBMC of 192 HIV infected individuals under antiretroviral therapy, using Qiagen blood mini kit following the manufacturer's instructions. The C2-V3 region of the envelope gp120 was amplified in a nested PCR using standard PCR reagents and cycling conditions detailed in chapter 2.

3.4.3 Library preparation and MiniSeq sequencing

The HIV V3 loop libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina San Diego, California, USA) following the manufacture's instructions with 20% Phix as control (Illumina San Diego, California, USA). Agarose electrophoresis (E-gels) was used to verify the size of the libraries, and normalized to 4nM each to ensure equal library representation in the pool. The prepared libraries were sequenced in an Illumina MiniSeq using a high output kit for 300 cycles (Illumina San Diego, California, USA) and generating paired-end 2x150bp long sequence reads for each.

3.4.4 De-multiplexing, sequence quality control evaluation and mapping

Sequences were de-multiplexed automatically on the MiniSeq as part of the data processing steps and ends pairing. Fastq files were generated for each sample representing the two paired-end reads. Sequence quality was validated using the FastQC programme. The fastq files were imported into Geneious® software version 8.1.5 for filtering and trimming of sequences as well as downstream analysis. All sequences were trimmed at the ends as part of the assembly process using the modified-Mott algorithm and quality scores assigned by the sequencing base caller. The two sequence lists from each sample were paired and mapped to HIV-1 subtype C envelope consensus sequence (Gene-Bank accession number: DQ275658) to identify Indels and mapped to HXB2 for subtyping confirmation. Figure 3.1

represents the bioinformatics processes employed for viral subtype and co-receptor determination.

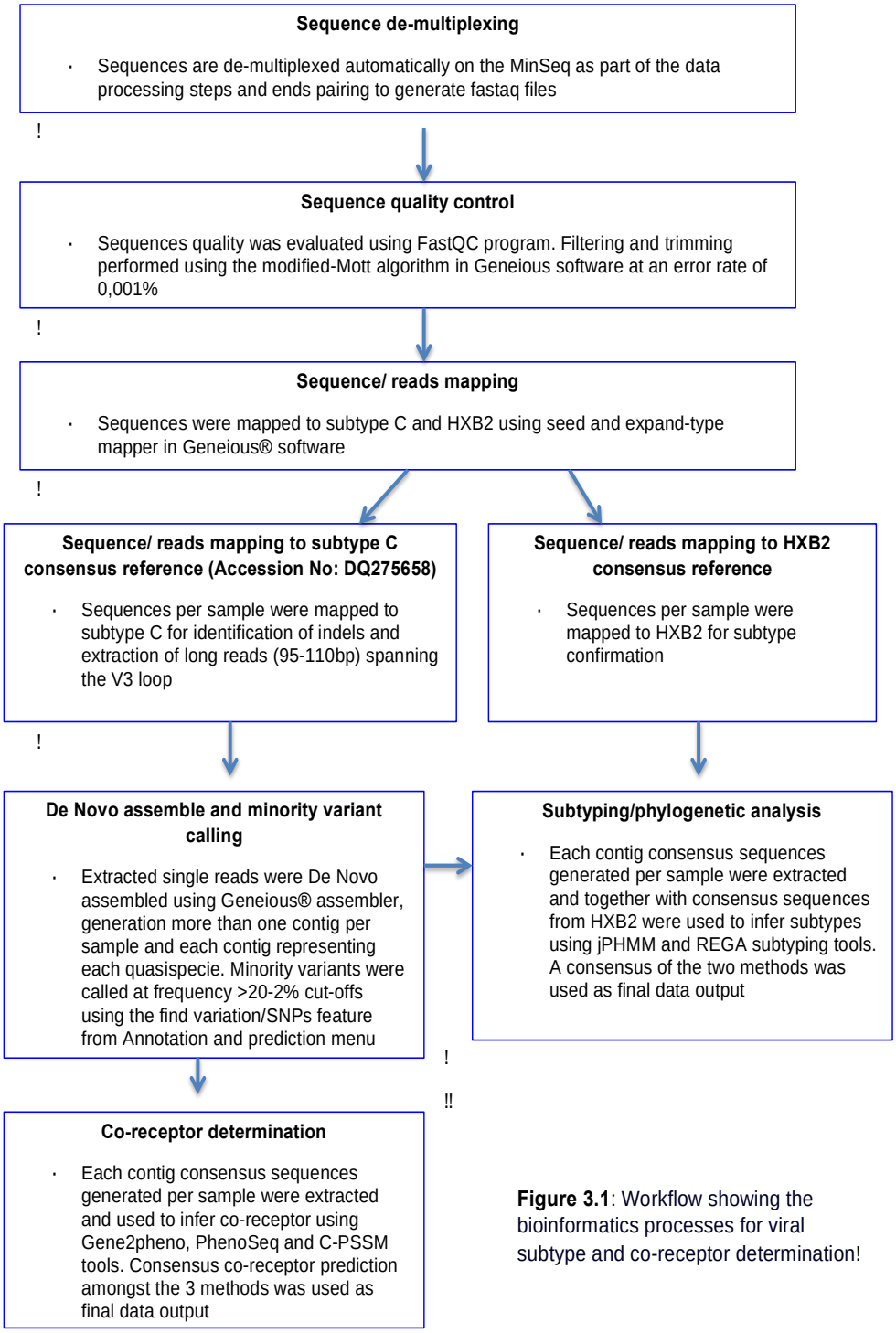


Figure 3.1: Workflow showing the bioinformatics processes for viral subtype and co-receptor determination!

3.4.5 Intra-patient quasispecies and co-receptor prediction

To achieve specific objective 3.2.1.2, all reads spanning the HIV envelope V3 region which were mapped to the HIV-1 subtype C envelope consensus were extracted as single reads and De Novo assembled using Geneious® assembler and generating contigs for each quasispecies in each sample. Minority variants were called based on the variant read depth, variant frequency and consensus base frequency with >20-2% cut-offs using the find variation/SNPs feature from the Annotation and prediction menu in Geneious® software version 8.1.5. This parameter also calculates the p-value (maximum variant p-values set to 10^{-6}; 0.0001% to see variant by chance) for the specified variant frequency call, with lower p-values representing a real variant at the given position (http://www.geneious.com). All consensus contigs generated were aligned using Geneious® multiple alignments and translated into amino acids. HIV subtype was inferred with maximum likelihood tree and confirmed with Geno2pheno co-receptor determination tool, the jumping profile Hidden Markov Model (jpHMM) and REGA subtyping tools. Co-receptor tropism was inferred using PSSMsinsi (Jensen *et al.* 2006). Geno2pheno [co-receptor] (10% false-positive rate) (Thielen and Lengauer 2012) and PhenoSeq-C (Cashin *et al.* 2015) web based tools and a consensus co-receptor result amongst the 3 tools were used as final output data. Association of predicted co-receptor usage and $CD4^+$ cell count; viral load and number of years on treatment were determined using Chi-square computation.

3.4.6 V3 loop amino acid sequence analysis

To achieve specific objective 3.2.1.3, the envelope V3 loop amino acid sequence variability at each position of R5 and X4 viruses was determined using the Los Alamos HIV Sequence Database Entropy- one tool (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html) and WebLogo (http://weblogo.threeplusone.com/). The N-glycosylation sites were determent using the Los Alamos HIV sequence Database N-GlycoSite (http://www.hiv.lanl.gov/content/sequence/GLYCOSITE/glycosite.html)

3.5. RESULTS

3.5.1 Patients' characteristics

Patients' clinical data including viral loads, $CD4^+$ cell counts, number of years on treatment, age and treatment regimen are shown in table 3.1. The median viral load log10 copies/mL was 3.9 (range, 0.7-5.4); the median $CD4^+$ cell/ mm^3 was 256 (range, 5-1022); median age was 38, (range, 4-72) median duration on treatment was 5 years (range, 0-16). There were 43 patients on first-line antiretroviral therapy, 15 on second-line regimens, 2 have not started treatment, 1 stopped treatment and 11 patients with unknown treatment regimen.

Table 3.1: Characteristics of individuals predicted to harbor R5 or X4 monotropic and those that were predicted to harbor dual/mixed R5+X4 and R5X4+ viruses.

		Biotype			
Parameter	**Total**	**X4**	**R5**	**R5+X4**	**R5X4+**
Number of Patients	72	15 (20.8%)	39 (54.2%)	5 (6.9%)	13 (18.1%)
Median viral load Log10 copies/mL (Range)	3.9 (0.7-5.4)	4.1 (0.7-5.2)	3.8 (1- 6.3)	3.6 (1.5-4.6)	4.4 (2.5-5.4)
Median CD4 counts cells/mm^3 (Range)	256 (5-1022)	228.5 (13-381)	347 (43-1022)	64 (41-483)	228.5 (5-362)
Median age (Range)	38 (4-72)	36 (8-60	38 (4-72)	34 (4-55)	39 (13-60)
Median years on treatment (Range)	5 (0-16)	5 (1-8)	6 (0-16)	5 (2-6)	5 (1-11)
Number of patients on first-line Antiretroviral therapy *	43 (59.7%)	10 (23.3%)	22 (51.2%)	3 (7%)	8 (18.5%)
Number of patients on Second-line Antiretroviral therapy #	15 (20.8%)	4 (26.6%)	7 (46.6%)	-	4 (26.6%)
Unknown treatment regimen	11 (15.2)	2 (18.1%)	6 (54.5%)	2 (18.1%)	1 (9)
Not started treatment	2 (2.7)	-	2 (100)	-	-
Stopped treatment	1 (1.4)	-	-	-	1 (100)

* First-line regimen includes; TDF+FTC+EFV, ABC+3TC+EFV, D4T+3TC+EFV, TDF+3TC+NVP, TDF+TFC+EFV, AZT+3TC+NVP, AZT+DDI+EFV, D4T+3TC+EFV.

Second-line regimen; FTC+TDF+LPV/r, ABC+3TC+ LPV/r, AZT+3TC+ LPV/r.

Note: R5+X4 refers to the quasispecies of R5 and X4 viruses per patient with majority R5 and minority X4 viruses; and R5X4+ refers to the quasispecies with moajority X4 and minority R5 viruses. R5 and X4 refer to viruses that exclusively use CCR5 or CXCR4 respectively.

3.5.2 Frequency and distribution of predicted X4 and R5 viruses

Quality V3 loop sequences were obtained for 72 individuals. Subtypes A1, B and C viruses were identified at frequencies of 4% (3/72), 4% (3/72) and 92% (66/72) respectively. Table 3.2 shows 14 viruses whose V3 loops gave discordant subtype inferences.

Table 3.2: Comparison of subtype prediction with Geno2pheno co-receptor determination tool, the jumping profile Hidden Markov Model (jpHMM) and REGA subtyping tools.

Samples	jpHMM	REGA	Geno2pheno	Phylogenetic tree	Biotype
LP1111	C	C	A/AG	D	$R5^{+}X4$
LP1112	C	C	A/AG	D	X4
DF10	C	C	A/AG	D	$X4^{+}$ R5
DF48	D	C	A/AG	D	X4
DF22	A1	A1	B	D	R5
DF84	C	C	A/AG	D	$X4^{+}$ R5
DF30	A1	A1	B	J	X4
DF33	C	C/N	D	O/N	X4
DF58	C	C	A/AG	C	X4
DF09	C	C	A/AG	C	X4
BB26	B	B	B	B	X4
DF81	C	B	B	B	$R5^{+}X4$
DF56	B	B	B	B	R5
DF85	A1	A1	A/AG	J	$X4^{+}$ R5

Note: These are samples that were non-C with phylogenetic analysis and had to be predicted using other tools. Output data was from a consensus of two tools.

Fifty four percent (39/72) of patients were predicted to exclusively harbor R5 viral quasispecies; and 21% (15/72) predicted to exclusively harbor X4 quasispecies. Twenty five percent of patients (18/72) were predicted to harbor R5X4 dual/mixture of R5 and X4 quasispecies. Of these 18 patients, 28% (5/18) were predicted to harbor the R5+X4, a mixture with majority R5 and minority X4 viruses while 72% (13/18) were predicted to harbor the R5X4+ mixture with majority X4 and minority R5 viruses. The proportion of all patients who were predicted to harbor X4 viruses either exclusively or dual/mixture was 46% (33/72) shown in table 3.3. Thirty-five percent (23/66) of the patients who were of HIV-1 subtype C were predicted to harbor X4 viruses (χ^2=3.58; p=0.05), and 57% of these (13/23) predicted to harbor X4 viruses exclusively. The X4 viruses were found to be associated with low $CD4^+$ cell counts (X^2= 6.11, p= 0.0134), and no association was found with the duration on treatment (X^2= 0.4213, p= 0.81006). The three algorithms used were compared to each other to validate the most likely co-receptor prediction. PhenoSeq-C was 76% in concordance with Geno2Pheno and 83% with PSSMsinsi. Geno2Pheno was 81.9% in concordance with C-PSSMsinsi.

Table 3.3: Frequency of predicted R5, X4, R5+X4 and R5X4+ viruses in the study subjects.

Exclusively		**Dual/Mixed R5+X4**		**Dual/Mixed R5X4+**	
R5	**X4**	**R5**	**X4**	**R5**	**X4**
54% **(39/72)**	21% (15/72)	72%(13/18)	28% (5/18)	28% (5/18)	72% (13/18)

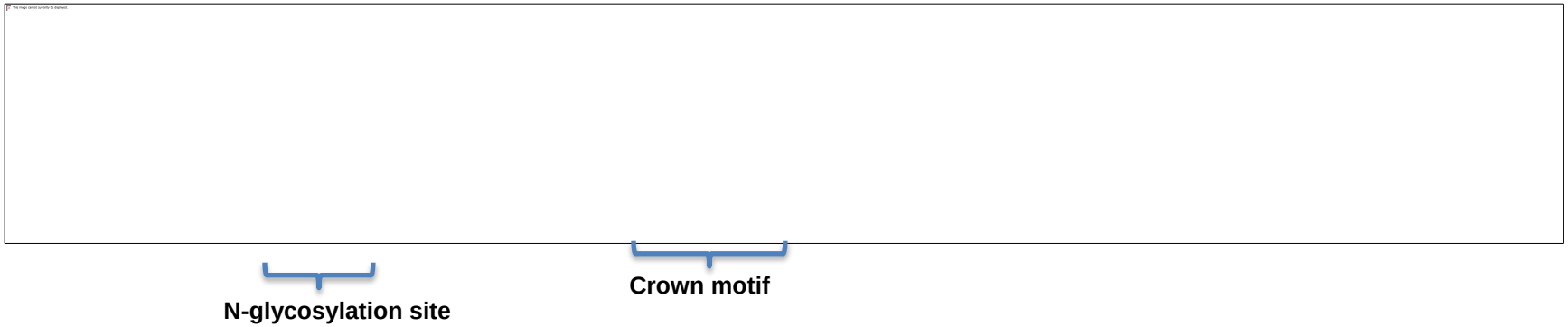
3.5.3 V3 loop amino acid sequence diversity

A higher level of amino acid variation was observed in predicted X4 viruses compared to predicted R5 viruses. Entropy plots were performed to compare the degree of V3 amino acid variability between the R5 and X4 viruses. There were higher entropy levels at 29 amino acid positions in X4 viruses compared to 17 amino acid positions in R5 viruses (X^2= 9.13, p=0.002) (figure 3.2).

The V3 loop sequences of R5 viruses had between 34-35 amino acids with deletions but no insertions; whereas sequences of X4 viruses were 31-37 amino acids long with insertions and deletions (figure 3.3). The consensus X4 sequence had insertions of Valine, Isoleucine or Threonine and Glycine or Arginine at positions 14 and 15 respectively. As an indicator of X4 viruses, a significantly higher average net charge was observed in X4 than R5 viruses (X^2= 35.6, $p<0.05$) (figure 3.3). All the R5 viruses had conserved N-glycosylation sites (NXT) and loss of this site was associated with X4 viruses (X^2= 36.02, $p<0.05$). The GPGQ crown motif remained conserved in R5 viruses; while in X4 viruses there were changes to GPGX (where X was usually an Arginine or Histidine). In addition, X4 consensus sequences had an insertion of Arginine and Glycine within the crown motif between position 19 and 20.

The V3 loop sequences of R5 viruses had between 34-35 amino acids with deletions but no insertions; whereas V3 loop sequences of X4 viruses were 31-37 amino acids long with insertions and deletions (figure 3.3). The consensus X4 sequence had insertions of Valine, Isoleucine or Threonine and an insertion of Glycine or Arginine between positions 14 and 15. X4 viruses had a significantly higher average net charge than R5 viruses (X^2= 35.6, $P<0.05$). All the R5 using viruses had conserved N-glycosylation sites (NXT) and loss of this site was associated with X4 viruses (X^2= 36.02, $p<0.05$). The GPGQ crown motif remained conserved in R5 viruses; while in X4 viruses there were changes to GPGX (where X was usually an Arginine or Histidine). In addition, some X4 viruses had an insertion of Arginine and Glycine within the crown motif.

A



B

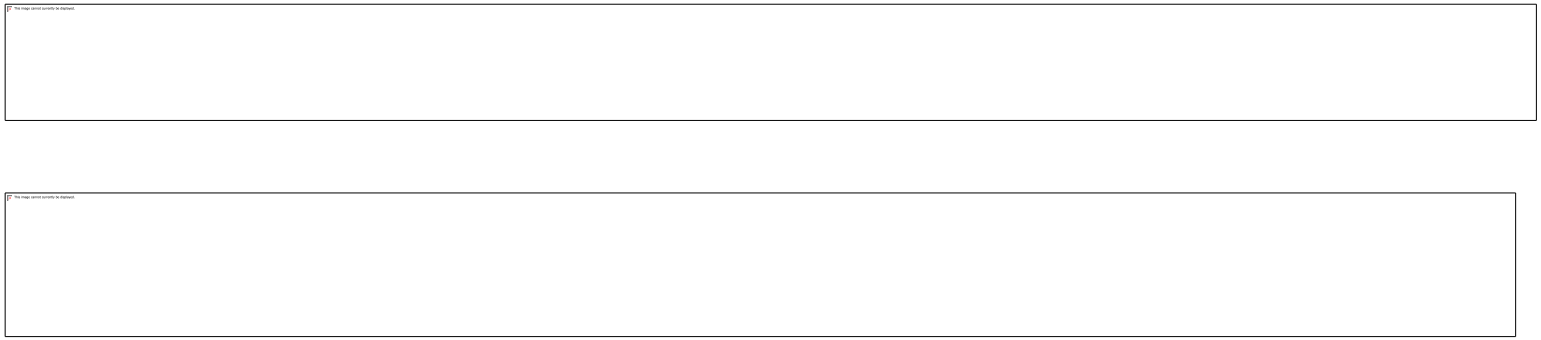
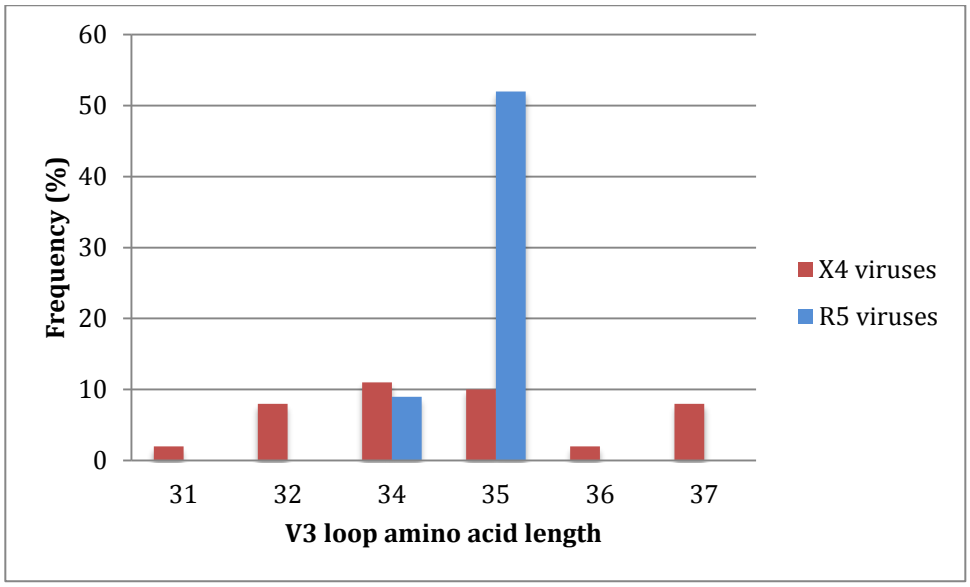


Figure 3.2: Variation within the V3 region of Subtype C R5 and X4 viruses. (2A) Entropy plots representing the variation at each amino acid position site from amino acid position 1 to 37 for R5 and 1 to 39 for X4 viruses. The N-glycosylation site and crown motif are indicated. (2B) Sequence logos of subtype C V3 residues of R5 and X4 consensus sequences. The size of each logo represents the frequency of the amino acid at each site. X4 viruses show more variability than R5. # **Note** (-) Indicates insertion at position 14, 15, 19 and 20

A



B

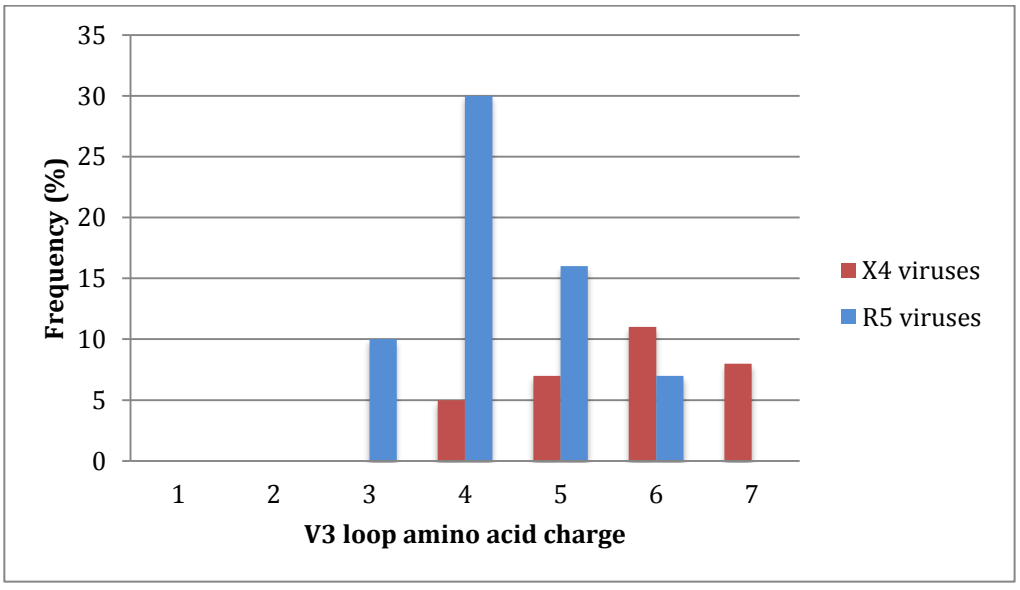


Figure 3.3: Comparison of (A) amino acid length and (B) amino acid net charge of HIV-1 subtype C isolates capable of using CCR5 and CXCR4.

3.6 DISCUSSION AND CONCLUSIONS

Maraviroc binds to CCR5 and prevents its usage by the infecting virus. As a result, X4 viruses will not be sensitive to Maraviroc (Garcia-Perez *et al.* 2011). It has been shown that viruses in primary infection preferably use CCR5 exclusively, but as disease progresses viruses switch or extend co-receptor usage to CXCR4. In this study, we investigated the frequency and distribution of R5 and X4 viruses in chronically infected HIV patients on Maraviroc-free first and second line treatment regimens in South Africa. The aim was to predict Maraviroc utility in the study population, but more importantly in those who are under the second line treatment regimen, and for whom the introduction of a Maraviroc-containing regimen may be the next option, as salvage therapy.

The highest frequency of X4 viruses (30%) thus far reported from South Africa was from treatment-experienced patients in KwaZulu Natal (Connell *et al.* 2008). In the current study, predicted X4 viruses were detected at a higher frequency of 46%. It was interesting to observe that in this study 21% of the patients were predicted to harbor X4 viruses exclusively and have been on treatment for at least one year. Pure X4 viral populations are frequently detected and often restricted to late stages particularly in subtype C (Melby *et al.* 2006). In our study, about 19% of patients predicted to harbor X4 viruses were in the initial stages of the infection (stage 1 or 2). Our observation is similar to that of (Frange *et al.* 2008) in which a relatively high prevalence (15.9%) of predicted X4 viruses was documented during early infection in a retrospective evaluation of a large number of patients enrolled in the PRIMO cohort in France. It is not known whether the X4 strains detected in early stages were transmitted or evidence that co-receptor switching is not restricted to late stages of the infection. In the current study 22% of patients were predicted to harbor an R5+X4 mixture, with X4 viruses as a minority population, which could easily be missed when population based sequencing, is applied and might outgrow the R5 viruses and impact negatively on the use of Maraviroc. Also of note is that 53.3 % (8/15) of those with X4 viruses were on a second line treatment regimen (AZT+3TC+LPV/r or FTC+TDF+LPV/r or TDF+3TC+NVP), and for whom Maraviroc might be the next option. The predicted X4 viruses were statistically associated with a low $CD4^+$ cell count (<350 cells/µl), but not with viral load. Co-receptor switching has been associated with disease progression, treatment, and viral genotype (reviewed in Schuitemaker, Van 't Wout and Lusso, 2010). Overall, subtype C viruses appear to have the least propensity to switch from R5 to X4. However, in the current

study, approximately 35% (23/66) of the patients infected with subtype C viruses were predicted to harbor CXCR4 utilizing viruses and 57% (13/23) of these were X4 exclusively. The mechanism underling the emergence of X4 viruses in treatment-experienced patients is not yet clear; and the question remains whether the identification of co-receptor switching and the emergence of drug resistance mutations are not coincidental events (Singh *et al.* 2011; Ketseoglou *et al.* 2014). On the other hand, there is data showing that the up regulation of CCR5 is associated with HIV infection, and that this is reversed in the course of treatment (Ostrowski *et al.* 1998; Weinberger *et al.* 2009). Treatment may encourage the flourishing of dual tropic viruses, since down regulation of CCR5 can be perceived as an increase in the utility of CXCR4.

Nevertheless, the high frequency of X4 viruses may not be related to drug exposure but an ongoing evolution of HIV-1 subtype C. Studies have reported a high proportion of X4 viruses in treatment naïve individuals (Van Rensburg *et al.* 2002; Cilliers *et al.* 2003). Continuous studies of HIV-1 co-receptor tropism on treatment experienced and naïve individuals will be necessary to unravel the emergence of high frequency of X4 viruses in subtype C infections. This is necessary since C viruses drive the epidemic in regions with the highest burden of infection. Sequence analysis of the V3 region showed that X4 and dual/mixture R5X4 viruses were associated with an increased number of positively charged amino acids, loss of a potential glycosylation site, as well as variable V3 loop amino acid lengths in contrast to R5 viruses. The significance of Arginine and Glycine insertions observed in the crown motif at positions 19 and 20 respectively in X4 viruses needs to be determined.

Concordance among three algorithms was used to predict the most likely co-receptor preferences. Predictions with PhenoSeq-C had an 83% concordance with Geno2Pheno, and 76% concordance with PSSMsinsi. Geno2Pheno had an 81.9% concordance with PSSMsinsi. Previous studies have demonstrated best concordance (>85%) between Geno2Pheno and C-PSSMsinsi, (Seclén *et al.* 2011; Sechet *et al.* 2015; Kalu *et al.* 2017). In this study we have observed the best concordance with Geno2Pheno and PhenoSeq-C. These tools are used to detect nonrandom distribution of amino acid at adjacent sites associated with empirically determined groupings of sequences; the tools may be limited in dealing with mixed bases.

In this study, the application of De Novo assembly approaches with reads spanning the V3 loop resolved the issue of mixed bases in consensus sequences. And to reduce false positive variants calling, all sequences were trimmed at the ends as part of the assembly process using the modified-Mott algorithm and quality scores assigned by the sequencing base caller. This algorithm trims the ends and improves the error rate by more than the error probability limit threshold set, in this case set to 0.001%, and we have considered only those variants with a read depth of at least 100 and present in at least 50 reads. The cut-offs for depth and read length were selected after a manual inspection of test batches of samples and ensured that at least 5 reads are present to call a variant. Minority variants were called based on the variant read depth, variant frequency and consensus base frequency with >20-2% cut-offs using the find variation/SNPs feature from Annotation and prediction menu in Geneious® software version 8.1.5 (http://www.geneious.com).

In the current study, viral DNA was used from peripheral blood mononuclear cells (PBMCs) to enhance the success of PCR amplification since the study subjects were on treatment and some had very low or undetectable viral loads. HIV-1 tropism can be evaluated in plasma or PBMCs. However, only tropism testing of HIV obtained from plasma RNA has been validated as a tool to predict virological response to CCR5 antagonists in clinical trials. The use of proviral DNA from PBMC for co-receptor determination (an approach that still has to be validated) may be a limitation of the current study. However, a number of studies, employing Sanger sequencing, have shown a strong correlation in co-receptor preference prediction between HIV variants obtained from plasma and PBMC (Paar *et al.* 2011; Parisi *et al.* 2011; Svicher *et al.* 2014; Baumann *et al.* 2015).

Next Generation Sequencing of HIV-1 from plasma has been shown to predict virologic response to Maraviroc as well as enhanced sensitivity in the Trofile phenotypic assay and also sensitive to minority non-R5 variants than population based sequencing (Swenson *et al.* 2011, 2012; Kagan *et al.* 2012). In addition, other studies using Sanger and NGS, have shown that viruses from plasma and PBMC compartments can be used to reliably predict virologic success to Maraviroc, although with an apparent improved outcome with viruses obtained from plasma (Pou *et al.* 2013; Swenson *et al.* 2013). It is also known that viruses in plasma are ‘replenished’ from archives such, as PBMCs. Hence viruses emanating from PBMCs are clinically relevant.

In conclusion, our findings show that a highly significant number of chronically HIV-1 subtype C infected patients on Maraviroc-free treatment are predicted to harbor CXCR4 utilizing viruses. The data is useful in the consideration of whether to include entry antagonists such as Maraviroc in alternative forms of treatment for patients failing second line treatment regimen in the study setting. The determination of co-receptor usage prior to initiation of therapy consisting of Maraviroc is suggested.

3.7 REFERENCES

Baumann, R.E., Rogers, A.A., Hamdan, H.B., Burger, H., Weiser, B., Gao, W., Anastos, K., Young, M., Meyer, W.A., Pesano, R.L., Kagan, R.M. (2015) 'Determination of HIV-1 coreceptor tropism using proviral DNA in women before and after viral suppression', *AIDS Research and Therapy*, 12(1).

Berger, E.A., Doms, R.W., Fenyö, E.-M., Korber, B.T.M., Littman, D.R., Moore, J.P., Sattentau, Q.J., Schuitemaker, H., Sodroski, J., Weiss, R.A. (1998) 'A new classification for HIV-1', *Nature*, 391(6664), 240–240.

Björndal, a, Deng, H., Jansson, M., Fiore, J.R., Colognesi, C., Karlsson, a, Albert, J., Scarlatti, G., Littman, D.R., Fenyö, E.M. (1997) 'Coreceptor usage of primary human immunodeficiency virus type 1 isolates varies according to biological phenotype.', *Journal of virology*, 71(10), 7478–7487.

Cashin, K., Gray, L.R., Harvey, K.L., Perez-Bercoff, D., Lee, G.Q., Sterjovski, J., Roche, M., Demarest, J.F., Drummond, F., Harrigan, P.R., Churchill, M.J., Gorry, P.R. (2015) 'Reliable genotypic tropism tests for the major HIV-1 subtypes', *Scientific Reports*, 5.

Cecilia, D., Kulkarni, S.S., Tripathy, S.P., Gangakhedkar, R.R., Paranjape, R.S., Gadkari, D.A. (2000) 'Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India', *Virology*, 271(2), 253–258.

Choge, I., Cilliers, T., Walker, P., Taylor, N., Phoswa, M., Meyers, T., Viljoen, J., Violari, A., Gray, G., Moore, P.L., Papathanosopoulos, M., Morris, L. (2006) 'Genotypic and phenotypic characterization of viral isolates from HIV-1 subtype C-infected children with slow and rapid disease progression.', *AIDS research and human retroviruses*, 22(5), 458–65.

Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W.C., Moore, J.P., Trkola, A., Morris, L. (2003) 'The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C', *J Virol*, 77(7), 4449–4456.

Connell, B.J., Michler, K., Capovilla, A., Venter, W.D.F., Stevens, W.S., Papathanasopoulos, M.A. (2008) 'Emergence of X4 usage among HIV-1 subtype C: Evidence for an evolving epidemic in South Africa', *AIDS*, 22(7), 896–899.

Connor, R.I., Sheridan, K.E., Ceradini, D., Choe, S., Landau, N.R. (1997) 'Change in Coreceptor Use Correlates with Disease Progression in HIV-1-Infected Individuals', *Journal of Experimental Medicine*, 185(4), 621–628.

Doms, R.W., Peiper, S.C. (1997) 'Unwelcomed guests with master keyes: How HIV uses chemokine receptors for cellular entry', *Virology*.

Engelbrecht, S., de Villiers, T., Sampson, C.C., zur Megede, J., Barnett, S.W., van Rensburg, E.J. (2001) 'Genetic analysis of the complete gag and env genes of HIV type 1 subtype C primary isolates from South Africa.', *AIDS research and human retroviruses*, 17(16), 1533–1547.

Esbjörnsson, J., Månsson, F., Martínez-Arias, W., Vincic, E., Biague, A.J., da Silva, Z.J., Fenyö, E.M., Norrgren, H., Medstrand, P. (2010) 'Frequent CXCR4 tropism of HIV-1 subtype A and CRF02_AG during late-stage disease - indication of an evolving epidemic in West Africa', *Retrovirology*, 7.

Frange, P., Galimand, J., Vidal, N., Goujard, C., Deveau, C., Souala, F., Peeters, M., Meyer, L., Rouzioux, C., Chaix, M.L. (2008) 'New and old complex recombinant HIV-1 strains among patients with primary infection in 1996-2006 in France: The French ANRS CO06 primo cohort study', *Retrovirology*, 5.

Garcia-Perez, J., Rueda, P., Alcami, J., Rognan, D., Arenzana-Seisdedos, F., Lagane, B., Kellenberger, E. (2011) 'Allosteric model of maraviroc binding to CC Chemokine Receptor 5 (CCR5)', *Journal of Biological Chemistry*, 286(38), 33409–33421.

Gilliam, B.L., Riedel, D.J., Redfield, R.R. (2010) 'Clinical use of CCR5 inhibitors in HIV and beyond', *Journal of Translational Medicine*.

Gulick, R.M., Lalezari, J., Goodrich, J., Clumeck, N., DeJesus, E., Horban, A., Nadler, J., Clotet, B., Karlsson, A., Wohlfeiler, M., Montana, J.B., McHale, M., Sullivan, J., Ridgway, C., Felstead, S., Dunne, M.W., van der Ryst, E., Mayer, H. (2008) 'Maraviroc for Previously Treated Patients with R5 HIV-1 Infection', *New England Journal of Medicine*, 359(14), 1429–1441.

Gulick, R.M., Wilkin, T.J., Chen, Y.Q., Landovitz, R.J., Amico, K.R., Young, A.M., Richardson, P., Marzinke, M.A., Hendrix, C.W., Eshleman, S.H., McGowan, I., Cottle, L.M., Andrade, A., Marcus, C., Klingman, K.L., Chege, W., Rinehart, A.R., Rooney, J.F., Andrew, P., Salata, R.A., Magnus, M., Farley, J.E., Liu, A., Frank, I., Ho, K., Santana, J., Stekler, J.D., McCauley, M., Mayer, K.H. (2017) 'Phase 2 Study of the Safety and Tolerability of Maraviroc-Containing Regimens to Prevent HIV Infection in Men Who Have Sex With Men (HPTN 069/ACTG A5305)', *The Journal of infectious diseases*, 215(2), 238–246.

Hatse, S., Princen, K., Vermeire, K., Gerlach, L.O., Rosenkilde, M.M., Schwartz, T.W., Bridger, G., De Clercq, E., Schols, D. (2003) 'Mutations at the CXCR4 interaction sites for AMD3100 influence anti-CXCR4 antibody binding and HIV-1 entry', *FEBS Letters*, 546(2–3), 300–306.

Huang, W., Eshleman, S.H., Toma, J., Fransen, S., Stawiski, E., Paxinos, E.E., Whitcomb, J.M., Young, A.M., Donnell, D., Mmiro, F., Musoke, P., Guay, L.A., Jackson, J.B., Parkin, N.T., Petropoulos, C.J. (2007) 'Coreceptor Tropism in Human Immunodeficiency Virus Type 1 Subtype D: High Prevalence of CXCR4 Tropism and Heterogeneous Composition of Viral Populations', *Journal of Virology*, 81(15), 7885–7893.

Irlbeck, D.M., Amrine-Madsen, H., Kitrinos, K.M., Labranche, C.C., Demarest, J.F. (2008) 'Chemokine (C-C motif) receptor 5-using envelopes predominate in dual/mixed-tropic HIV from the plasma of drug-naive individuals', *AIDS*, 22(12), 1425–1431.

Jensen, M.A., Coetzer, M., van 't Wout, A.B., Morris, L., Mullins, J.I. (2006) 'A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences.', *Journal of virology*, 80(10), 4698–704.

Kagan, R.M., Johnson, E.P., Siaw, M., Biswas, P., Chapman, D.S., Su, Z., Platt, J.L., Pesano, R.L. (2012) 'A Genotypic Test for HIV-1 Tropism Combining Sanger Sequencing with Ultradeep Sequencing Predicts Virologic Response in Treatment-Experienced Patients', *PLoS ONE*, 7(9).

Kalu, A.W., Telele, N.F., Gebreselasie, S., Fekade, D., Abdurahman, S., Marrone, G., Sönnerborg, A. (2017) 'Prediction of coreceptor usage by five bioinformatics tools in

a large Ethiopian HIV-1 subtype C cohort', *PLoS ONE*, 12(8).

Ketseoglou, I., Lukhwareni, A., Steegen, K., Carmona, S., Stevens, W.S., Papathanasopoulos, M.A. (2014) 'Viral tropism and antiretroviral drug resistance in HIV-1 subtype C-infected patients failing highly active antiretroviral therapy in Johannesburg, South Africa.', *AIDS research and human retroviruses*, 30(3), 289–93.

Meintjes, G., Black, J., Conradie, F., Cox, V., Dlamini, S., Fabian, J., Maartens, G., Manzini, T., Mathe, M., Menezes, C., Moorhouse, M., Moosa, Y., Nash, J., Orrell, C., Pakade, Y., Venter, F., Wilson, D., Society, S.A.H.I.V.C. (2014) 'Adult antiretroviral therapy guidelines 2014', *Southern African Journal of HIV Medicine*, 15(4), 121–143.

Melby, T., DeSpirito, M., DeMasi, R., Heilek-Snyder, G., Greenberg, M.L., Graham, N. (2006) 'HIV-1 Coreceptor Use in Triple-Class Treatment–Experienced Patients: Baseline Prevalence, Correlates, and Relationship to Enfuvirtide Response', *The Journal of Infectious Diseases*, 194(2), 238–246.

Ostrowski, M. a, Justement, S.J., Catanzaro, a, Hallahan, C. a, Ehler, L. a, Mizell, S.B., Kumar, P.N., Mican, J. a, Chun, T.W., Fauci, a S. (1998) 'Expression of chemokine receptors CXCR4 and CCR5 in HIV-1-infected and uninfected individuals.', *Journal of immunology (Baltimore, Md. : 1950)*, 161(6), 3195–201.

Paar, C., Geit, M., Stekel, H., Berg, J. (2011) 'Genotypic prediction of human immunodeficiency virus type 1 tropism by use of plasma and peripheral blood mononuclear cells in the routine clinical laboratory', *Journal of Clinical Microbiology*, 49(7), 2697–2699.

Papathanasopoulos, M. a, Cilliers, T., Morris, L., Mokili, J.L., Dowling, W., Birx, D.L., McCutchan, F.E. (2002) 'Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa', *AIDS research and human retroviruses*, 18(12), 879–86.

Parisi, S.G., Andreoni, C., Sarmati, L., Boldrin, C., Buonomini, A.R., Andreis, S., Scaggiante, R., Cruciani, M., Bosco, O., Manfrin, V., D'Ettorre, G., Mengoli, C., Vullo, V., Palù, G., Andreoni, M. (2011) 'HIV coreceptor tropism in paired plasma,

peripheral blood mononuclear cell, and cerebrospinal fluid isolates from antiretroviral-naïve subjects', *Journal of Clinical Microbiology*, 49(4), 1441–1445.

Ping, L.H., Nelson, J. a, Hoffman, I.F., Schock, J., Lamers, S.L., Goodman, M., Vernazza, P., Kazembe, P., Maida, M., Zimba, D., Goodenow, M.M., Eron, J.J., Fiscus, S. a, Cohen, M.S., Swanstrom, R. (1999) 'Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants.', *Journal of virology*

Pou, C., Codoñer, F.M., Thielen, A., Bellido, R., Pérez-Álvarez, S., Cabrera, C., Dalmau, J., Curriu, M., Lie, Y., Noguera-Julian, M., Puig, J., Martínez-Picado, J., Blanco, J., Coakley, E., Däumer, M., Clotet, B., Paredes, R. (2013) 'HIV-1 Tropism Testing in Subjects Achieving Undetectable HIV-1 RNA: Diagnostic Accuracy, Viral Evolution and Compartmentalization', *PLoS ONE*, 8(8).

R., G., T.J., W., Y., C., R.J., L., K.R., A., A., Y., P., R., M.A., M., M., M. (2016) 'HPTN 069/ACTG 5305: Phase ii study of maraviroc-based regimens for HIV prep in MSM', *Topics in Antiviral Medicine*, 24(E-1), 41–42.

Raymond, S., Delobel, P., Mavigner, M., Cazabat, M., Encinas, S., Souyris, C., Bruel, P., Sandres-Sauné, K., Marchou, B., Massip, P., Izopet, J. (2010) 'CXCR4-using viruses in plasma and peripheral blood mononuclear cells during primary HIV-1 infection and impact on disease progression', *AIDS*, 24(15), 2305–2312.

Van Rensburg, E.J., Smith, T.L., Zeier, M., Robson, B., Sampson, C., Treurnicht, F., Engelbrecht, S. (2002) 'Change in co-receptor usage of current South African HIV-1 subtype C primary isolates', *AIDS*, 16(18), 2479–2480.

Richman, D.D., Bozzette, S.A. (1994) 'The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression', *J Infect Dis*, 169(5), 968–974.

Riemenschneider, M., Cashin, K.Y., Budeus, B., Sierra, S., Shirvani-Dastgerdi, E., Bayanolhagh, S., Kaiser, R., Gorry, P.R., Heider, D. (2016) 'Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C', *Scientific Reports*, 6.

van Rij, R.P., Visser, J.A., van Praag, R.M.E., Rientsma, R., Prins, J.M., Lange, J.M.A., Schuitemaker, H. (2002) 'Both R5 and X4 Human Immunodeficiency Virus Type 1

Variants Persist during Prolonged Therapy with Five Antiretroviral Drugs', *J. Virol.*, 76(6), 3054–3058.

Schuitemaker, H., Van 't Wout, A.B., Lusso, P. (2010) 'Clinical significance of HIV-1 coreceptor usage', *Journal of Translational Medicine*.

Sechet, M., Roussel, C., Schmit, J.L., Saroufim, C., Ghomari, K., Merrien, D., Cordier, F., Pik, J.J., Landgraf, N., Douadi, Y., Liné, D., Duverlie, G., Castelain, S. (2015) 'X4 tropic virus prediction is associated with a Nadir CD4 T-cell count below 100 cells/mm^3', *Intervirology*, 58(3), 155–159.

Seclén, E., Soriano, V., González, M.M., Gómez, S., Thielen, A., Poveda, E. (2011) 'High concordance between the position-specific scoring matrix and geno2pheno algorithms for genotypic interpretation of HIV-1 tropism: V3 length as the major cause of disagreement', *Journal of Clinical Microbiology*, 49(9), 3380–3382.

Singh, A., Sunpath, H., Green, T.N., Padayachi, N., Hiramen, K., Lie, Y., Anton, E.D., Murphy, R., Reeves, J.D., Kuritzkes, D.R., Ndung'u, T. (2011) 'Drug resistance and viral tropism in HIV-1 subtype C-infected patients in KwaZulu-Natal, South Africa: implications for future treatment options.', *Journal of acquired immune deficiency syndromes (1999)*, 58(3), 233–40.

Svicher, V., Alteri, C., Montano, M., Nori, A., D'Arrigo, R., Andreoni, M., Angarano, G., Antinori, A., Antonelli, G., Allice, T., Bagnarelli, P., Baldanti, F., Bertoli, A., Borderi, M., Boeri, E., Bon, I., Bruzzone, B., Barresi, R., Calderisi, S., Callegaro, A.P., Capobianchi, M.R., Gargiulo, F., Castelli, F., Cauda, R., Ceccherini-Silberstein, F., Clementi, M., Chirianni, A., Colafigli, M., D'Arminio Monforte, A., De Luca, A., Di Biagio, A., Di Nicuolo, G., Di Perri, G., Di Santo, F., Fadda, G., Galli, M., Gennari, W., Ghisetti, V., Costantini, A., Gori, A., Gulminetti, R., Leoncini, F., Maffongelli, G., Maggiolo, F., Maserati, R., Mazzotta, F., Meini, G., Micheli, V., Monno, L., Mussini, C., Nozza, S., Paolucci, S., Palù, G., Parisi, S., Parruti, G., Pignataro, A.R., Quirino, T., Re, M.C., Rizzardini, G., Sanguinetti, M., Santangelo, R., Scaggiante, R., Sterrantino, G., Turriziani, O., Vatteroni, M.L., Viscoli, C., Vullo, V., Zazzi, M., Lazzarin, A., Perno, C.F. (2014) 'Genotypic testing on HIV-1 DNA as a tool to assess HIV-1 co-receptor usage in clinical practice: Results from the DIVA study group', *Infection*, 42(1), 61–71.

Swenson, L.C., Däumer, M., Paredes, R. (2012) 'Next-generation sequencing to assess HIV tropism', *Current Opinion in HIV and AIDS*.

Swenson, L.C., Dong, W.W.Y., Mo, T., Demarest, J., Chapman, D., Ellery, S., Heera, J., Valdez, H., Poon, A.F.Y., Harrigan, P.R. (2013) 'Use of cellular HIV DNA to predict virologic response to maraviroc: Performance of population-based and deep sequencing', *Clinical Infectious Diseases*, 56(11), 1659–1666.

Swenson, L.C., Mo, T., Dong, W.W.Y., Zhong, X., Woods, C.K., Thielen, A., Jensen, M.A., Knapp, D.J.H.F., Chapman, D., Portsmouth, S., Lewis, M., James, I., Heera, J., Valdez, H., Harrigan, P.R. (2011) 'Deep third variable sequencing for HIV type 1 tropism in treatment-naive patients: A reanalysis of the MERIT trial of maraviroc', *Clinical Infectious Diseases*, 53(7), 732–742.

Tersmette, M., Gruters, R.A., de Wolf, F., de Goede, R.E., Lange, J.M., Schellekens, P.T., Goudsmit, J., Huisman, H.G., Miedema, F. (1989) 'Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates.', *Journal of virology*, 63(5), 2118–25.

Thielen, A., Lengauer, T. (2012) 'Geno2pheno[454]: A web server for the prediction of HIV-1 coreceptor usage from next-generation sequencing data', *Intervirology*, 55(2), 113–117.

Weinberger, A.D., Perelson, A.S., Ribeiro, R.M., Weinberger, L.S. (2009) 'Accelerated Immunodeficiency by Anti-CCR5 Treatment in HIV Infection', *PLoS Computational Biology*, 5(8).

Woollard, S.M., Kanmogne, G.D. (2015) 'Maraviroc: a review of its use in HIV infection and beyond.', *Drug design, development and therapy*, 9, 5447–68.

Yi, Y., Isaacs, S.N., Williams, D. a, Frank, I., Schols, D., De Clercq, E., Kolson, D.L., Collman, R.G. (1999) 'Role of CXCR4 in cell-cell fusion and infection of monocyte-derived macrophages by primary human immunodeficiency virus type 1 (HIV-1) strains: two distinct mechanisms of HIV-1 dual tropism.', *Journal of Virology*, 73(9), 7117–7125.

CHAPTER FOUR

Genetic Variation within the Coding Region Of CCR5 In Chronically HIV-1 Infected South African Individuals, The majority of whom were drug-experienced

ABSTRACT

Background: The human co-receptor CCR5 is an important target for viral and cellular antiretroviral interventions. To date, several mutations and single-nucleotide polymorphisms (SNPs) in this gene have been discovered and found to be important genetic factors capable of influencing susceptibility to HIV-1 infection or affecting the rate of disease progression. In this study we characterize polymorphism in the coding region of CCR5 in an HIV-1 infected South African population with a majority of drug-experienced individuals.

Methods: Genomic DNA was isolated from peripheral blood mononuclear cells of 192 HIV infected patients and CCR5 was sequenced on an Illumina MiniSeq platform. Sequences were analyzed using Geneious® software version 8.1.5. Single nucleotide polymorphisms (SNPs) and Indels were detected using Bayesian Haplotype-based polymorphism discovery and genotyping tool within Geneious software. The SNPs were confirmed and compared to SNPs in other populations reported in the 1000 Genome Phase III and HapMap. Hardy-Weinberg Equilibrium was calculated, Haplotypes and Linkage disequilibrium were inferred using LDlink 3.0 web tool.

Results: The following SNPs; P35P, S75S, Y89Y, A335V and Y339F and their varying frequencies were detected in this study at varying frequencies, The A335V variant was detected at the highest frequency 17.4% (29/167), whereas Y339F was detected in 2 patients 1.2% (2/167). The G265S variant is reported for the first time in this study at 0.6% (1/167) frequency. The SNPs detected were in strong LD (D'= 1, R^2 = 0.0) with minor deviation from the Hardy-Weinberg Equilibrium. These variants were not located at the binding motif of Maraviroc.

Conclusion: Two variants were detected at a higher frequency in this study compared to previously report in South Africa. The variants detected in this study are not located in the binding motif of Maraviroc, therefore those infected with R5 viruses may benefit from this drug.

Key words: CCR5; Single nucleotide polymorphism; Maraviroc; South Africa

4.1 INTRODUCTION

The entry of HIV-1 into the host cell is mediated by host cell CD4 and co-receptors either CCR5 or CXCR4. Several additional co-receptors have been identified, but only CCR5 and CXCR4 are bonafide co-receptors in the attachment and entry of HIV into the host (Deng *et al.* 1996; Dragic *et al.* 1996). Although CXCR4 using viruses are transmissible, the majority of the infections are established by the CCR5 using viruses. Because CCR5 is the predominant co-receptor for clinical HIV isolates, it has become a very attractive target for anti HIV-1 therapy. A number of small molecules have demonstrated potent antiviral inhibition both in vitro and in clinical trails. Maraviroc is one of the small molecules approved by the US Food and Drug Administration as CCR5 antagonist, inhibiting HIV-1 entry into the host (Dorr *et al.* 2005; Watson 2005; Garcia-Perez *et al.* 2011).

The CCR5 is a 40.6 kDa protein consisting of 352 amino acids (Mueller and Strange 2004). Its structure consists of the N-terminal extra cellular domain, seven transmembrane domains made up of hydrophobic residues with, three extra cellular loops, and three intracellular loops and a C-terminal tail. It carries specific motifs important for chemokine ligand binding, functional responses of the receptor and HIV-1 co-receptor activity. Several amino acid residues including the N-terminal and position 2, 3, 10, 11 and 18 were found to interact CCR5 using viruses, and are also important for HIV envelope binding affinity (Dragic *et al.* 1998; Rabut *et al.* 1998; Blanpain *et al.* 1999; Farzan *et al.* 1999).

Shortly after the role of CCR5 in HIV infection was discovered, studies were conducted that led to the discovery of the CCR5Δ32 mutant and its association with protection to HIV-1 infection (Picton *et al.* 2010). This discovery provided the first genetic evidence of natural protection to HIV-1 infection and prompted further studies investigating the effects of polymorphism in CCR5 and how it influences the outcome of HIV-1 exposure and infection. To date, several mutations and single-nucleotide polymorphisms (SNPs) in this gene have been discovered and found to be important genetic factors capable of influencing susceptibility to HIV-1 infection or affecting the rate of disease progression (Picton *et al.* 2010). Mutations in the coding region can affect the CCR5 protein structure, its expression, chemokine binding, and transport on signaling.

One of the factors that play a role in SNPs frequency distribution within the CCR5 gene is ethnicity. The most studied variant that demonstrate these differences is the CCR5Δ32 that exists mostly in the Caucasian population and is rare in the African population. There are also SNPs that vary in frequency in African populations. Therefore, characterizing variation within the coding region of the CCR5 genes in a diverse country with a rich collection of ethnic background such as South Africa will provide good characterization of the variation that exists in CCR5 within the South African population.

4.2 Hypothesis

There is limited variability in the CCR5 Maraviroc binding motif in HIV-1 chronically infected individuals.

4.3 Objective

The objective of the study is to characterize variation within CCR5 coding region and determine the SNP frequency distribution in a South African population.

4.3.1 Specific objectives

4.3.1.1 To characterize and determine the frequency of the polymorphisms in the CCR5 gene in HIV-1 infected South African

4.3.1.2 To compare the frequency of the polymorphism to other populations

4.3.1.3 To determine whether the identified mutations were located at the Maraviroc binding motifs.

4.4 MATERIALS AND METHODS

4.4.1 Study population and sample collection

Characteristics of the study population and sample collection procedures are as described in chapter two, section 2.3.2.

4.4.2 Primer design, DNA extraction and Polymerase Chain Reaction (PCR) amplification of CCR5 gene

Primers and protocols to amplify the CCR5 gene are also described in chapter two, from section 2.3.3.1.3 to sections 2.3.3.1.8. Briefly, DNA was extracted from peripheral blood mononuclear cells (PBMC) of 192 patients and the CCR5 gene amplified in a nested PCR using standard PCR reagents to yield a 1203bp product.

4.4.3 Library preparation and MiniSeq sequencing

The CCR5 amplicons were purified using Ampure XP beads (Beckman coulter) and quantified using a Qubit 3 dsDNA HS kit with detection range from 10pg/µl to 100ng/µl (Invitrogen). The concentration of each amplicon was diluted to 0.2ng/µl with 10mM Tris HCl. The Nextera XT DNA sample preparation kit (Illumina k.k Tokyo, Japan) was used to fragment 1ng input DNA and tag with sequencing adaptors. The libraries were amplified in a limited PCR cycle program to add Illumina index 1 (i7) and index 2 (i5) barcodes for sample identification, and again purified using Ampure XP beads. Agarose electrophoresis (E-gels) was used to verify the size of the libraries, and normalized to 4nM each to ensure equal library representation in the pool. Five microliter of each normalized library was mixed, heat denatured and diluted to 1.8pM with 20% PhiX as control (Illumina k.k Tokyo,Japan). The prepared libraries were sequenced in an Illumina MiniSeq using a high output kit for 300 cycles (Illumina k.k Tokyo, Japan) and generating paired-end 2x150 bp long sequence reads for each.

4.4.4 De-multiplexing and sequence quality control evaluation

Sequences are de-multiplexed automatically on the MiniSeq as part of the data processing steps and ends pairing. Fastq files were generated for each sample representing the two paired-end reads. Sequence quality was validated using the FastQC

programme. The fastq files were imported into Geneious® software version 8.1.5 for filtering and trimming of sequences as well as downstream analysis. All sequences were trimmed at the ends as part of the assembly process using the modified-Mott algorithm and quality scores assigned by the sequencing base caller. This algorithm trims the ends and improves the error rate by more than the error probability limit threshold set, in this case 0.001%. The two sequence lists from each sample were paired and mapped to the human genome h19 to demultiplex reads for CCR5 with the following parameters: Gaps were allowed with a maximum of 15% per read and gap size of 15, word length was set to 14, maximum mismatches per read were set to 25%, minimum overlap identity was set to 80%, maximum ambiguity was set to 16. In general, the reference assembler uses a seed and expand-type mapper, followed by a fine-tuning step that was set to none (fast / read mapping) for this analysis.

4.4.5 Single nucleotide polymorphisms (SNPs) and indels detection and verification

Variants or SNPs were detected using two methods within Geneious®; (1) the Find SNP option, detecting variant frequency at 5% cutoff, maximum variant P-value 10^{-6} and maximum strand-Bias P-value 10^{-5} when exceeding 65% bias. The variants were detected only inside the CDS. (2) The other option was Bayesian haplotype-based polymorphism discovery and genotyping tool (http://arxiv.org/abs/1207.3907), with the following parameters: ploidy and minimum Alternate Count set at 2, minimum Alternate fraction set at 0.2, minimum probability at 0, Combined nearby variants (haplotype-length) set at 2. The effect of variants on translation was analyzed.

The GenBank NCBI SNPs database (dbSNP) and ENSEMBLE genome browser were searched for all reported SNPs in CCR5 to determine whether variants detected in this study had been previously reported. Frequencies of SNPs detected were compared to other populations published in the 1000 genome browser, HapMap and center for Biotechnology information (NCBI) SNPs database. Chi-square test/fisher exact test were performed to determine if there was any significant differences between these populations.

4.4.6 Hardy-Weinberg Equilibrium (HWE)

All variants loci detected within the coding regions of these genes were tested for deviation from Hardy-Weinberg Equilibrium (HWE) using an excel HWE calculator and chi-square test with P<0.05 showing non-consistence with HWE (Court *et al*. 2012).

4.4.7 Haplotypes and association mapping / Linkage disequilibrium inference

Linkage disequilibrium (LD) analysis between SNP per and haplotype assignments were determined using LDLink 3.0 web tool; exploring the LDmatrix, LDpair and LDhap modules (https://analysistools.nci.nih.gov/LDlink/?tab=home). This tool investigates patterns of linkage disequilibrium across a variety of ancestral population group. In this case African population from the 1000 Genome project phase 3 (version 5). Linkage disequilibrium was measured by calculated D prime (D'), R squared (R^2) and goodness-of-fit (chi-square and p-value) with variant Rs number assigned by dbSNP used as input. The D' is an indicator of allelic segregation for two genetic variants, which ranges from 0 to 1 with higher values indicating tight linkage of alleles and low or a D' value of 0 indicates no linkage of alleles. R squared (R^2) is a measure of correlation of alleles for two genetic variants. R^2 values range from 0 to 1 with higher values indicating a higher degree of correlation. An R^2 value of 0 indicates alleles are independent, whereas an R^2 value of 1 indicates an allele of one variant perfectly predicts an allele of another variant. Haplotypes were determined using the LDhap module, which calculates population specific haplotypes frequencies of all haplotypes observed for a list of query variants. Publicly available reference haplotypes from the 1000 Genome project phase 3 (version 5) are used by LDlink to calculate population-specific measures of linkage disequilibrium (Machiela and Chanock 2015). Variant of minor allele frequencies <5% and those not assigned in the dbSNP (novel SNPs) were excluded from LD determination and haplotype assignment.

4.5 RESULTS

4.5.1 PCR products and sequence analysis

From a total of 192 samples, the complete 1000bp CCR5 coding region was obtained in from 92.2% (177/192) individuals and quality sequences were obtained from 94.4% (167/177).

4.5.2 Single Nucleotide Polymorphisms (SNPs) identification, Hardy Weinberg Equilibrium and Linkage Disequilibrium associations

Across the coding region of the gene, 19.2% (32/167) of the individuals had nonsynonymous changes compared to the consensus sequence, 5.98 % (10/167) had synonymous changes and 74.9% (125/167) of the individuals had no changes (table 4.1). There were no insertion and deletions observed in sequences from our study population. SNPs in this gene are in strong linkage disequilibrium (D'= 1, R^2 = 1.0) and have not deviated from the Hardy Weinberg equilibrium. Variant of minor allele frequencies <5% and those not assigned in the dbSNP (novel SNPs) were excluded from LD determination and haplotype assignment and only one haplotype of the A335V variant was observed at 16.8 heterozygous and 0.6% homozygous allele frequency.

Table 4.1: Human CCR5 nonsynonymous and synonymous changes, genotype frequencies and Hardy Weinberg Equilibrium calculations in the study population.

CCR5 Nonsynonymous changes (n= 167)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
G265S (**NI**)	**G**GC→**A**GC Transition	793 G/G 793 G/A	3	166 (99.4) 1 (0.6)	P-value = 0.96 X^2 = 0.001
A335V **(rs1800944)**	G**C**C→G**T**A Transition	1004 C/C 1004 C/T 1004 T/T	3	138 (82.6) 28 (16.8) 1 (0.6)	P-value = 0.8 X^2 = 0.06
Y339F **(rs1800945)**	T**A**C→T**T**C Transversion	1016 A/A 1016 A/T	3	165 (98.8) 2 (1.2)	P-value = 0.94 X^2 = 0.006
CCR5 Synonymous changes					
P35P **(rs113471490)**	CC**G** →CC**A** Transition	105 G/G 105 G/A	2	161 (97.4) 6 (3.6)	P-value = 0.81 X^2 = 0.05
S75S **(rs1800941)**	TC**T**→TC**C** Transition	225 T/T 225 T/C	3	165 (98.9) 2 (1.2)	P-value =0.84 X^2 = 0.04
Y89Y **(rs750809479)**	TA**T**→TA**C** Transition	267 T/T 267 T/C	3	165 (98.9) 2 (1.2)	P-value =0.94 X^2 = 0.006
No mutation				**125 (74.9)**	

Note: rs= Reference single nucleotide polymorphism; NI= Not Identified (novel variants)

The SNPs were compared to other SNPs in other populations in the 1000 genome browser. The G265S variant has not been reported before, and is reported for the first time in the study. The A335V, Y339F and S75S have been reported before in African population. The observed A335V frequency was much higher in this study than previously reported (table 4.2 and table 4.3).

Table 4.2: Comparison of CCR5 allele frequencies between a South African population and East Asian (EAS), European (EUR), African (AFR), Ad Mixed American (AMR) and South Asian (SAS) populations.

CCR5 Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
G265S **(NI)**	G (**G**) A (**S**)	99.4 0.6	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)
A335V **(rs1800944)**	C (**A**) T (**V**)	82.6 17.4	100 0	100 0	99.4 0.4	98.7 1.3	100 0
Y339F **(1800945)**	A (**Y**) T (**F**)	98.8 1.2	100 0	100 0	99 1	99.7 0.3	100 0
CCR5 synonymous allele frequencies (%)							
P35P **(rs113471490)**	G (**P**) A (**P**)	100 0	100 0	100 0	99.9 0.1	100 0	100 0
S75S **(rs1800941)**	T (**S**) C (**S**)	98.9 1.2	100 0	100 0	99 1	99.6 4.3	100 0
Y89Y **(rs750809479)**	T (**Y**) A (**Y**)	98.9 1.2	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)

Note: rs= Reference single nucleotide polymorphism; NI= Not Identified (novel variants); NA= No data available

Table 4.3: Comparison of CCR5 allele frequencies between a South African population and other Africa population, which comprises of the African Carribeans in Barbados (**ACB**), Americans of African Ancestry in USA (**ASW**), Esan in Nigeria (**ESN**), Gambian in the Western Gambia (**GWD**), Luhya in Webuye, Kenya (**LWK**), Mende in Seirra Leone (**MSL**), Yoruba in ibadan, Nigeria (**YRI**).

CCR5 Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
G265S (**NI**)	G (**G**) A (**S**)	99.4 0.6	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)
A335V **(rs1800944)**	G (**A**) T (**V**)	82.6 17.4	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)
Y339F **(rs1800945)**	A (**Y**) T (**F**)	98.8 1.2	98 2	100 0	98 2	99 1	98 2	100 0	100 0
CCR5 synonymous allele frequencies (%)									
P35P **(rs113471490)**	G (**P**) A (**P**)	97.4 3.6	100 0	100 0	100 0	100 0	99 1	100	100
S75S **(rs1800941)**	T (**S**) C (**S**)	98.9 1.2	99 1	99 1	99 1	99 1	99 1	100 0	99 1
Y89Y **(rs750809479)**	T (**Y**) A (**Y**)	98.9 1.2	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)	(NA) (NA)

Note: rs= Reference single nucleotide polymorphism; NI= Not Identified (novel variants); NA= No data available

The mutations were then analyzed for their location on CCR5 (figure 4.1). They are located at the C-terminus of the protein and are not close to the binding motif of Maraviroc and V3 loop. The A335V was the most frequent in this study and we evaluated if there is any association with declined $CD4^+$ cell count and found no association (p= value 0.35, $X^2 = 2.1$)

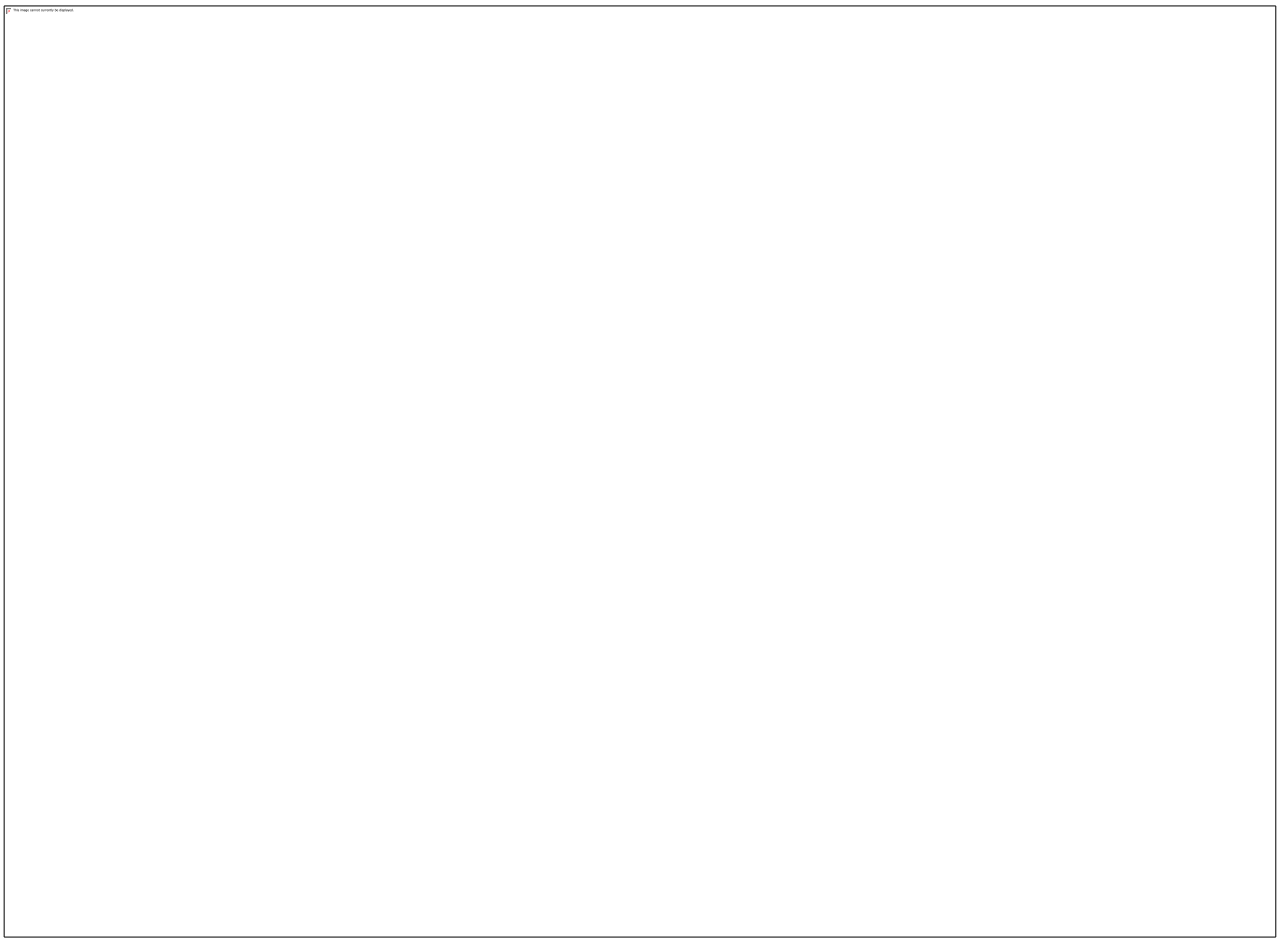


Figure 4.1: Graphic illustration of transmebrane organization of human CCR5 showing the location of the non-synonymous mutations identified in the study (highlighted in red). Residues involved in Maraviroc binding are: W86 in TM2, Y108 in TM3, I198 in TM5, Y251 in TM6 and E283 in TM7. None of the mutations observed were in these residues or close to them.

4.6 DISCUSSION AND CONCLUSION

The CCR5 co-receptor has become a target for antiretroviral therapy. Polymorphism within the coding region of this gene may affect its expression and binding affinities to inhibitors such as Maraviroc. In this study we have characterized polymorphisms within the CCR5 coding region in HIV-1 infected South Africans. The objective was to document the polymorphism and their frequency in a South African population and compare with other African populations. To date polymorphisms in the coding region of CCR5 have been studied and found to be important factors influencing susceptibility to HIV-1 infection and the progression of the disease (Picton *et al.* 2010). Six polymorphisms at frequencies ranging from 0.6 to 16.8 were observed in the population. Only 3 of these resulted in amino acid changes.

Several factors such as ethnicity and geographic locations may contribute to the distribution of SNP frequencies. This is observed in the CCR5Δ32 mutant, which occurs in variable frequencies of 4-15% in Caucasian populations, with an average 10% in European population (Galvani and Novembre 2005). It is rarely found in or African population. In this current study, conducted in Limpopo Province, we detected higher SNP frequencies than previously identified in studies by Petersen et al, 2001 and Picton et al, 2010 conducted in Gauteng and Western Cape provinces, which are widely separated geographic locations. The A335V and Y339F variants were reported to be highest in this study at frequencies 17% and 1.2% compared to 7.1 and 0.008% respectively in the previous studies. (Petersen *et al.* 2001; Picton *et al.* 2010)

The A335V polymorphism was previously reported by Ansari-Lari et al in 1997 and found at higher frequency in African American population compared to Caucasian population (Ansari-Lari *et al.* 1997; Carrington *et al.* 1997, 1999). The Y339F variant was reported in a South African population in 2001 (Petersen *et al.* 2001). In this study the frequency of this variant was found to closely relate to the Esan in Nigeria and Luyha in Webuye, Kenya. These polymorphisms are located in the open reading frame, and many studies have been conducted to elucidate the effects of this polymorphism on the structure and functions of the CCR5 protein. These have not been found to affect protein function, and they do not appear to contribute to changes in HIV disease progression (Carrington *et al.* 1997; Blanpain *et al.* 2002; Tamamis and Floudas 2014) which supports the hypothesis that there is limited variation in the CCR5 gene in the study cohort. The association of

A335V polymorphism and parameters of AIDS progression such as decline $CD4^+$ cells was determined and found no association (P= value 0.35, X^2 = 2.1). Further studies are needed to characterize the effects of the novel G265S mutation on the protein.

Given the importance of allele frequencies in genetic studies, it is critically important to be able to estimate them reliably. Factors that could be a possible contributor to the varying allele frequency distribution include the sequencing technology applied. Traditionally, allele frequencies were simply estimated by SNP genotype data and Sanger data, but using these approaches may underestimate allele frequencies when compared to Next Generation Sequencing (Bao *et al.* 2009; Holt *et al.* 2009; Ingman and Gyllensten 2009; Koboldt *et al.* 2009; Lynch 2009; Kim *et al.* 2011). In the current study Next Generation Sequencing was applied which relies crucially on the accurate calling of SNPs and genotypes. Next Generation Sequencing coverage levels often determines whether variant discovery can be made with a certain degree of confidence at particular base position (10X-30X depth of coverage). Further studies comparing CCR5 allele frequency distribution from different sequencing technologies will elucidate this aspect.

Residues to which Maraviroc binds are; the W86 binds in TM2, Y108 in TM3, I198 on TM5, Y251 in TM6 and E283 in TM7 (Seibert *et al.* 2006; Kondru *et al.* 2008; Garcia-Perez *et al.* 2011). In the current study none of the identified mutations are located at or near the binding sites of Maraviroc.

In conclusion mutations in the CCR5 coding region were observed at higher frequencies compared to other studies conducted in South African populations at different locations. The geographic location or the deep sequencing we used could be playing a role in the varying SNP frequencies. The variants detected in this study are not located in the binding motif of Maraviroc, therefore those infected with R5 viruses may benefit from this drug. Therefore the subset of patients infected by R5 viruses in our population may benefit from this drug.

4.7 REFERENCES

Ansari-Lari, M.A., Liu, X.M., Metzker, M.L., Rut, A.R., Gibbs, R.A. (1997) 'The extent of genetic variation in the CCR5 gene', *Nature Genetics*.

Bao, H., Xiong, Y., Guo, H., Zhou, R., Lu, X., Yang, Z., Zhong, Y., Shi, S. (2009) 'MapNext: a software tool for spliced and unspliced alignments and SNP detection of short sequence reads.', *BMC genomics*, 10 Suppl 3(Suppl 3), S13.

Blanpain, C., Libert, F., Vassart, G., Parmentier, M. (2002) 'CCR5 and HIV infection', *Receptors and Channels*.

Blanpain, C., Migeotte, I., Lee, B., Vakili, J., Doranz, B.J., Govaerts, C., Vassart, G., Doms, R.W., Parmentier, M. (1999) 'CCR5 binds multiple CC-chemokines: MCP-3 acts as a natural antagonist.', *Blood*, 94(6), 1899–905.

Carrington, M., Dean, M., Martin, M.P., O'Brien, S.J. (1999) 'Genetics of HIV-1 infection: Chemokine receptor CCR5 polymorphism and its consequences', *Human Molecular Genetics*.

Carrington, M., Kissner, T., Gerrard, B., Ivanov, S., O'Brien, S.J., Dean, M. (1997) 'Novel Alleles of the Chemokine-Receptor Gene CCR5', *The American Journal of Human Genetics*, 61(6), 1261–1267.

Court MH, M.H. (n.d.) Court's (2005–2008) Online Calculator. Tuft University Web Site. [online], *2012*.

Deng, H.K., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Di Marzio, P., Marmon, S., Sutton, R.E., Mark Hill, C., Davis, C.B., Peiper, S.C., Schall, T.J., Littman, D.R., Landau, N.R. (1996) 'Identification of a major co-receptor for primary isolates of HIV-1', *Nature*, 381(6584), 661–666.

Dorr, P., Westby, M., Dobbs, S., Griffin, P., Irvine, B., Macartney, M., Mori, J., Rickett, G., Smith-Burchnell, C., Napier, C., Webster, R., Armour, D., Price, D., Stammen, B., Wood, A., Perros, M. (2005) 'Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity', *Antimicrobial Agents and Chemotherapy*, 49(11), 4721–4732.

Dragic, T., Litwin, V., Allaway, G.P., Martin, S.R., Huang, Y., Nagashima, K.A., Cayanan, C., Maddon, P.J., Koup, R.A., Moore, J.P., Paxton, W.A. (1996) 'HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5', *Nature*, 381(6584), 667–673.

Dragic, T., Trkola, a, Lin, S.W., Nagashima, K. a, Kajumo, F., Zhao, L., Olson, W.C., Wu, L., Mackay, C.R., Allaway, G.P., Sakmar, T.P., Moore, J.P., Maddon, P.J. (1998) 'Amino-terminal substitutions in the CCR5 coreceptor impair gp120 binding and human immunodeficiency virus type 1 entry.', *Journal of virology*, 72(1), 279–285.

Farzan, M., Mirzabekov, T., Kolchinsky, P., Wyatt, R., Cayabyab, M., Gerard, N.P., Gerard, C., Sodroski, J., Choe, H. (1999) 'Tyrosine sulfation of the amino terminus of CCR5 facilitates HIV-1 entry', *Cell*, 96(5), 667–676.

Galvani, A.P., Novembre, J. (2005) 'The evolutionary history of the CCR5-Delta 32 HIV-resistance mutation', *Microbes and Infection*, 7(2), 302–309 .

Garcia-Perez, J., Rueda, P., Alcami, J., Rognan, D., Arenzana-Seisdedos, F., Lagane, B., Kellenberger, E. (2011) 'Allosteric model of maraviroc binding to CC Chemokine Receptor 5 (CCR5)', *Journal of Biological Chemistry*, 286(38), 33409–33421.

Holt, K.E., Teo, Y.Y., Li, H., Nair, S., Dougan, G., Wain, J., Parkhill, J. (2009) 'Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA', *Bioinformatics*, 25(16), 2074–2075.

Ingman, M., Gyllensten, U. (2009) 'SNP frequency estimation using massively parallel sequencing of pooled DNA', *European Journal of Human Genetics*, 17(3), 383–386.

Kim, S., Lohmueller, K., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., Nielsen, R. (2011) 'Estimation of allele frequency and association mapping using next-generation sequencing data', *BMC bioinformatics*, 12(1), 231, available: http://www.biomedcentral.com/1471-2105/12/231.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L. (2009) 'VarScan: Variant detection in massively parallel sequencing of individual and pooled samples', *Bioinformatics*,

25(17), 2283–2285.

Kondru, R., Zhang, J., Ji, C., Mirzadegan, T., Rotstein, D., Sankuratri, S., Dioszegi, M. (2008) 'Molecular interactions of CCR5 with major classes of small-molecule anti-HIV CCR5 antagonists.', *Molecular pharmacology*, 73(3), 789–800.

Lynch, M. (2009) 'Estimation of allele frequencies from high-coverage genome-sequencing projects', *Genetics*, 182(1), 295–301.

Machiela, M.J., Chanock, S.J. (2015) 'LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants', *Bioinformatics*, 31(21), 3555–3557.

Mueller, A., Strange, P.G. (2004) 'The chemokine receptor, CCR5', *International Journal of Biochemistry and Cell Biology*.

Petersen, D.C., Kotze, M.J., Zeier, M.D., Grimwood, A., Pretorius, D., Vardas, E., Van Rensburg, E.J., Hayes, V.M. (2001) 'Novel mutations identified using a comprehensive CCR5-denaturing gradient gel electrophoresis assay', *AIDS*, 15(2), 171–177.

Picton, A.C.P., Paximadis, M., Tiemessen, C.T. (2010) 'Genetic variation within the gene encoding the HIV-1 CCR5 coreceptor in two South African populations.', *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 10(4), 487–94.

Rabut, G.E., Konner, J.A., Kajumo, F., Moore, J.P., Dragic, T. (1998) 'Alanine substitutions of polar and nonpolar residues in the amino-terminal domain of CCR5 differently impair entry of macrophage- and dualtropic isolates of human immunodeficiency virus type 1.', *Journal of virology*, 72(4), 3464–8.

Seibert, C., Ying, W., Gavrilov, S., Tsamis, F., Kuhmann, S.E., Palani, A., Tagat, J.R., Clader, J.W., McCombie, S.W., Baroudy, B.M., Smith, S.O., Dragic, T., Moore, J.P., Sakmar, T.P. (2006) 'Interaction of small molecule inhibitors of HIV-1 entry with CCR5', *Virology*, 349(1), 41–54.

Tamamis, P., Floudas, C.A. (2014) 'Molecular recognition of CCR5 by an HIV-1 gp120 V3 loop', *PLoS ONE*, 9(4).

Watson, C. (2005) 'The CCR5 Receptor-Based Mechanism of Action of 873140, a Potent Allosteric Noncompetitive HIV Entry Inhibitor', *Molecular Pharmacology*, 67(4), 1268–12825.

CHAPTER FIVE

Characterization of APOBEC3 Variations in a South African Population

---

Nontokozo D. Matume, Denis M. Tebit, Laurie R. Gray, Stephen D. Turner, David Rekosh, Pascal O. Bessong, Marie-Louise Hammarskjöld. Characterization of APOBEC3 variation in a South African population. **Prepared for submission to Journal of Human Genetics, 2018 (Appendix II)**

All of the individual patient sequences used in this study have been submitted to the NCBI Sequence Read Archive (project number: PRJNA429751) and can be accessed using the following link; http://www.ncbi.nlm.nih.gov/bioproject/429751, pending manuscript acceptance. (**Accession numbers in Appendix II**)

ABSTRACT

The apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) genes A3D, A3F, A3G and A3H have all been implicated in the restriction of human immunodeficiency virus type 1 (HIV-1) replication. Polymorphisms in these genes may impact viral replication and fitness and contribute to viral diversity. To characterize polymorphisms in the coding regions of these APOBEC3 genes in a population from the Limpopo Province of South Africa, APOBEC3 gene fragments were amplified from genomic DNA of 192 HIV-1 infected subjects and sequenced on an Illumina MiSeq platform. SNPs were confirmed and compared to SNPs in other populations reported in the 1000 Genome Phase III and HapMap databases. Hardy-Weinberg Equilibrium was calculated and haplotypes and Linkage Disequilibrium (LD) were inferred using the LDlink 3.0 web tool.

Known variants compared to the GRCh37 consensus genome sequence were detected at relatively high frequencies (>5%) in all of the Apobec genes. A3H showed the most variation, with several of the variants present in both alleles in all of the patients. Several minor allele variants (<5%) were also detected in A3D, A3F and A3G. In addition, novel R6K, L221R and T238I variants in A3D and I117I in A3F were observed. Most of the SNPs were in strong LD with minor deviation from the Hardy-Weinberg Equilibrium. Four, five, four, and three haplotypes were identified for A3D, A3F, A3G, and A3H respectively. In general, polymorphisms in APOBEC3D, 3F, 3G and 3H were higher in our South African cohort than previously reported among other African, European and Asian populations.

Key words: APOBEC3; Single nucleotide polymorphism; South Africa

5.1 INTRODUCTION

The genes for the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like protein gene family (APOBEC3), a family of seven members (APOBEC3 A, B, C, D, F, G and H), are situated on human chromosome 22. The proteins encoded by these genes are cytidine deaminases that have been classified as restriction factors because of their role as innate immunity factors. They provide host cell defense against a diverse set of retroviruses, endogenous retroelements and DNA viruses, including human immunodeficiency virus (HIV) (Sheehy *et al.* 2002; Harris and Liddament 2004; Chiu and Greene 2008). APOBEC proteins restrict HIV through deamination of cytosines in viral cDNA during reverse transcription, causing G-to-A hypermutations in the viral DNA product, which results in degradation and viral inhibition (Zhang *et al.* 2003). The Vif protein of HIV has evolved to counteract this restriction by binding to APOBEC proteins leading to proteasomal degradation.

One of the most studied APOBEC proteins and the first that was discovered to restrict HIV-1 replication is APOBEC3G. In the absence of the HIV-1 Vif protein, APOBEC3G is efficiently packaged into viral particles, causing restriction during reverse transcription. The gene was originally identified as an HIV restriction factor because its expression converted a T-cell line that could support the replication of an HIV lacking *vif* into one that had a non-permissive phenotype (Sheehy *et al.* 2002).

Three other members of the APOBEC family, APOBEC3D (A3D), APOBEC3F (A3F), and APOBEC3H (A3H) can also be packaged into HIV particles and inhibit viral replication, when stably expressed in human T-cell lines (Hultquist *et al.* 2011). Endogenous A3D and A3F combine to generate the 5'-GA-to-AA mutation pattern observed in *vif*-negative HIV grown in the non-permissive T-cell line CEM2n (Refsland *et al.* 2012; An *et al.* 2016). Of the seven different human haplotypes of APOBEC3H, only hapII, hapV and hapVII are stable at the protein level and capable of HIV restriction (Dang *et al.* 2006; OhAinle *et al.* 2008; Harari *et al.* 2009; Wang *et al.* 2011; Ooms *et al.* 2013).

Several APOBEC3 (A3D, A3F, A3G and A3H) genes are known to possess common polymorphisms that render them defective with reduced antiviral activity and increased sensitivity to HIV-1 Vif (An *et al.* 2004; Ooms *et al.* 2010; Hultquist *et al.* 2011; Duggal *et al.* 2013; Refsland *et al.* 2014). The genetic associations between natural

polymorphisms in APOBEC genes and the ability of the resulting proteins to restrict HIV and the contribution of polymorphisms to overall HIV diversity and disease progression have not received widespread attention. Polymorphisms in APOBEC genes could play a significant role in HIV-1 evolution and diversity, especially in African populations, where the prevalence of HIV-1 is still increasing.

African populations are characterized by a high level of genetic diversity owing to a large number of variable genes and alleles (Cavalli-Sforza 1991; Jorde *et al.* 2000; Tishkoff and Williams 2002; Jakobsson *et al.* 2008). Patterns of genetic variation in the African population are influenced by a demographic history that includes changes in population size, admixture and locus-specific forces such as natural selection, recombination and mutation. Genetic studies of structural variation of genes across ethnically diverse populations have been conducted (Conrad and Hurles 2007). Many population genetic studies of African populations are based on analysis of genetic markers genotyped in a small number of populations in projects such as the 1000 Genomes Project (2010) and the International Haplotype Map (HapMap) Project (Batzer and Deininger 2002; Rosenberg *et al.* 2002; Li *et al.* 2008). Although these projects are valuable in their description of the overall human genetic diversity, they are limited in their coverage of African populations (Altshuler *et al.* 2010). Thus, it is important to continue to add information about African populations that are underrepresented in human genomic studies, such as the South African population.

South Africa embodies a rich collection of ethnic backgrounds in addition to the more recent Caucasian immigrants. The major ethnic groups include the Bapedi, Basotho, Ndebele, Swati, Tsonga, Tswana, Xhosa, Venda and Zulu. The genetic substructure of these populations has been assessed by studying the Y-chromosome and autosomal DNA resulting into a cluster of three specific groups: Tswana/Sotho, Nguni and Venda (Lane *et al.* 2002; Mitchell 2010). It is of clear interest to characterize the APOBEC3 gene polymorphisms existing in these various populations, since they may play a crucial role in the restriction and evolution of HIV-1.

In the current study, we characterized the genetic variability within the coding regions of A3D, A3F, A3G and A3H to document the level of diversity in samples obtained from HIV-1 positive individuals attending three HIV clinics in the Limpopo Province of Northern South Africa.

5.2 Hypothesis

There are variations in APOBEC3 genes in South African populations that could significantly affect viral replication, transmission, fitness and evolution.

5.3 Objective

The objective of the study is to characterize the genetic variability within the coding regions of A3D, A3F, A3G and A3H in order to determine the level of diversity in a South African population.

5.3.1 Specific objectives

5.3.1.1 To sequence the coding regions of A3D, A3F, A3G and A3H.
5.3.1.2 To characterize variation in the coding regions of A3D, A3F, A3G and A3H.
5.3.1.3 To compare APOBEC3 variation observed in the study population with other populations.

5.4 MATERIALS AND METHODS

5.4.1 Study population and DNA extraction

Characteristics of the study population and samples collection are described in chapter two, section 2.3.2.

5.4.2 Primer design, Polymerase Chain Reaction (PCR) to amplify A3D, A3F, A3G and A3H genes

Primer design and amplification of the APOBEC3 genes is detailed in chapter two, from section 2.3.3.1.3 to 2.3.3.1.4. Briefly, the Takara (LA) PCR Kit Ver. 2.1 for long DNA fragments amplification (Clontech) was used to amplify the complete gene in a 1st round 30 cycle PCR. The 1st round PCR products of each complete gene were used as templates in a nested PCR to generate amplicons consisting of the coding regions. The amplicons were purified using AMpure XP beads (Beckman coulter) and quantified using Qubit 3 with the dsDNA HS kit (Invitrogen). Amplicons per sample were pooled and normalized to 10ng using 10mM Tris elusion.

5.4.3 Fragmentation, tagmentation and addition of Illumina indices

The Tn5 tagmentation based protocol described in chapter two was used to fragment 1ng amplicon DNA products into smaller size fragments ranging from 35bp to 700bp, tagging them at 55^{o}C for 5 min with the Illumina sequencing adapters at both ends of each amplicon. Following this step, unique Illumina dual-index barcodes Index 1 (i7) and index 2 (i5) were added to the sample in a short PCR of 12 cycles, followed by the second AMpure XP bead purification. Using the full complement of Nextera XT indices, up to 96 samples were pooled for each run.

5.4.4 Library normalization, pooling and sequencing

The libraries were nomalized, pooled and sequenced following protocols from chapter two. Biological sample sheets were created in Basespace by giving each sample the appropriate index and setting up a sequencing run for the MiSeq. The run generated about 25 million reads/sequences per sample. All of the individual patient sequences generated and used in this study have been submitted to the NCBI Sequence Read

Archive (project number: PRJNA429751) and will be released and accessible using the following link; http://www.ncbi.nlm.nih.gov/bioproject/429751, pending manuscript acceptance (see sample accession numbers in appendix ii).

5.4.5 Demultiplexing and sequence quality control evaluation

Sequences were demultiplexed automatically on the MiSeq as part of the data processing steps and ends pairing. Fastq files were generated for each sample representing the two paired-end reads. Sequence quality was validated using the Galaxy NGS platform: QC and manipulation and fastQC program.

5.4.6 Sequence filtering, trimming, mapping and variant calling

Sequencing data quality, including the duplication rate, percent GC, and read quality was assessed by quality control tools for high throughput sequencing data (Andrews 2010; Ewels *et al.* 2016). After filtering low coverage samples, reads were aligned against the human genome with the BWA-MEM tool (Li 2013). Alignments were sorted, marked for duplicates, and indexed using SAMtools (Li *et al.* 2009), and variants were called using Freebayes, a Haplotype-based variant detection from short-read sequencing tool (Tan *et al.* 2015). Variant calls were normalized and decomposed with vt, a unified representation of genetic variants, and functionally annotated using SnpEff, a program for annotating and predicting the effects of single nucleotide polymorphisms (Cingolani *et al.* 2012). Comprehensive annotation and prioritization was performed using the GEMINI framework for Integrative Exploration of Genetic Variation and Genome Annotations (Paila *et al.* 2013). All further data manipulation and analysis was performed using R, a Language and Environment for Statistical Computing (R Development Core Team 2010).

5.4.7 Statistical analysis

5.4.7.1 Hardy-Weinberg Equilibrium (HWE)

All variants loci detected within the coding regions of APOBEC3 genes were tested for deviation from Hardy-Weinberg Equilibrium (HWE) using an excel HWE calculator and chi-square test with $P<0.05$ showing non-consistence with HWE (Court *et al*. 2012).

5.4.7.2 Pairwise linkage disequilibrium (LD) and Haplotype assignment

Pairwise linkage disequilibrium (LD) analysis between SNPs of each gene was done to test if they were in LD and haplotype assignments. Both the LD and the haplotypes were determined using the LDLink 3.0 web tool; exploring the LDmatrix, LDpair and LDhap modules (https://analysistools.nci.nih.gov/LDlink/?tab=home) detailed in chapter four, section 4.3.7. Variants of minor allele frequencies <5% and those not assigned in the dbSNP (novel SNPs) were excluded from LD determination and haplotype assignment

5.5 RESULTS

5.5.1 Single nucleotide polymorphisms (SNPs), detection of indels and verification

There is limited availability of APOBEC3 gene sequences from African populations, and when sequencing has been performed, it has generally been limited to A3G (Reddy *et al.* 2010). In this study, we applied next generation sequencing to determine variation in the coding exons of the APOBEC genes A3D, A3F, A3G and A3H in DNA from 192 HIV positive individuals residing in the Limpopo province of northern South Africa. The proteins expressed from these genes have all been shown to be capable of HIV restriction (Hultquist *et al.* 2011).

The A3D gene is 12.1kb long and has seven exons with exon 5 shown to display the most variation. Good quality A3D sequences after targeted DNA implication of the exons were successfully obtained for 165/192 subjects. In the DNA from these 165 individuals, 8 nonsynonymous and 2 synonymous changes were identified when compared to the GRCh37 build of the human genome (Table 5.1). Of the 165 subjects analyzed, 50.3% (83/165) were identified with nonsynonymous or synonymous changes in many positions in the coding region of the A3D gene, while no changes were detected in the remaining 49.7% (82/165). There were no insertions or deletions observed in A3D in the sequenced samples. Variant R248K was observed in 18.8% (31/165) of the patients, with 17.6% being heterozygotes. The R248K variant together with T316T was in strong linkage disequilibrium (LD) (D' =1, R^2= 0.122). Minor allele frequencies (MAF<5%) were observed in the case of L221L, C224Y, T238I and C320Y, therefore their LD profiles could not be calculated. Although no variants deviated from the Hardy-Weinberg Equilibrium (HWE) (P-value <0.05), R248K showed a tendency towards deviation with P-value =0.09. In addition, R6K, and T238I, variants that have not been reported elsewhere, were observed as heterozygotes in 9.7 and 4.8 % respectively (Table 5.1).

The A3F gene is 13.3kb long and contains seven exons, with the most variation observed in exon 4. The A3F exons were successfully amplified and sequenced from a total of 171/192 subjects. Synonymous or nonsynonymous changes were observed in 95.9% (164/171) of the subjects while 4.1% (7/171) had no change relative to the GRCh37 human genome build. In the 171 samples successfully sequenced, there were seven nonsynonymous changes (R48P, A78V, I87L, P97L, A108S, V231I and Y307C)

and seven synonymous changes (I117I, S118S, R143R, Y196Y, S229S, S327S and E245E). A108S and A78V were the most frequent nonsynonymous changes in A3F, found in 64.3% and 39.1% of the subjects respectively (Table 5.1). No insertions or deletions were observed for A3F in the sequenced samples. All A3F variants were in strong LD (D'= 1, R^2 range from 0.001- 0.899). A few variants (A108S, R143R and E245E) deviated from the HWE (P-value <0.05). The synonymous I117I mutation has not been reported previously (Table 5.1).

The A3G gene was the first one that was described as encoding an HIV restriction factor and it remains the most studied among Africans. The gene is 10.7 kb and has 8 exons. We successfully amplified A3G from 171/192 subjects. The most variation has been observed in exons 3, 6 and 7. A total of four nonsynonymous (H186R, R256H, Q275E and G363R) and four synonymous changes (S60S, A109A, F119F and L371L) were observed in A3G with the most frequent being H186R (53.8%) and Q275E (32.7%), Table 3. In total, nonsynonymous or synonymous changes were observed in 95.9% (164/171), whereas 4.1% (7/171) had no changes relative to the reference GRCh37 human genome. There were no insertions or deletions observed in this gene. Variants S60S, A109A, F119F H186R, R256H, and Q275E were in strong to partial LD (D'= 1, R2 range from 0-0.068). None of the variants deviated from the HWE (P-value > 0.05) (Table 5.1).

A3H is the shortest, but most polymorphic of the APOBEC3 genes we analyzed. It is 6.8kb in length (Table 1) and contains 5 exons, with the most variation in exons 1, 2 and 3. We observed nonsynonymous or synonymous changes in all the study subjects (175/175). We found 6 nonsynonymous (N15Δ, R18L, G105R, K121E, E140K and E178D) and 1 synonymous (T43T) changes. The N15Δ deletion was the only deletion observed and it occurred in 107 of 175 subjects (61.1%) either in a homozygous (57) or heterozygous (50) form. The G105R, K121E, E140K and E178D variants occurred mostly as homozygous forms in 90.3-100% of all subjects (Table 3). No insertions were found. All variants in this gene are in strong LD (D'=1, R^2 range from 0.203- 1). The N15Δ and R18L variants deviated from the HWE (P-value >0.05), (Table 5.1).

Table 5.1: APOBEC 3D, 3F, 3G and 3H nonsynonymous and synonymous changes, genotype frequencies and Hardy Weinberg Equilibrium calculations from the study population.

APOBEC 3D Nonsynonymous changes (n= 165)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
R6K (**NI**)	A**G**A→A**A**A Transition	17G/G 17G/A	1	149 (90.3) 16 (9.7)	P-value = 0.51 X^2 = 0.43
R97C **(rs75858538)**	**C**GC→**T**GC Transition	289 C/C 289 C/T 289 T/T	1	148 (89.7) 16 (9.7) 1 (0.6)	P-value = 0.44 X^2 = 0.59
L221R **(NI)**	C**T**G→C**G**G Transition	662T/T 662T/G	5	162 (98.2) 3 (1.8)	P-value = 0.9 X^2 = 0.01
C224Y **(rs772893975)**	T**G**T→T**A**T Transition	671G/G 671G/A	5	159 (96.4) 6 (3.6)	P-value = 0.812 X^2 = 0.056
T238A **(rs201709403)** T238I **(NI)**	**A**CA→**G**CA Transition **A**CA→A**T**A Transition	712A/A 712A/G 713A/T	5	142 (86.1) 15 (9.1) 8 (4.8)	P-value = 0.218 X^2 = 1.5
R248K **(rs61748819)**	A**G**G→A**A**G Transition	743 G/G 743 G/A	5	134 (81.2) 29 (17.6)	P-value = 0.09

		743 A/A		2 (1.2)	$X^2 = 0.76$
C320Y **(rs61999342)**	T**G**C→T**A**C Transition	959 G/G 959 G/A	6	164 (99.4) 1 (0.6)	P-value = 0.94 $X^2 = 0.006$
APOBEC3D Synonymous changes					
L221L **(rs769426665)**	CT**G** →CT**C** Transition	663G/G 663G/C	5	162 (98.2) 3 (1.8)	P-value = 0.97 $X^2 = 0.02$
T316T **(rs184448269)**	AC**C**→AC**T** Transition	948 C/C 948 C/T	6	158 (95.7) 7 (4.3)	P-value =0.84 $X^2 = 0.04$
No mutation				**82 (49.7)**	
APOBEC 3F Nonsynonymous changes (n= 171)					
R48P **(rs35053197)**	C**G**T→C**C**C Transversion	143 G/G 143 G/C	2	156 (91.2) 15 (8.8)	P-value = 0.54 $X^2 = 0.36$
A78V **(rs5750728)**	G**C**C→G**T**C Transition	233 C/C 233 C/T 233 T/T	4	108 (63.2) 59 (36.8) 4 (2.3)	P-value = 0.9 $X^2 = 0.04$
I87L **(rs146543452)**	**A**TC→**C**TC Transition	260 A/A 260 A/C	4	162 (94.7) 9 (5.3)	P-value = 0.72 $X^2 = 0.12$
P97L **(rs201939303)**	C**C**G→C**T**G Transition	290 C/C 290 C/T	4	168 (98.2) 3 (1.8)	P-value = 0.9 $X^2 = 0.04$
A108S **(rs2020390)**	**G**CT→**T**CT	322 G/G		61 (35.7)	P-value =

	Transition	322 G/T 322 T/T	4	92 (53.8) 18 (10.5)	0.05 $X^2 = 3.77$
V231I **(rs2076101)**	**G**TC→**A**TC Transition	898 G/G 898 G/A 898 A/A	5	139 (81.3) 30 (17.5) 2 (1.2)	P-value = 0.99 $X^2 = 0.07$
Y307C **(rs12157816)**	T**A**C→T**G**C Transition	920 A/A 920 A/G	6	156 (91.2) 15 (8.8)	P-value = 0.55 $X^2 = 0.36$
APOBEC3F Synonymous changes					
I117I **(NI)**	AT**C**→AT**T** Transition	351 C/C 351 C/T	4	169 (98.8) 2 (1.2)	P-value =0.94 $X^2 = 0.003$
S118S **(rs35928287)**	TC**C**→TC**T** Transition	354 C/C 354 C/T	4	125 (73.1) 46 (26.9)	P-value =0.04 $X^2 = 4.69$
R143R **(rs4821862)**	CG**C**→CG**T** Transition	429 C/C 429 C/T 429 T/T	4	23 (13.5) 97 (56.7) 51 (29.8)	P-value = 0.03 $X^2 = 4.69$
Y196Y **(rs765418322)**	TA**T**→TA**C** Transition	588 T/T 588 T/C 588 C/C	4	118 (69) 49 (28.7) 4 (2.3)	P-value = 0.7 $X^2 = 0.17$
S229S **(rs549550231)**	TC**A**→TC**G** Transition	894 A/A 894 A/G	5	169 (98.8) 2 (1.2)	P-value = 0.94 $X^2 = 0.005$
E245E **(rs113109079)**	GA**G**→GA**A** Transition	942 G/G 942 G/A 942 A/A	5	161 (94.2) 9 (5.2) 1 (0.6)	P-value = 0.04 $X^2 = 4.08$

S327S **(rs35895636)**	TC**C**→TC**T** Transition	981 C/C 981 C/T 981 T/T	5	143 (83.6) 25 (14.6) 3 (1.8)	P-value = 0.14 X^2 = 2.189
No mutation				**7 (4.1)**	
APOBEC 3G Nonsynonymous changes (n= 171)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
H186R **(rs8177832)**	C**A**C→C**G**C Transition	557 A/A 557 A/G 557 G/G	4	79 (46.2) 74 (43.3) 18 (10.5)	P-value = 0.89 X^2 = 0.02
R256H **(rs17000736)**	C**G**C→C**A**C Transition	767 G/G 767 G/A	6	167(97.7) 4 (2.3)	P-value = 0.97 X^2 = 0.00
Q275E **(rs17496046)**	**C**AG→**G**AG Transversion	823 C/C 823 C/G 823 G/G	6	115 (67.3) 50 (29.2) 6 (3.5)	P-value = 0.84 X^2 = 0.03
G363R **(rs148267053)**	**G**GA→**A**GA Transition	1087 G/G 1087 G/A	7	154 (90.1) 17 (9.9)	P-value = 0.5 X^2 = 0.3
APOBEC3G Synonymous changes					
S60S **(rs112603901)**	TC**C**→TC**T** Transition	180 C/C 180 C/T	3	152 (88.9) 19 (11.1)	P-value =0.59 X^2 = 0.4
A109A **(rs375760983)**	GC**C**→GC**T**	327 C/C		170 (99.4)	P-value =0.97

	Transition	327 C/T	3	1 (0.6)	$X^2 = 0.00$
F119F **(rs5757465)**	TT**T**→TT**C** Transition	357 T/T 357 T/C	3	170 (99.4) 1 (0.6)	P-value =0.97 $X^2 = 0.00$
L371L **(rs11545130)**	**C**TG→**T**TG Transition	1111 C/C 1111 C/T	7	164 (95.9) 7 (4.1)	P-value =0.8 $X^2 = 0.005$
No mutation				**7 (4.1)**	
APOBEC 3H Nonsynonymous changes (n= 175)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
N15Δ **(rs139292)**	**-CAA** Deletion	45 CAA/CAA 45 CAA/ Δ 45 Δ / Δ	1	68 (38.9) 50 (28.6) 57 (32.5)	P-value = 0.00 $X^2 = 31.8$
R18L **(rs139293)**	C**G**C→C**T**C Transvertion	53 G/G 53 G/T 53 T/T	1	156 (89.1) 15 (8.6) 4 (2.3)	P-value = 0.00 $X^2 = 15.9$
G105R **(rs139297)**	**G**GC→**C**GC Transvertion	313 G/G 313 G/C 313 C/C	2	0 4 (2.3) 171 (97.7)	P-value = 0.87 $X^2 = 0.022$
K121E **(rs139298)**	**A**AG→**G**AG Substitution	361 A/A 361 A/G	2	0 175 (100)	P-value = 0.59 $X^2 = 0.3$
E140K **(rs139300)**	**G**AG→**A**AG Transvertion	419 G/G 419 A/A	2	0 175 (100)	P-value = 0.47 $X^2 = 0.3$

E178D **(rs139302)**	**G**AG→GA**C** Transvertion	534 G/G 534 G/C 534 C/C	3	0 17 (9.7) 158 (90.3)	P-value = 0.497 $X^2 = 0.45$
APOBEC3H Synonymous changes					
T43T **(rs139294)**	AC**G**→AC**C** Transvertion	129 G/G 129 C/C	1	31 (17.7) 144 (82.3)	P-value =0.97 $X^2 = 0.00$
No mutation				**0**	

#Note NI= Not Identified (novel variants); Nucleotide in bold: Codon specific change; Rs= Reference single nucleotide polymorphism

In order to better understand the A3 genetic changes observed in each subject, all clusters of variation per gene were assigned into haplotypes as described in materials and methods and their frequencies calculated. These haplotypes were classified as either confirmed or unconfirmed based on the number of heterozygous variants. This classification was necessary due to the fact that NGS reads were short and we could not confirm if SNPs occurred in the same allele (Table 5.2). Nonsynonymous variants were considered and their genotypes (homozygous or heterozygous) were indicated. Low frequency variants (MAF<5%) were excluded from the haplotype assignment. Comparisons were made to the GRCh37 human genome whose combinations are represented as haplotypes in A3D, A3F and A3G (Table 5.2). We identified four haplotypes for A3D, five for A3F and four for A3G (Table 5.2). The A3D, A3F and A3G haplotypes were comprised of 3, 6 and 3 variants respectively. It is worth noting that only haplotypes for A3G and A3H have been described previously (Feng and Chelico 2011; Ooms *et al.* 2013; Refsland *et al.* 2014; K. *et al.* 2016), (Table 5.2). The seven identified haplotypes of A3H were recently described as having an impact on the genetic diversity of HIV-1 Vifs in the global pandemic (Ooms *et al.* 2010, 2013; Refsland *et al.* 2014). The seven known A3H haplotypes (I-VII) are combinations of 5 nucleotide changes located on exons 2, 3 and 4. Haplotypes II, V, and VII have been termed stable, because of the observed relatively long half-lives of the encoded proteins, enabling them to restrict HIV-1. The other four haplotypes (haplotypes I, III, IV, VI) have been termed unstable, since

the encoded protein half-lives have been shown to be short, resulting in complete loss of the ability to restrict HIV (Reddy *et al.* 2010; Ooms *et al.* 2013) In our subjects, we identified 3 haplotypes for A3H: the stable haplotype II (15N, 18R, 105R, 121E 178D), haplotype III (15 Δ , 18R, 105R, 121E, 178D) and haplotype IV (15 Δ, 18L, 105R, 121E, 178D) (Table 4) (Reddy *et al.* 2010; Wang *et al.* 2011; Ooms *et al.* 2013) The prevalence of these haplotypes was: 30.8% haplotype II; 56.6% haplotype III and 8.6% confirmed and 4% unconfirmed haplotype IV (Table 5.2). Thus the majority of the subjects in our patient population are not expressing ApoBec3H proteins that can restrict HIV.

Table 5.2: Haplotype formation and the frequencies with respect to the A3D, A3F, A3G and A3H

Confirmed APOBEC3D Haplotypes		
Variation	**Frequency (%)**	**Haplotype**
97R, 238T, 248R	82 (49.7)	Haplotype i
97C (hom)**, 238T, 248R** 97C (het)**, 238T, 248R**	1 (0.6) 16 (9.7)	Haplotype ii
97R, 238A (het)**, 248R**	11 (6.7)	Haplotype iIi
97R, 238T, 248K (hom) **97R, 238T,** 248K (het)	2 (1.2) 29 (17.6)	Haplotype iv
Minor variant frequency <5%	8 (4.8))	Not assigned
Others[a]	16 (9.7)	Not assigned
Unconfirmed APOBEC3D Haplotypes None		
Confirmed APOBEC3F Haplotypes		

48R, 78A, 87I, 108A, 231V, 307Y	7 (4.1)	Haplotype i
48R, 78V (hom)**, 87I,** 108S (hom)**,** 231I (hom)**, 307Y** **48R**, 78V (het)**, 87I,** 108S (hom)**,** 231I (hom)**, 307Y**	1 (0.6) 1 (0.6)	Haplotype ii
48R, 78A, 87I, 108S (hom)**, 231V, 307Y** **48R, 78A, 87I,** 108S (het)**, 231V, 307Y**	2 (1.2) 37 (21.6)	Haplotype iii
48R, 78A, 87I, 108A, 231V, 307C (het)	4 (2.3)	Haplotype iv
48R, 78V (hom)**, 87I,** 108S (hom)**, 231V, 30Y**	1 (0.6)	Haplotype vi
Minor variant frequency <5%	7 (4.1)	Not assigned
Others[a]	47 (27.5)	Not assigned
Unconfirmed APOBEC3F Haplotypes		
48R, 78V (het)**, 87I,** 108S (het)**,** 231I (het)**, 307Y**	30 (17.5)	Haplotype ii
48P (het)**, 78A, 87I,** 108S (het)**, 231V, 307Y**	9 (5.3)	Haplotype v
48R, 78V (het)**, 87I,** 108S (het)**, 231V, 307Y**	25 (14.6)	Haplotype vi
Confirmed APOBEC3G Haplotypes		
186H, 275Q, 363G	7 (4.1)	Haplotype i
186R (hom)**, 275Q, 363G** 186R (het)**, 275Q, 363G**	20 (11.7) 42 (24.6)	Haplotype ii
186H, 275E (hom)**, 363G** **186H,** 275E (het)**, 363G**	7 (4.1) 15 (8.8)	Haplotype iii

186H, 275Q, 363R (het)	12 (7)	Haplotype iv
Minor variant frequency <5%	12 (7)	Not assigned
Others[a]	56 (32.7)	Not assigned
Unconfirmed APOBEC3G Haplotypes None		
Confirmed APOBEC3H Haplotypes[b]		
15N, 18R, 105R (hom)**,** 121E (hom)**,** 178D (hom)	54 (30.8)	Haplotype ii
15Δ (hom)**, 18R,** 105R (hom)**,** 121E (hom)**,** 178E (hom) 15Δ (het)**, 18R,** 105R (hom)**,** 121E (hom)**,** 178D (hom)	50 (28.6) 49 (28)	Haplotype iii
15Δ**,** (hom), 18L (hom)**,** 105R (hom)**,** 121E (hom)**,** 178D (hom) 15Δ**,** (hom), 18L (het)**,** 105R (hom)**,** 121E (hom)**,** 178D (hom)	7 (4) 8 (4.6)	Haplotype iv
Unconfirmed APOBEC3H Haplotypes[b]		
15Δ**,** (het), 18L (het)**,** 105R (hom)**,** 121E (hom)**,** 178D (hom)	7 (4)	Haplotype iv

#Note

[a] Refers to the haplotypes with synonymous changes and those of novel SNPs (not reported on the dpSNP)

[b]A3H haplotypes were determined using previous classification from references (Wang *et al*. 2011; Ooms *et al*. 2013; Refsland *et al*. 2016)

hom=homozygous; het=heterozygous

Bold defines variants that are different from those listed in haplotype I in each gene.

Haplotypes are unconfirmed in our population due to more than 1 heterozygous SNPs in the cluster

5.5.2 Allele frequencies and their comparison with other populations

We next compared the nonsynonymous and synonymous variant frequencies in the South African population in our study to previously reported variant frequencies in the following populations: African (AFR), East Asian (EAS), European (EUR), Ad Mixed American (AMR), and South Asian (SAS), as reported in the 1000 Genome Project phase III, the HapMap project (NCBI), the dsSNP database and the Ensembl genome browser (Table 5.3). Our data show that with the exception of, V231I, (A3F) and F119F (A3G), all the variants identified showed a higher frequency than previously reported for the overall AFR population (Table 5.3). In a previous study by Duggal and colleagues comparing A3 variations between Africans, Asian and Europeans, nonsynonymous variation in A3D (R97C, R248K); A3F (A108S, V231I, Y307C); A3G (H186R, E275Q) and A3H (15Δ, R18L, R105G, E121E/K, E178D) were reported (Reddy *et al.* 2010). Our data suggest that several variants occur more frequently in our South African population than in the "African" population previously studied (Duggal *et al.* 2013). This included R97C and R248K in A3D; A108S in A3F; H186R, Q275E in A3G and N15Δ, R18L, G105R, K121E and E178D in A3H (Table 5.3). Of particular interest was the A3F A108S variant, which occurs in about 80% of Asians and has been reported in only 30% of Africans (Duggal *et al.* 2013). We found this variant in 64.3% of our subjects, more than twice the prevalence previously reported among Africans by Duggal and colleagues (Duggal *et al.* 2013).

Overall, the EAS, EUR, AMR and SAS populations showed a higher level of A3 conservation than our study population. For example, the A3D sequences in these populations were more closely related to the reference GRCh37 human genome (98-100%). In the case of A3F and A3G, the percent identity was between 92-100%, with the exception of the A3F variants A78V, A108S, V231I, R143R and the A3G variant F119F (Table 5). The percentage identities among the A3H variants were highly variable, with the N15**Δ** variant clearly present in higher frequency in our population compared to the others. (Table 5.3).

Table 5.3: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population and total populations from East Asian (**EAS**), European (**EUR**), African (**AFR**), Ad Mixed American (**AMR**) and South Asian (**SAS**).

APOBEC 3D Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
R6K (**NI**)	G (**R**) A (**K**)	90.3 9.7	NA	NA	NA	NA	NA
R97C **(rs758585538)**	C (**R**) T (**C**)	89.7 10.3	100 0	100 0	96.6 3.4	100 0	100 0
L221R **(NI)**	T (**L**) G (**R**)	98.2 1.8	NA	NA	NA	NA	NA
C224Y **(rs772893975)**	G (**C**) A (**Y**)	96.4 3.6	100 0	100 0	100 0	100 0	100 0
T238A **(rs201709403)**	A (**T**) G (**A**)	86.1 9.1	100 0	100 0	100 0.001	100 0	100 0
T238I (NI)	A (**T**) T (**I**)	86.1 4.8	NA	NA	NA	NA	NA
R248K **(rs61748819)**	G (**R**) A (**K**)	81.2 18.8	100 0	100 0	89. 11	98.9 4	100 0
C320Y **(rs61999342)**	G (**C**) A (**Y**)	99.4 0.6	NA	NA	NA	NA	NA
APOBEC3D synonymous allele frequencies (%)							
L221L **(rs769426665)**	G (**L**) C (**L**)	98.2 1.8	100 0	100 0	100 0	100 0	100 0

T316T **(rs184448269)**	C (**T**) T (**T**)	95.7 4.3	100 0	100 0	98.8 1.2	99.6 4.3	100 0
APOBEC 3F Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
R48P **(rs35053197)**	G (**R**) C (**P**)	91.2 8.8	100 0	NA	97 3	99 1	99 1
A78V **(rs5750728)**	C (**A**) T (**V**)	63.2 36.8	29 79.3	58.8 49.1	79.8 22	37.6 62.4	39.1 60.8
I87L **(rs114704208)**	A (**I**) C (**L**)	94.7 5.3	100 0	100 0	99 1	100 0	100 0
P97L **(rs201939303)**	C (**P**) T (**L**)	98.2 1.8	NA	NA	NA	NA	NA
A108S **(rs2020390)**	G (**A**) T (**S**)	35.7 64.3	29 71	51 49	68 32	37 63	40 60
V231I **(2076101)**	G (**V**) A (**I**)	81.3 18.7	71 29	51 49	81 19	38 62	39 61
Y307C **(rs12157816)**	A (**Y**) G (**C**)	91.2 8.8	100 0	98 2	97 3	98 2	100 0
APOBEC3F synonymous allele frequencies (%)							
I117I **(NI)**	C (**I**) T (**I**)	98.8 1.2	NA	NA	NA	NA	NA
S118S **(rs35928287)**	C (**S**) T (**S**)	73.1 26.9	100 0	100 0	97 3	100 0	100 0
R143R **(rs4821862)**	C (**R**)	13.5	29	51	45	36	39

	T (**R**)	86.5	71	49	55	64	61
Y196Y **(rs765418322)**	T (**Y**) C (**Y**)	69 31	100 0	100 0	100 0	100 0	100 0
S229S **(549550231)**	A (**S**) T (**S**)	98.8 1.2	NA	NA	NA	NA	NA
E245E **(113109079)**	G (**E**) A (**E**)	94.2 5.8	100 0	100 0	99 1	100 0	100 0
S327S **(35895636)**	C (**S**) T (**S**)	83.6 16.4	100 0	100 0	99 2	100 0	100 0
APOBEC 3G Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
H186R **(rs8177832)**	A (**H**) G (**R**)	46.2 53.8	92.7 7.2	97.2 2.98	56.9 43	92.5 7.5	99.1 0.82
R256H **(rs17000736)**	G (**R**) A (**H**)	97.7 2.3	100 0	100 0	98.6 1.44	100 0	100 0
Q275E **(rs17496046)**	C (**Q**) G (**E**)	67.3 32.7	97.3 2.68	94.6 5.3	87.5 12.5	96 4	98.7 1.3
G363R **(rs148267053)**	G (**G**) A (**R**)	90.1 9.9	100 0	100 0	98.6 1.4	99.9 1	100 0
APOBEC3G synonymous allele frequencies (%)							
S60S **(rs112603901)**	C (**S**) T (**S**)	88.9 11.1	100 0	100 0	99.7 0.7	100 0	100 0
A109A **(rs375760983)**	C (**A**) T (**A**)	99.4 0.6	100 0	100 0	NA	NA	NA
F119F **(rs5757465)**	T (**F**)	99.4	77.6	55.3	97.1	60.2	55.5

	C (**F**)	0.6	22.4	44.7	28.7	39.7	44.5
L371L **(rs11545130)**	C (**L**) T (**L**)	95.9 4.1	97 3	100 0	100 0	100 0	100 0
APOBEC 3H Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
N15Δ **(rs139292)**	CAA(**N**) Δ	38.9 61.1	69 31	72 28	74 26	66 34	60 40
R18L **(rs139293)**	G (**R**) T (**L**)	89.1 10.9	84.1 15.9	70.7 29.3	93.3 0.6	75.8 24.2	69.4 30.6
G105R **(rs139297)**	C (**G**) G (**R**)	0 100	33.1 66.9	46.3 53.7	87.5 12.5	38.3 61.7	42.9 57.1
K121E **(rs139298)**	C (**K**) G (**E**)	0 100	33.1 66.9	46.3 53.7	84.7 15.3	34.6 65.4	43.1 56.9
E140K **(rs139300)**	G (**E**) A (**K**)	0 100	0 100	0 100	0 100	0 100	0 100
E178D **(rs139302)**	G (**E**) C (**D**)	0 100	83.3 66.6	45.4 54.6	82.7 17.3	37.7 62.3	43.9 56.1
APOBEC3H synonymous allele frequencies (%)							
T43T **(rs139294)**	G (**T**) C (**T**)	17.7 82.3	33.3 66.6	45.4 54.6	82.7 17.3	37.7 62.3	43.8 56.2

Note: NI= Not Identified

(NA)= No data available

The term “Africans” has been loosely used to describe datasets generated from different parts of the African continent. To provide a more accurate comparison, we compared the variants detected in our study with previous datasets derived from other more specific studies in African populations or involving people of African descent (Table 5.4). These included Americans of African Ancestry in USA (ASW); African Caribbeans in Barbados (ACB); Gambians in the Western Gambia (GWD); Esan in Nigeria (ESN); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL) and Yoruba in Ibadan, Nigeria (YRI). We noticed significantly higher level of single nucleotide changes in our population for the following variants: R48P, A108S, R143R, S327S in A3F; Q275E, G363R in A3G) and G105R, K121E, E178D in A3H (Table 5.4). However, in most cases, the variant allele frequencies in A3D, A3F, A3G and A3H were closer (Table 5.4). Notably, the variant frequencies of R97C and R248K in A3D are almost close to frequencies in ASW and ESN and MSL respectively, whereas the frequency of R48P in A3F is closer to that reported in, ESN, LWK and YRI. The frequency of I87L was close to that observed in GWD, LWK and MSL. The frequency of V231I was close to that observed in all Africans, whereas the frequency of Y307C was close to the frequencies observed in ACB and YRI. The frequency of E245E was close to that observed in ASW, GWD, MSL and YRI. For A3G variants, the frequency observed for H186R is similar to that observed in ESN and MSL. The frequency of R256H and L371L were similar among all Africans. In the case of A3H, information about the N15Δ variant was not available from any of the previous African studies, while the R18L variant was similar in prevalence among all Africans (Table 5.4).

Table 5.4: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population and other Africa population, which comprises of the African Carribeans in Barbados (**ACB**), Americans of African Ancestry in USA (**ASW**), Esan in Nigeria (**ESN**), Gambian in the Western Gambia (**GWD**), Luhya in Webuye, Kenya (**LWK**), Mende in Seirra Leone (**MSL**), Yoruba in ibadan, Nigeria (**YRI**).

APOBEC 3D Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
R6K (**NI)**	G (**R**) A (**K**)	90.3 9.7	NA	NA	NA	NA	NA	NA	NA
R97C **(rs758585538)**	C (**R**) T (**C**)	89.7 10.3	97 3	94 6	97 3	98 2	94 6	98 2	97 3
L221R **(NI)**	T (**L**) G (**R**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
C224Y **(rs772893975)**	G (**C**) A (**Y**)	96.4 3.6	NA	NA	NA	NA	NA	NA	NA
T238A **(rs201709403)**	A (**T**) G (**A**)	86.1 9.1	100 0	99 1	100 0	100 0	100 0	100 0	100 0
T238I (NI)	A (**T**) T (**I**)	86.1 4.8	NA	NA	NA	NA	NA	NA	NA
R248K **(rs61748819)**	G (**R**) A (**K**)	81.2 18.8	91 9	96 4	86 14	92 8	87 13	86 14	87 13
C320Y	G	99.4	NA	NA	NA	NA	NA	NA	NA

(rs61999542)	(**C**) A (**Y**)	0.6							
APOBEC3D synonymous allele frequencies (%)									
L221L **(rs769426665)**	G (**L**) C (**L**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
T316T **(rs184448269)**	C (**T**) T (**T**)	95.7 4.3	99 1	99 1	98 2	99 1	99 1	98 2	99 1
APOBEC 3F Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
R48P **(rs35053197)**	G (**R**) C (**P**)	91.2 8.8	98 2	98 2	95 5	99 1	95 5	98 2	95 5
A78V **(rs5750728)**	C (**A**) T (**V**)	63.2 36.8	NA	NA	NA	NA	NA	NA	NA
I87L **(rs146543452)**	A (**I**) C (**L**)	99.4 0.6	100 0	100 0	100 0	98 2	99 1	99 1	100 0
P97L **(rs201939303)**	C (**P**) T (**L**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
A108S **(rs2020390)**	G (**A**) T (**S**)	35.7 64.3	67 33	66 34	69 31	62 38	65 35	76 24	70 30
V231I **(2076101)**	G (**V**) A (**I**)	81.3 18.7	79 21	73 27	84 16	78 22	80 20	85 15	87 13
Y307C **(rs12157816)**	A (**Y**) G	91.2	95	98	96	98	98	96	95

	(**C**)	8.8	5	2	4	2	2	4	5
APOBEC3F synonymous allele frequencies (%)									
I117I **(NI)**	C (**I**) T (**I**)	98.8 1.2	NA	NA	NA	NA	NA	NA	NA
S118S **(rs35928287)**	C (**S**) T (**S**)	73.1 26.9	NA	NA	NA	NA	NA	NA	NA
R143R **(rs4821862)**	C (**R**) T (**R**)	13.5 86.5	45 55	45 55	39 61	50 50	45 55	46 54	47 53
Y196Y **(rs765418322)**	T (**Y**) C (**Y**)	69 31	NA	NA	NA	NA	NA	NA	NA
S229S **(549550231)**	A (**S**) T (**S**)	98.8 1.2	NA	NA	NA	NA	NA	NA	NA
E245E **(113109079)**	G (**E**) A (**E**)	94.2 5.8	99 1	98 2	100 0	98 2	99 1	98 2	98 2
S327S **(35895636)**	C (**S**) T (**S**)	83.6 16.4	98 2	99 1	96 4	100 0	97 3	100 0	98 2
APOBEC 3G Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
H186R **(rs8177832)**	A (**H**) G (**R**)	46.2 53.8	56 44	75 25	49 51	57 43	68 32	49 51	52 48
R256H **(rs17000736)**	G (**R**) A (**H**)	97.7 2.3	98 2	98 2	99 1	97 3	98 2	99 1	99 1

Q275E **(rs17496046)**	C (**Q**) G (**E**)	67.3 32.7	90 10	91 9	86 14	87 13	83 17	91 9	87 13
G363R **(rs148267053)**	G (**G**) A (**R**)	90.1 9.9	98 2	99 1	100 0	98 2	99 1	98 2	98 2
APOBEC3G synonymous allele frequencies (%)									
S60S **(rs112603901)**	C (**S**) T (**S**)	88.9 11.1	99 1	98 2	100 0	100 0	99 1	100 0	100 0
A109A **(rs375760983)**	C (**A**) T (**A**)	99.4 0.6	NA	NA	NA	NA	NA	NA	NA
F119F **(rs5757465)**	T (**F**) C (**F**)	99.4 0.6	93 7	89 11	100 0	98 2	98 2	99 1	99 1
L371L **(rs11545130)**	C (**L**) T (**L**)	95.9 4.1	98 2	98 2	97 3	98 2	95 5	98 2	95 5
APOBEC 3H Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
N15Δ **(rs139292)**	CAA (**N**) Δ	38.9 61.1	NA	NA	NA	NA	NA	NA	NA
R18L **(rs139293)**	G (**R**) T (**L**)	89.1 10.9	93 7	87 13	93 7	94 6	94 6	93 7	96 4
G105R **(rs139297)**	C (**G**) G (**R**)	0 100	85 15	75 25	90 10	87 13	91 9	89 11	91 9

K121E **(rs139298/99)**	A (**K**) G (**E**)	0 100	85 15	75 25	90 10	87 13	91 9	89 11	91 9
E140K **(rs139300)**	G (**E**) A (**K**)	0 100	100 0	100 0	100 0	100 0	100 0	100 0	100 0
E178D **(rs139302)**	G (**E**) C (**D**)	0 100	17 83	29 71	11 89	16 84	16 84	13 87	11 89
APOBEC3H synonymous allele frequencies (%)									
T43T **(rs139294)**	G (**T**) C (**T**)	17.7 82.3	22 78	30 70	16 84	16 84	13 87	17 83	13 87

Note: NI= Not identified

(NA)= No data available

5.6 DISCUSSION AND CONCLUSION

In this study we characterized SNPs and indels within the coding exons of several human APOBEC3 genes (A3D, A3F, A3G and A3H) to document the level of diversity in these genes in a diverse South African population residing in the Limpopo Province in Northern South Africa. We observed a high level of A3 diversity and a higher prevalence of certain variants than has previously been observed in other African populations. Interestingly, some of these variants have previously been linked to HIV disease progression (An *et al.* 2004; Reddy *et al.* 2010; Compaore *et al.* 2016). The use of next generation sequencing also allowed the identification of genetic variants that were not previously identified using methods such as TaqMan assays, restriction fragment length polymorphism (RFLP) or Sanger sequencing (Reddy *et al.* 2010).

Common variants in APOBEC3 genes have been intensively studied and many have been found to have differential effects on antiviral activity (An *et al.* 2004, 2016; Reddy *et al.* 2010; Duggal *et al.* 2013; Compaore *et al.* 2016). For example, the variants R97C and R248K in A3D were shown to moderately decrease antiviral activity (Duggal *et al.* 2013). In contrast, the A3F variants A108S, V231I and Y307C have been reported to have potent antiviral activity against HIV-1 ΔVif strains (Mulder *et al.* 2010; Duggal *et al.* 2011). Among the A3F variants, A108S was the most interesting in our study, as its prevalence was twice that reported previously for the “African” population (Table 5.4), but very similar to reports from EAS, EUR, AMR and SAS populations (Table 5.3). It will be of clear interest to perform additional studies to determine the prevalence of this variant in other regions of South Africa. SNPs in A3G can alter its antiviral activity and sometimes enhance the rate of HIV-1 disease progression, as reported in a cohort of HIV-1 subtype C infected South African women and a US based cohort of African Americans (An *et al.* 2004; Reddy *et al.* 2010). In particular, the H186R variant has previously been associated with more rapid decline in CD4+ cells and accelerated disease progression (An *et al.* 2004; Reddy *et al.* 2010; Compaore *et al.* 2016). Our study show that this variant is similar in prevalence in our population as in some other African populations (Table 5.4), but significantly higher than reported in a previous study performed in Durban, South Africa (Reddy *et al.* 2010). This difference could be explained either by differences in the ethnicity of the study population or the genotyping method used, since the previous study used TaqMan genotyping rather than the NGS methodology used in our study.

Recent studies have shown A3H as the most polymorphic member of the A3 family. The A3H variants (15Δ, R18L, G105R, K121E, E178D) which make up 7 different haplotypes have been functionally described in other studies, showing varying protein expression and stability (Harari *et al.* 2009; Tan *et al.* 2009; Li *et al.* 2010; Ooms *et al.* 2010; Zhen *et al.* 2010, 2012; Wang *et al.* 2011). Data from the 1000 genome project suggest that stable A3H haplotypes (II, V and VII) predominate in Africa while unstable haplotypes (I, III, IV, VI) are more prevalent in Asia (Refsland *et al.* 2014), Interestingly, the unstable A3H haplotypes III and IV (which encode complete loss-of-function proteins) were unexpectedly dominant among our study population. This can be attributed mainly to the high prevalence (69.2%) of the deletion at amino acid residue 15 (Tables 5.2, 5.3 and 5.4). This prevalence is inversely proportional to that reported in previous studies of Africans in which stable A3H haplotypes were dominant (56%) (Refsland *et al.* 2014). Data from two recent studies illustrate that stable A3H haplotypes may function as contemporary HIV-1 restriction factors, contributing to limiting viral replication and rates of transmission (Ooms *et al.* 2013; Refsland *et al.* 2014). It is unclear what role, if any, the unstable A3H haplotype III and IV may play in the high prevalence and transmission of HIV-1 in Limpopo.

Because HIV-1 Vif acts as an antagonist to APOBEC proteins including A3H, we speculate that the distribution of stable versus unstable A3H haplotypes in our study might also influence Vif variation in HIV in our study population. Studies performed in primary CD4+ lymphocytes have shown that HIV-1 Vif variants with certain amino acid residues (F39 and H48), known as hyper Vifs, are better capable of neutralizing stable A3H genotypes, implying that HIV-1 Vif might adapt to the A3H haplotype in a particular population (Refsland *et al.* 2014). We are presently analyzing HIV-1 Vif sequences from our study subjects in order to determine a possible correlation between the A3H haplotypes and HIV-1 Vif genetic variation in this rural area of South Africa.

One limitation of our study was that all the subjects were HIV infected and were mostly at the chronic stage of infection. It is possible that HIV-1 negative subjects would present a different A3 profile. If this is the case, it could imply that these A3 genotypes either alone or in combination might influence HIV transmission. This is thus worthy of further studies. Although we did not quantify the A3 mRNA levels in this study, reports comparing HIV-1 non-controllers versus long-term non-progressors (LTNP) have shown

that LTNPs express a higher level of A3G and A3F proteins (Jin, X., Brooks, A., Chen, H., Bennett, R., Reichman 2005) .

In conclusion, we have shown that significant A3 variation exists among HIV patients in an ethnically diverse population in Northern South Africa, by providing extensive data for 4 different A3 genes that are known to restrict HIV infection, but have previously only been sparsely studied in African populations. Our NGS results provide a baseline for future studies that could functionally characterize the SNPs identified in the APOBEC proteins in this population and specifically analyze how they affect restriction of HIV replication and Vif function. Such studies will serve to increase our understanding of how the APOBEC3 landscape might have shaped the HIV epidemic in Northern South Africa.

5.7 REFERENCES

Altshuler, D., Lander, E., Ambrogio, L. (2010) 'A map of human genome variation from population scale sequencing', *Nature*, 476(7319), 1061–1073.

An, P., Bleiber, G., Duggal, P., Nelson, G., May, M., Mangeat, B., Alobwede, I., Trono, D., Vlahov, D., Donfield, S., Goedert, J.J., Phair, J., Buchbinder, S., O'Brien, S.J., Telenti, A., Winkler, C.A. (2004) 'APOBEC3G Genetic Variants and Their Influence on the Progression to AIDS', *Journal of Virology*, 78(20), 11070–11076.

An, P., Penugonda, S., Thorball, C.W., Bartha, I., Goedert, J.J., Donfield, S., Buchbinder, S., Binns-Roemer, E., Kirk, G.D., Zhang, W., Fellay, J., Yu, X.F., Winkler, C.A. (2016) 'Role of APOBEC3F Gene Variation in HIV-1 Disease Progression and Pneumocystis Pneumonia', *PLoS Genetics*, 12(3).

Andrews, S. (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data. [online], *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/*.

Batzer, M.A., Deininger, P.L. (2002) 'Alu repeats and human genomic diversity', *Nature Reviews Genetics*.

Cavalli-Sforza, L.L. (1991) 'Genes, People and Languages', *Scientific American*, 265(5), 104–110.

Chiu, Y.-L., Greene, W.C. (2008) 'The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements', *Annual Review of Immunology*, 26(1), 317–353.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M. (2012) 'A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3', *Fly*, 6(2), 80–92.

Compaore, T.R., Soubeiga, S.T., Ouattara, A.K., Obiri-Yeboah, D., Tchelougou, D., Maiga, M., Assih, M., Bisseye, C., Bakouan, D., Compaore, I.P., Dembele, A., Martinson, J., Simpore, J. (2016) 'APOBEC3G variants and protection against HIV-1 infection in Burkina Faso', *PLoS ONE*, 11(1).

Conrad, D.F., Hurles, M.E. (2007) 'The population genetics of structural variation',

Nature Genetics, 39(7S), S30–S36.

Court MH, M.H. (n.d.) Court's (2005–2008) Online Calculator. Tuft University Web Site. [online], *2012*.

Dang, Y., Wang, X., Esselman, W.J., Zheng, Y.-H. (2006) 'Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family.', *Journal of virology*, 80(21), 10522–10533.

Duggal, N.K., Fu, W., Akey, J.M., Emerman, M. (2013) 'Identification and antiviral activity of common polymorphisms in the APOBEC3 locus in human populations', *Virology*, 443(2), 329–337.

Duggal, N.K., Malik, H.S., Emerman, M. (2011) 'The Breadth of Antiviral Activity of Apobec3DE in Chimpanzees Has Been Driven by Positive Selection', *Journal of Virology*, 85(21), 11361–11371.

Ewels, P., Magnusson, M., Lundin, S., Käller, M. (2016) 'MultiQC: Summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), 3047–3048.

Feng, Y., Chelico, L. (2011) 'Intensity of deoxycytidine deamination of HIV-1 proviral DNA by the retroviral restriction factor APOBEC3G is mediated by the noncatalytic domain', *Journal of Biological Chemistry*, 286(13), 11415–11426.

Harari, A., Ooms, M., Mulder, L.C.F., Simon, V. (2009) 'Polymorphisms and Splice Variants Influence the Antiretroviral Activity of Human APOBEC3H', *Journal of Virology*, 83(1), 295–303.

Harris, R.S., Liddament, M.T. (2004) 'Retroviral restriction by APOBEC proteins', *Nature Reviews Immunology*.

Hultquist, J.F., Lengyel, J.A., Refsland, E.W., LaRue, R.S., Lackey, L., Brown, W.L., Harris, R.S. (2011) 'Human and Rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H Demonstrate a Conserved Capacity To Restrict Vif-Deficient HIV-1', *Journal of Virology*, 85(21), 11220–11234.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C.,

Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., Van De Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A., Singleton, A.B. (2008) 'Genotype, haplotype and copy-number variation in worldwide human populations', *Nature*, 451(7181), 998–1003.

Jin, X., Brooks, A., Chen, H., Bennett, R., Reichman, R.& S.H. (2005) 'APOBEC3G/CEM15 (hA3G) mRNA levels associate inversely with human immunodeficiency virus viremia.', *Journal of viriology*, 79, 11513–11516.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T., Batzer, M.A. (2000) 'The Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data', *The American Journal of Human Genetics*, 66(3), 979–988.

K., M., K., D., D., M., C., M., B., S.-P. (2016) 'Evaluating the contribution of APOBEC3G haplotypes on influencing HIV infection in a Zimbabwean paediatric population', *South African Medical Journal*.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M.E., Jonker, E., Freeman, C., Young, L., Morar, B., Toffie, L. (2002) 'Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies', *American Journal of Physical Anthropology*, 119(2), 175–185.

Li, H. (2013) 'Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.'

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), 2078–2079.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., Myers, R.M. (2008) 'Worldwide human relationships inferred from genome-wide patterns of variation', *Science*, 319(5866), 1100–1104.

Li, M.M.H., Wu, L.I., Emerman, M. (2010) 'The range of human APOBEC3H sensitivity to lentiviral Vif proteins.', *Journal of virology*, 84(1), 88–95.

Machiela, M.J., Chanock, S.J. (2015) 'LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants', *Bioinformatics*, 31(21), 3555–3557.

Mitchell, P. (2010) 'Genetics and southern African prehistory: An archaeological view', *Journal of Anthropological Sciences*.

Mulder, L.C.F., Ooms, M., Majdak, S., Smedresman, J., Linscheid, C., Harari, A., Kunz, A., Simon, V. (2010) 'Moderate influence of human APOBEC3F on HIV-1 replication in primary lymphocytes', *J Virol*, 84(18), 9613–9617.

OhAinle, M., Kerns, J.A., Li, M.M.H., Malik, H.S., Emerman, M. (2008) 'Antiretroelement Activity of APOBEC3H Was Lost Twice in Recent Human Evolution', *Cell Host and Microbe*, 4(3), 249–259.

Ooms, M., Brayton, B., Letko, M., Maio, S.M., Pilcher, C.D., Hecht, F.M., Barbour, J.D., Simon, V. (2013) 'HIV-1 Vif adaptation to human APOBEC3H haplotypes', *Cell Host and Microbe*, 14(4), 411–421.

Ooms, M., Majdak, S., Seibert, C.W., Harari, A., Simon, V. (2010) 'The Localization of APOBEC3H Variants in HIV-1 Virions Determines Their Antiviral Activity', *Journal of Virology*, 84(16), 7961–7969.

Paila, U., Chapman, B.A., Kirchner, R., Quinlan, A.R. (2013) 'GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations', *PLoS Computational Biology*, 9(7).

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R. (2014) 'Full-length RNA-seq from single cells using Smart-seq2', *Nature Protocols*, 9(1), 171–181.

R Development Core Team (2010) 'R: A Language and Environment for Statistical Computing', *R Foundation for Statistical Computing Vienna Austria*.

Reddy, K., Winkler, C.A., Werner, L., Mlisana, K., Abdool Karim, S.S., Ndung'U, T. (2010) 'Apobec3g expression is dysregulated in primary hiv-1 infection and polymorphic variants influence cd4+ t-cell counts and plasma viral load', *AIDS*, 24(2), 195–204.

Refsland, E.W., Hultquist, J.F., Harris, R.S. (2012) 'Endogenous origins of HIV-1 G-to-A hypermutation and restriction in the nonpermissive T cell line CEM2n', *PLoS Pathogens*, 8(7), 39.

Refsland, E.W., Hultquist, J.F., Luengas, E.M., Ikeda, T., Shaban, N.M., Law, E.K., Brown, W.L., Reilly, C., Emerman, M., Harris, R.S. (2014) 'Natural Polymorphisms in Human APOBEC3H and HIV-1 Vif Combine in Primary T Lymphocytes to Affect Viral G-to-A Mutation Levels and Infectivity', *PLoS Genetics*, 10(11).

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., Foster, M.W., Sharp, R.R., Cann, H.M., Lewontin, R.C., Latter, B.D.H., Barbujani, G., Magagni, A., Minch, E., Cavalli-Sforza, L.L., Jorde, L.B., Brown, R.A., Armelagos, G.J., Romualdi, C., Bowcock, A.M., Mountain, J.L., Cavalli-Sforza, L.L., Pritchard, J.K., Stephens, M., Donnelly, P., Qamar, R., Rosenberg, N.A., Weber, J.L., Broman, K.W., Rogers, A.R., Jorde, L.B., Urbanek, M., Goldman, D., Long, J.C., Wilson, J.F., Thomas, D.C., Witte, J.S., Wacholder, S., Rothman, N., Caporaso, N., Pritchard, J.K., Donnelly, P., Ptak, S.E., Przeworski, M., Jin, L., Chakraborty, R., Calafell, F. (2002) 'Genetic structure of human populations.', *Science (New York, N.Y.)*, 298(5602), 2381–5.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H. (2002) 'Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein', *Nature*, 418(6898), 646–650.

Tan, A., Abecasis, G.R., Kang, H.M. (2015) 'Unified representation of genetic variants', *Bioinformatics*, 31(13), 2202–2204.

Tan, L., Sarkis, P.T.N., Wang, T., Tian, C., Yu, X.-F. (2009) 'Sole copy of Z2-type human cytidine deaminase APOBEC3H has inhibitory activity against retrotransposons and HIV-1.', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 23(1), 279–87.

Tishkoff, S.A., Williams, S.M. (2002) 'Genetic analysis of African populations: human evolution and complex disease', *Nat Rev Genet*, 3(8), 611–621.

Wang, X., Abudu, A., Son, S., Dang, Y., Venta, P.J., Zheng, Y.-H. (2011) 'Analysis of human APOBEC3H haplotypes and anti-human immunodeficiency virus type 1

activity.', *Journal of virology*, 85(7), 3142–52.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., Gao, L. (2003) 'The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA', *Nature*, 424(6944), 94–98.

Zhen, A., Du, J., Zhou, X., Xiong, Y., Yu, X.F. (2012) 'Reduced APOBEC3H variant anti-viral activities are associated with altered RNA binding activities', *PLoS ONE*, 7(7).

Zhen, A., Wang, T., Zhao, K., Xiong, Y., Yu, X.-F. (2010) 'A single amino acid difference in human APOBEC3H variants determines HIV-1 Vif sensitivity.', *Journal of virology*, 84(4), 1902–11.

CHAPTER SIX

Genetic Characterization of HIV-1 *Vif* and Correlation with APOBEC3 Variation in Chronically Infected South African Treatment Experienced Patients

ABSTRACT

Background: The viral infectivity factor (Vif) protein of HIV-1 is essential for viral replication and has the ability to counteract the restriction of several human APOBEC3 proteins. Polymorphisms in *vif* are associated with differences in the ability to counter APOBEC3 proteins. In this study, we characterized variation in *vif* and correlated this with variation in the *apobec3* genes in a South African population.

Methods: The HIV-1 *vif* gene was successfully amplified from 86 HIV-1 infected drug experience individuals and sequenced on an Illumina MiSeq platform. Consensus sequences for majority and minority variants were derived for each patient using Geneious® software version 8.1. Correlation of APOBEC3 mutations with low CD4+ counts was derived using Chi-square computation.

Results: Functional Vif without stop codon was observed in all except from 4 individuals. The functional domain and motifs, such as Zn binding motifs, the proline rich domain, the human casein kinase, and the N and C-terminal core binding factor (CBF) interaction sites were highly conserved. The APOBEC3 binding motifs and the nuclear localization signal were less conserved in the South African HIV-1 Vif. APOBEC3 H variation was shown to strongly correlate with Vif variation. South African Vif appeared to be under purifying selection, with the dS= 0.2581 and dN= 0.0684 and the dN/dS value = 0.265

Conclusion: There is a high genetic diversity in the South African *vif* gene, which may reflect variation in the *apobec3* genes.

Key words

HIV-1 *vif*, *apobec3* genes, Variation, South Africa

6.1 INTRODUCTION

Several human APOBEC3 proteins can enter assembling viral particles, causing subsequent effects when the virus infects new target cells. The APOBEC3 proteins deaminate viral cDNA (cytosine to uracil) during reverse transcription causing G-to-A hypermutations and degradation, which leads to viral inhibition (Zhang *et al.* 2003). The viral infectivity factor (Vif) protein of HIV-1 has evolved to counteract this restriction.

The Vif protein is a 192 amino acid (23-kDa) protein that is required for pathogenesis in *vivo* as well as for virus replication *ex vivo* in monocytes, macrophages, primary CD4+ T lymphocytes, and non-permissive T-cell lines (Ratner *et al.* 1985; Sodroski *et al.* 1986; Fisher *et al.* 1987; Strebel *et al.* 1987; Gabuzda *et al.* 1992; von Schwedler *et al.* 1993). Studies have shown that during counter restriction, Vif forms an E3 ligase complex by binding to cellular proteins including Cullin 5, Elongin B and C to form a complex that induces polyubiquitination and directs the degradation of APOBEC3 by the 26S proteasesome (Sheehy *et al.* 2002; Conticello *et al.* 2003; Marin *et al.* 2003; Yu *et al.* 2003; Mehle *et al.* 2004).

Vif is a potential therapeutic and preventive intervention target in HIV disease pathogenesis, because it functions through protein-protein interaction with cellular proteins to form a complex that counteracts with the antiviral APOBEC3 proteins (Wang *et al.* 2014). Polymorphisms in the *vif* gene have been associated with better or worse ApoBec3 neutralization (Pace *et al.* 2006; Bizinoto *et al.* 2011; Ronsard *et al.* 2015) and contribute to the overall viral diversity.

HIV-1 subtype C is responsible for over half of HIV infections worldwide and it still remains the dominant strain in Southern Africa. Little is known about the genetic diversity in Vif among HIV strains circulating in regions where the HIV burden is high and how this affects Vif structure and function. In this study we characterize variation of the *vif* gene in a HIV infected drug experienced South African population and explore the correlation of polymorphisms in Vif and polymorphisms APOBEC3.

6.2 Hypothesis

Variations in HIV-1 *vif* influences variation in *apobec3* at a genetic level.

6.3 Objective

The main objective of the study is to characterize HIV-1 *vif* gene in drug experienced South African individuals to document the level of diversity and correlates with *Apobec3* genes and clinical parameters.

6.3.1 Specific objectives

6.3.1.1 To characterize variation in HIV-1 *vif* gene

6.3.1.2 To correlate Vif variation with APOBEC variation and clinical parameters

6.4 MATERIALS AND METHODS

6.4.1 Study population and sample collection

Characteristics of the study population and sample collection procedures are as described in chapter two, section 2.3.2.

6.4.2 Primer design and Polymerase Chain Reaction (PCR) amplification of HIV-1 *vif* gene

Primers and PCR protocols to amplify the subtype C 579 bp HIV-1 Vif Open Reading Frame (ORF) are detailed in chapter two, sections 2.3.3.2.2 to 2.3.3.2.3.

6.4.4 Library preparations for HIV-1 Vif and MiSeq sequencing

The HIV-1 Vif amplicons were prepared as detailed in chapter two, section 2.3.3.3.2.3. Libraries were prepared using methods from section 2.3.3.1.5 to section 2.3.3.1.8 and sequenced in an Illumina MiSeq V3 chemistry kit for 600 cycles (Illumina San Diego, California, USA) and generating paired-end 2x300 bp long sequence reads for each sample.

6.5 Sequence analysis

6.5.1 Demultiplexing and sequence quality control evaluation

Sequences were de-multiplexed automatically on the MiSeq as part of the data processing steps and ends pairing. Fastq files were generated for each sample representing the two paired-end reads. Sequence quality was validated using the FastQC programme.

6.5.2 Sequence filtering, trimming and mapping

The fastq files were imported into Geneious® software version 8.1 for filtering and trimming of sequences as well as downstream analysis. All sequences were trimmed at the ends as part of the assembly process using the modified-Mott algorithm and quality scores assigned by the sequencing base caller. This algorithm trims the ends and improves the error rate by more than the error probability limit threshold set, in this case

set at 0.001%. The two sequence lists from each sample were paired and mapped to HIV-1 subtype C Vif consensus (accession number: KU168308.1) using the following parameters: Gaps were allowed with a maximum of 15% per read and gap size of 15, word length was set to 14, maximum mismatches per read were set to 25%, minimum overlap identity was set to 80%, maximum ambiguity was set to 16. In general, the reference assembler uses a seed and expand-type mapper, followed by a fine-tuning step that was set to none (fast / read mapping) for this analysis.

Majority (>100-20% cut-offs) and Minority (>20-2% cut-offs) variants were called based on the variant read depth, variant frequency and consensus base frequency using the find variation/SNPs feature from Annotation and prediction menu in Geneious® software version 8.1.5. This parameter also calculates the p-value (maximum variant p-values set to 10^{-6}; 0.0001% to see variant by chance) for the specified variant frequency call, with lower p-values representing a real variant at the given position (http://www.geneious.com). Two types of consensuses were generated for each sample; a consensus with the majority (>100-20% cut-offs) variants and a consensus with minority variants (>20-2% cut-offs), because minority variants can easly be missed when general or one consensus sequence is used. All consensus contigs generated per sample were aligned using multiple alignment (global alignment with free end gaps, cost matrix: 65% similarity (5.0/-4.0) in Geneious® and translated into amino acids.

6.5.3 Phylogenetic analysis

Consensus sequences of subtypes A1, AG, B, C, D and O were retrieved from GenBank and used as reference sequences together with consensus sequences of test samples to construct a phylogenetic tree. Subtypes were also confirmed with jpHMM subtyping tools.

6.5.4 Detection of selective pressure

The SNAP (Synonymous Non-synonymous Analysis Program) version 2.1.1 in HIV database website (www.hiv.lanl.gov) was used to estimate the selection pressure on the Vif sequences by calculating the average dN/dS values. This program calculates the ratio of non-synonymous (dN) vs. synonymous (dS) substitution rates based on a set of codon-aligned nucleotide sequences to determine whether a gene is evolving under diversifying selection and identifying the sites within the gene at which diversifying

selection occurs (Nielsen and Yang 1998; Yang *et al.* 2000). Positive selection pressure acting on protein coding sequence is usually inferred when the rate of non-synonymous substitution is greater than the synonymous rate, under the assumption that synonymous substitutions are neutral and that the synonymous substitution rate therefore approximates the neutral rate of evolution. Diversifying selection can be inferred when the ratio is greater than 1.

6.5.5 HIV-1 vif amino acid analysis

Consensus amino acid frequency and Shannon entropy levels were calculated for each site as a measure of variation in the South African Vif protein sequence alignments using the Los Alamos HIV Sequence Database Entropy-Two tool (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html).

6.5.6 Correlation of APOBEC3 and Vif gene variation

APOBEC3D, F, G and H sequences from chapter five were paired with the corresponding patient's Vif sequences to correlate variation of both APOBEC and Vif using a Chi-square computation. Association of APOBEC variants and $CD4^+$ cell count were also determined using Chi-square computation.

6.6 RESULTS

Quality sequences were successfully amplified from 86 patient samples and sequence analysis revealed complete open reading frames (597bp) with no frame shift mutations, suggesting the presence of functional Vif protein in most samples except in 4 (4.5%) and 13 (15.1%) majority and minority derived consensus sequences respectively, which showed premature stop codons at the following positions +21, +38, +70, +89, +91, +174, and +192. There were no nucleotide deletion or insertion in the consensus sequences.

6.6.1 Phylogenetic analysis

The relationship or relatedness of the sequences were evaluated by phylogenetic analysis which revealed that all the consensus sequences clustered with subtype C consensus reference sequences with the following accession numbers, DQ275664, KY658706, DQ185451, KY112426, GU729791, DQ185455, KT881973, KT882202 and KT882199 from Botswana, Malawi and South African isolates. There was only one sample DF30 that clustered with subtype A1 and C consensus sequences and was confirmed by the jpHMM-subtyping tool to be an A1/C recombinant. Figure 6.1.

A

B

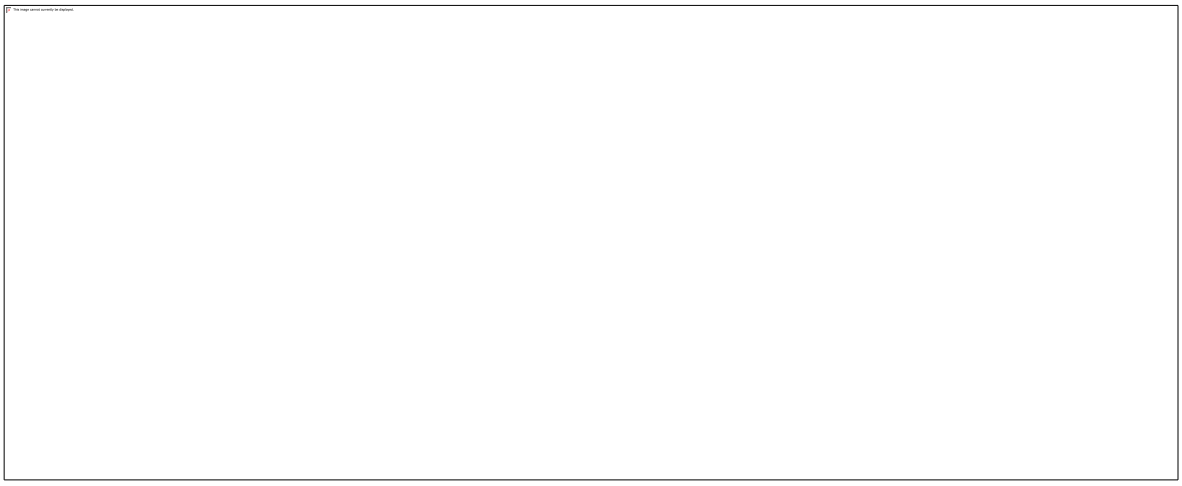


Figure 6.1: Phylogeny of 86 South African Vif consensus sequences with 10 subtype C, 2 each of subtype B, D, AG, A1 and 1 subtype O all highlighted in red were used as reference sequences (A). Sample DF30 was non-C and confirmed with jpHMM to be an A1/C recombinant (B).

6.6.2 Detection of selective pressure

The Synonymous Nonsynonymous Analysis Program (SNAP) tool was used to determine the selective pressure acting on the *vif* gene of the subtype C South African isolate. This tool calculated synonymous and nonsynonymous rates based on a set of codon-aligned nucleotide sequences to determine whether a gene is evolving under diversifying selection and identifying the sites within the gene at which diversifying selection occurs. Indels and stop codon are also indicated (figure 6.2). The type of selection pressure that might have occurred in Vif variants was calculated and the analysis suggested that the HIV-1 South African Vif was under purifying selection, with the dS= 0.2581 and dN= 0.0684 and the dN/dS value = 0.265 (figure 6.2). Purifying or negative selection pressure acting on protein- coding sequences is usually inferred when the rate of non-synonymous substitution is less than the synonymous rate (dN<dS), positive selection inferred when the rate of nonsynonymous mutation increases (dN>dS) and Neutral selection when synonymous and nonsynonymous mutation occur at the same rate (dN=dS).

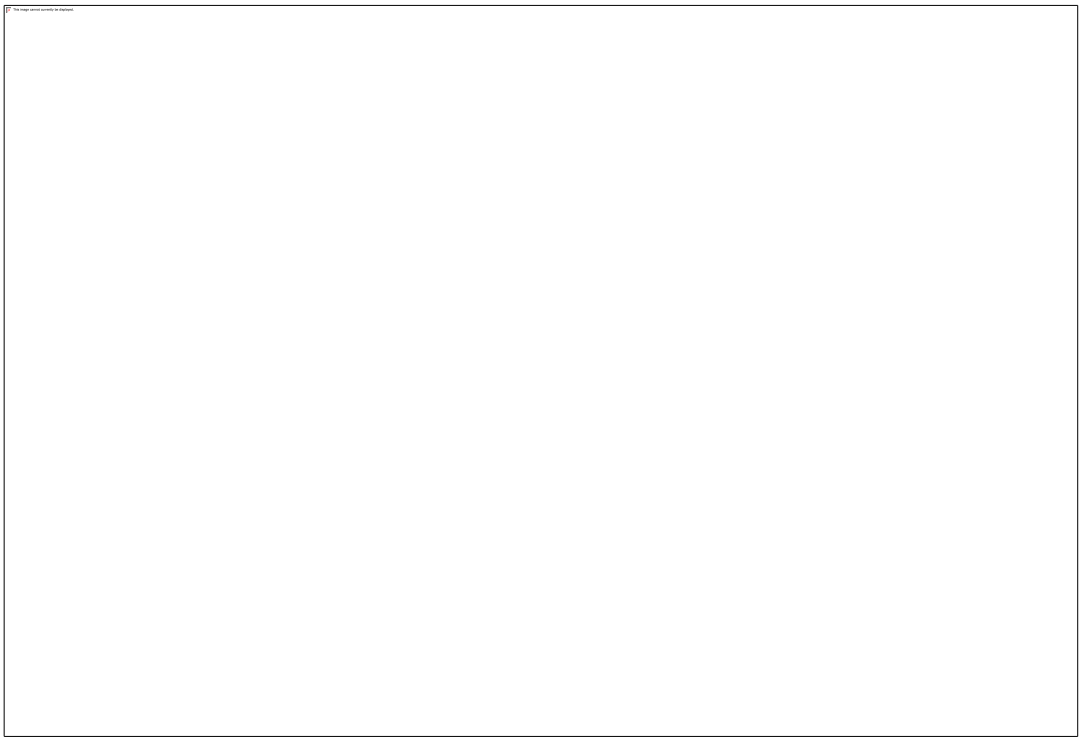


Figure 6.2: Cumulative dN/dS graph with indels and stop codon of 86 HIV-1 Vif consensus sequences. This graph indicates the rate of synonymous substitution (dS) versus non-synonymous (dN) substitution of each codon. The overall rate of dS= 0.2581 and dN= 0.0684 suggesting that the HIV-1 South African Vif was under purifying selection.

6.6.3 HIV-1 Vif amino acid analysis

The Vif nucleotides for both majority and minority-derived consensus were translated to a total of 192 amino acid and 74 (38.5%) amino acid were conserved, while 118 (61.5%) showed variable amino acid substitution indicated by high level of entropy shown in figure 6.3. Most of the sites involved in the functional binding domains such as the APOBEC3 binding sites, located at the N-terminus region, the Zinc domains, Cul5 binding regions, EloB/C Box and Vif dimerization site were conserved. There were polymorphisms or amino acid substitution within motifs implicated binding different APOBEC and the frequency of amino acid substitution at both majority and minority is displayed in figure 6.4.

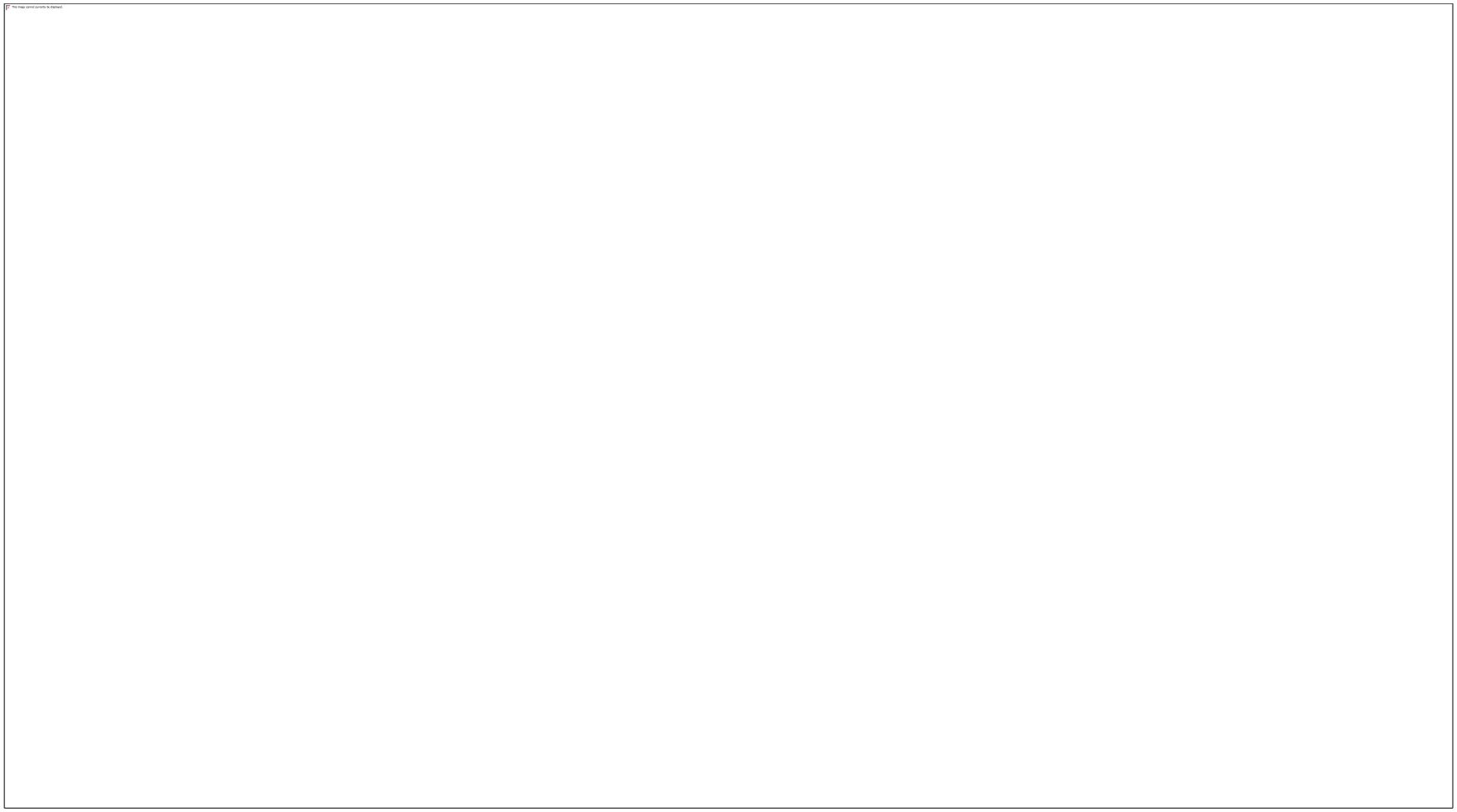


Figure 6.3: South African HIV-1 Vif protein sequence entropy of a total of 172 both minority and majority derived consensus sequence. Shannon entropy levels at each site are shown in bar graph. Boxes indicate functional binding domains exhibiting low entropy levels. Minority derived consensus show higher amino acid variability than majority-derived consensus.

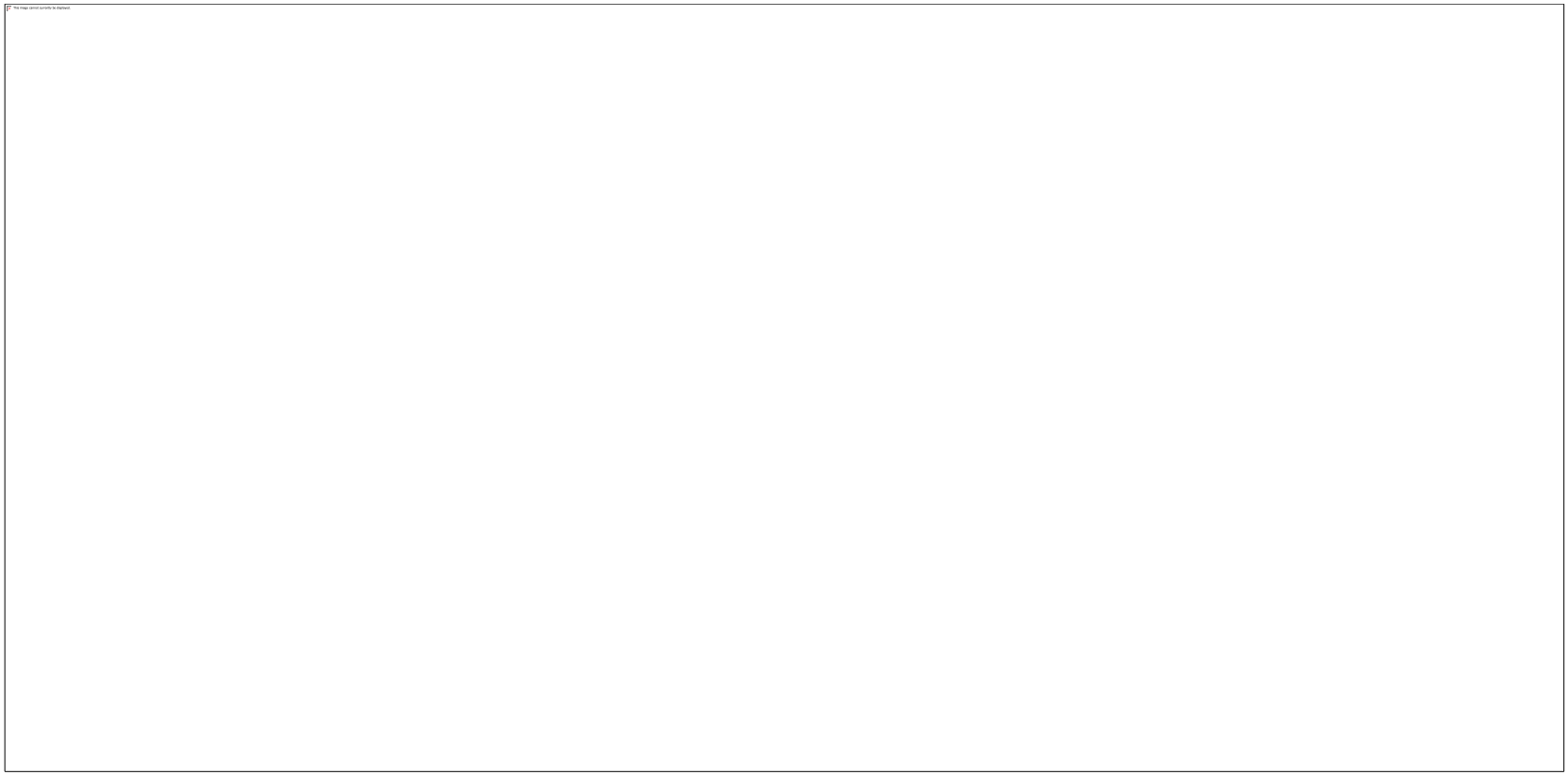


Figure 6.4: South African Vif protein sequence alignment summary from a total of 172 both majority and minority derived consensus sequences. Protein domains and functional regions are shaded to highlight functional regions and sites. These include tryptophans (W) involved in APOBEC3G binding, APOBEC3-protein family binding domains, the nuclear localization inhibitory signal, the MAPK phosphorylation sites, CBFb interaction sites, the Cullin5 zinc binding domain, the Elongin B/C (BC) box, the protease processing site, known phosphorylation sites, the Cullin5-box, and the dimerization site. The amino acid substitution frequency at each position of the *Vif* gene is indicated.

6.6.4 Correlation of APOBEC3 and variation Vif

The variation in the 86 Vif sequences was correlated with the APOBEC3 sequences in these patients. The following selected Vif variants; K22E, S32P, Y40H, E45G, F115S, G138R, L150P are shown to be responsible for inactivating intact Vif leading to failure to degrade A3G/F (Simon *et al.* 2005). These variants were not observed in the South African Vif but substitution of different amino acid were observed in those positions; N22H, S32T and Y40F.

An association of APOBEC3G variant; H186R and Q275E with low $CD4^+$ cell count ($p=0.067$, $X^2= 14.5$ and $p= 0.005$, $X^2= 2.6$) respectively was observed in this study. The G263R was associated with increased $CD4^+$ cell count. ($p= 0.0327$, $X^2= 6.9$)

It has been shown that vif variants with the ability to neutralize APOBEC3H predominate where stable A3H haplotype ii predominates. Vif residues F39 (A3H resistant) and V39 (A3H susceptible) from the study population were included in the 9713 HIV-1 isolates from different geographic regions represented in the Los Alamos database and compared to APOBEC3H haplotypes in the 1000 genome browser (Refsland *et al.* 2014). We found that F39 predominates at the higher frequency (70.5%) in our study population, which reflect what has been observed in HIV-1 vif sequences in Africa (80%). Also V39 was observed at a high frequency (17.1%) in our study population. However, this variant is common in HIV isolates in Asia (73%) (figure 6.5). The Vif variants correlate with APOBEC3H data. We have also observed that the N48 residue in Vif that confers sensitivity to stable A3H haplotype ii was present in 5.7% of our study population in comparison to 9% in other HIV-1 isolates from Africa.

A B

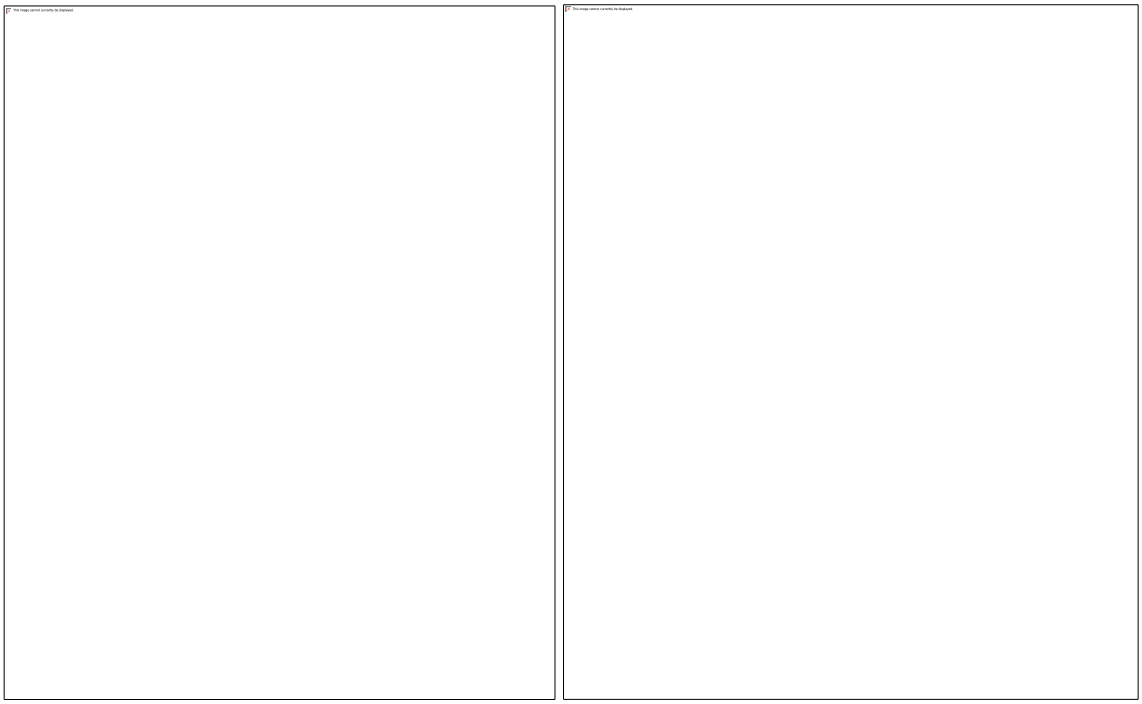


Figure 6.5: Correlation between the global distribution of HIV-1 Vif alleles and human A3H haplotypes. The left histogram depicts the frequency of HIV-1 isolates encoding a phenylalanine or valine at Vif residue 39 from the indicated geographic regions (n=9713; www.hiv.lanl.gov). The right histogram (B) shows the frequency of stable versus unstable A3H alleles from the same geographic regions (n=1092; www.1000Genomes.org).

6.7 DISCUSSION AND CONCLUSION

The HIV-1 viral infectivity factor (Vif) counteracts the activity of several APOBEC3 proteins and is considered a potential therapeutic intervention target against HIV replication. Polymorphisms in *vif* have been associated with better or worse efficacy in the ability to neutralize APOBEC3 proteins. In this study we characterized variation in HIV-1 Vif in a South African population and correlated it with variations in APOBEC3 D, F, G and H.

HIV-1 infectivity and processing are regulated by motifs such as the Proline rich self associated domain, the N and C terminal CBF β interaction site, the casein kinase II and the Zinc binding domain (Yang *et al.* 2001, 2003; Bieniasz 2003; Fujita *et al.* 2004, 2008; Mehle *et al.* 2004; Xiao *et al.* 2006; Donahue *et al.* 2008; Kataropoulou *et al.* 2009). These motifs were conserved with less than 5% and 10% variation at majority and minority-derived consensus respectively. Previous studies analyzing subtype C South African Vif sequences supports our finding where less variation was observed in these motifs (Scriba *et al.* 2002; Bell *et al.* 2007; Jacobs *et al.* 2008; Reddy *et al.* 2016), which indicates that HIV-1 is successfully expressed and its infectivity is not compromised in South African hosts.

The following motifs; ^{14}DRMR^{17}, ^{40}YRHHY^{44}, ^{74}TGERxW^{79}, ^{81}LGQGVSIEW^{89} implicated in binding A3D/F/G, and ^{39}F, ^{48}H, implicated in binding A3H, as well as the nuclear localization inhibition signal (^{90}RKKR^{93}) were highly variable in the viruses studied, which may affect the antiviral activity of the APOBEC3. The following selected Vif variants; K22E, S32P, Y40H, E45G, F115S, G138R, L150P were shown to be responsible for inactivating intact Vif leading to failure to degrade A3G/F (Simon *et al.* 2005; Ronsard *et al.* 2014, 2015). These variants were not observed in the current study. However, uncharacterized substitutions such as N22H, S32T and Y40F were observed. It will be of interest to investigate the phenotypic effects of these uncharacterized mutations in an HIV-1 subtype C based model.

Of all the members in the family that restrict HIV-1, APOBEC3H has been shown to be the most polymorphic with haplotype II showing stability and HIV-1 restriction ability whereas the unstable haplotype III and IV which have complete loss of protein function (Ooms *et al.* 2013; Refsland *et al.* 2014). Our data show high levels of A3H variability with a high frequency of unstable A3H and an observed significant correlation between

A3H haplotypes stability and predicted HIV-1 Vif. Our observation supports the finding by Refsland and colleagues (2014) where they found that stable A3H might constitute a replication barrier to a significant subset of circulating HIV-1 strain and findings in study supports the hypothesis that variations in HIV-1 *vif* influences variation in *Apobec3* at a genetic level.

Several studies have found an association between H186R and the Q275E of A3G with declined CD4+ cell count (An *et al.* 2004; Reddy *et al.* 2010; Bunupuradah *et al.* 2012; Singh *et al.* 2013). In the current study we found a similar association. Interestingly, the G363R was found to be associated with a higher CD4+ cell count. Studies have investigated the effects of A3G haplotypes on HIV-1 restriction and found that the H186R restricts HIV-1 less potently than wild type A3G both in the presence and absence of Vif. This indicates that H186R haplotype intrinsically has less antiviral activity (Reddy *et al.* 2010; Feng and Chelico 2011). Since the haplotypes of APOBEC3 D and F need to be established, future work will explore variations in Vif and associations with APOBEC3 D and F.

In conclusion, we have shown that there is a high level of HIV-1 Vif diversity in the study area. This diversity may impact the expression and packaging of Vif proteins, and the infectivity of HIV. In addition, a significant correlation was observed between HIV-1 Vif variation and APOBEC3 H haplotypes.

6.8 REFERENCE

An, P., Bleiber, G., Duggal, P., Nelson, G., May, M., Mangeat, B., Alobwede, I., Trono, D., Vlahov, D., Donfield, S., Goedert, J.J., Phair, J., Buchbinder, S., O'Brien, S.J., Telenti, A., Winkler, C.A. (2004) 'APOBEC3G Genetic Variants and Their Influence on the Progression to AIDS', *Journal of Virology*, 78(20), 11070–11076.

Bell, C.M., Connell, B.J., Capovilla, A., Venter, W.D.F., Stevens, W.S., Papathanasopoulos, M. a (2007) 'Molecular characterization of the HIV type 1 subtype C accessory genes vif, vpr, and vpu.', *AIDS research and human retroviruses*, 23(2), 322–30.

Bieniasz, P.D. (2003) 'Restriction factors: A defense against retroviral infection', *Trends in Microbiology*.

Bizinoto, M.C., Leal, É., Diaz, R.S., Janini, L.M. (2011) 'Loci Polymorphisms of the APOBEC3G Gene in HIV Type 1-Infected Brazilians', *AIDS Research and Human Retroviruses*, 27(2), 137–141.

Bunupuradah, T., Imahashi, M., Iampornsin, T., Matsuoka, K., Iwatani, Y., Puthanakit, T., Ananworanich, J., Sophonphan, J., Mahanontharit, A., Naoe, T., Vonthanak, S., Phanuphak, P., Sugiura, W. (2012) 'Association of APOBEC3G genotypes and CD4 decline in Thai and Cambodian HIV-infected children with moderate immune deficiency', *AIDS Research and Therapy*, 9.

Conticello, S.G., Harris, R.S., Neuberger, M.S. (2003) 'The Vif Protein of HIV Triggers Degradation of the Human Antiretroviral DNA Deaminase APOBEC3G', *Current Biology*, 13(22), 2009–2013.

Donahue, J.P., Vetter, M.L., Mukhtar, N.A., D'Aquila, R.T. (2008) 'The HIV-1 Vif PPLP motif is necessary for human APOBEC3G binding and degradation', *Virology*, 377(1), 49–53.

Duggal, N.K., Emerman, M. (2012) 'Evolutionary conflicts between viruses and restriction factors shape immunity', *Nature Reviews Immunology*.

Feng, Y., Chelico, L. (2011) 'Intensity of deoxycytidine deamination of HIV-1 proviral DNA by the retroviral restriction factor APOBEC3G is mediated by the noncatalytic domain', *Journal of Biological Chemistry*, 286(13), 11415–11426.

Fisher, A.G., Ensoli, B., Ivanoff, L., Chamberlain, M., Petteway, S., Ratner, L., Llo, R.C., Wong-Staal, F. (1987) 'The sor Gene of HIV-1 is Required for Efficient Virus Transmission in Vitro', *Science*, 237(4817), 888–893.

Fujita, J., Crane, A.M., Souza, M.K., Dejosez, M., Kyba, M., Flavell, R.A., Thomson, J.A., Zwaka, T.P. (2008) 'Caspase Activity Mediates the Differentiation of Embryonic Stem Cells', *Cell Stem Cell*, 2(6), 595–601.

Fujita, M., Akari, H., Sakurai, A., Yoshida, A., Chiba, T., Tanaka, K., Strebel, K., Adachi, A. (2004) 'Expression of HIV-1 accessory protein Vif is controlled uniquely to be low and optimal by proteasome degradation', *Microbes and Infection*, 6(9), 791–798.

Gabuzda, D.H., Lawrence, K., Langhoff, E., Terwilliger, E., Dorfman, T., Haseltine, W.A., Sodroski, J. (1992) 'Role of vif in replication of human immunodeficiency virus type 1 in CD4+ T lymphocytes.', *Journal of virology*, 66(11), 6489–95.

Jacobs, G.B., Nistal, M., Laten, A., van Rensburg, E.J., Rethwilm, A., Preiser, W., Bodem, J., Engelbrecht, S. (2008) 'Molecular analysis of HIV type 1 vif sequences from Cape Town, South Africa.', *AIDS research and human retroviruses*, 24(7), 991–4.

Kataropoulou, A., Bovolenta, C., Belfiore, A., Trabatti, S., Garbelli, A., Porcellini, S., Lupo, R., Maga, G. (2009) 'Mutational analysis of the HIV-1 auxiliary protein Vif identifies independent domains important for the physical and functional interaction with HIV-1 reverse transcriptase', *Nucleic Acids Research*, 37(11), 3660–3669.

Marin, M., Rose, K.M., Kozak, S.L., Kabat, D. (2003) 'HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation', *Nature Medicine*, 9(11), 1398–1403.

Mehle, A., Goncalves, J., Santa-Marta, M., McPike, M., Gabuzda, D. (2004) 'Phosphorylation of a novel SOCS-box regulates assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G degradation', *Genes and Development*, 18(23), 2861–2866.

Nielsen, R., Yang, Z. (1998) 'Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene', *Genetics*, 148(3), 929–936.

Ooms, M., Brayton, B., Letko, M., Maio, S.M., Pilcher, C.D., Hecht, F.M., Barbour, J.D.,

Simon, V. (2013) 'HIV-1 Vif adaptation to human APOBEC3H haplotypes', *Cell Host and Microbe*, 14(4), 411–421.

Pace, C., Keller, J., Nolan, D., James, I., Gaudieri, S., Moore, C., Mallal, S. (2006) 'Population level analysis of human immunodeficiency virus type 1 hypermutation and its relationship with APOBEC3G and vif genetic variation.', *Journal of virology*, 80(18), 9259–69.

Ratner, L., Haseltine, W., Patarca, R., Livak, K.J., Starcich, B., Josephs, S.F., Doran, E.R., Rafalski, J.A., Whitehorn, E.A., Baumeister, K., Ivanoff, L., Petteway, S.R., Pearson, M.L., Lautenberger, J.A., Papas, T.S., Ghrayeb, J., Chang, N.T., Gallo, R.C., Wong-Staal, F. (1985) 'Complete nucleotide sequence of the AIDS virus, HTLV-III', *Nature*, 313(6000), 277–284.

Reddy, K., Ooms, M., Letko, M., Garrett, N., Simon, V., Ndung'U, T. (2016) 'Functional characterization of Vif proteins from HIV-1 infected patients with different APOBEC3G haplotypes', *AIDS*, 30(11), 1723–1729.

Reddy, K., Winkler, C.A., Werner, L., Mlisana, K., Abdool Karim, S.S., Ndung'U, T. (2010) 'Apobec3g expression is dysregulated in primary hiv-1 infection and polymorphic variants influence cd4+ t-cell counts and plasma viral load', *AIDS*, 24(2), 195–204.

Refsland, E.W., Hultquist, J.F., Luengas, E.M., Ikeda, T., Shaban, N.M., Law, E.K., Brown, W.L., Reilly, C., Emerman, M., Harris, R.S. (2014) 'Natural Polymorphisms in Human APOBEC3H and HIV-1 Vif Combine in Primary T Lymphocytes to Affect Viral G-to-A Mutation Levels and Infectivity', *PLoS Genetics*, 10(11).

Ronsard, L., Lata, S., Singh, J., Ramachandran, V.G., Das, S., Banerjea, A.C. (2014) 'Molecular and genetic characterization of natural HIV-1 Tat exon-1 variants from North India and their functional implications', *PLoS ONE*, 9(1).

Ronsard, L., Raja, R., Panwar, V., Saini, S., Mohankumar, K., Sridharan, S., Padmapriya, R., Chaudhuri, S., Ramachandran, V.G., Banerjea, A.C. (2015) 'Genetic and functional characterization of HIV-1 Vif on APOBEC3G degradation: First report of emergence of B/C recombinants from North India', *Scientific Reports*, 5.

von Schwedler, U., Song, J., Aiken, C., Trono, D. (1993) 'Vif is crucial for human

immunodeficiency virus type 1 proviral DNA synthesis in infected cells.', *Journal of virology*, 67(8), 4945–55.

Scriba, T.J., de Villiers, T., Treurnicht, F.K., zur Megede, J., Barnett, S.W., Engelbrecht, S., van Rensburg, E.J. (2002) 'Characterization of the South African HIV type 1 subtype C complete 5' long terminal repeat, nef, and regulatory genes.', *AIDS research and human retroviruses*, 18(2), 149–59.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H. (2002) 'Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein', *Nature*, 418(6898), 646–650.

Simon, V., Zennou, V., Murray, D., Huang, Y., Ho, D.D., Bieniasz, P.D. (2005) 'Natural variation in vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification', *PLoS Pathogens*, 1(1), 0020–0028.

Singh, K.K., Wang, Y., Gray, K.P., Farhad, M., Brummel, S., Fenton, T., Trout, R., Spector, S.A. (2013) 'Genetic variants in the host restriction factor APOBEC3G are associated with HIV-1-related disease progression and central nervous system impairment in children', *Journal of Acquired Immune Deficiency Syndromes*, 62(2), 197–203.

Sodroski, J., Goh, W.C., Rosen, C., Tartar, A., Portetelle, D., Burny, A., Haseltine, W. (1986) 'Replicative and cytopathic potential of HTLV-III/LAV with sor gene deletions', *Science*, 231(4745), 1549–1553.

Strebel, K., Daugherty, D., Clouse, K., Cohen, D., Folks, T., Martin, M.A. (1987) 'The HIV A (sor) gene product is essential for virus infectivity', *Nature*, 328(6132), 728–730.

Wang, H., Lv, G., Zhou, X., Li, Z., Liu, X., Yu, X.F., Zhang, W. (2014) 'Requirement of HIV-1 Vif C-terminus for Vif-CBF-β interaction and assembly of CUL5-containing E3 ligase', *BMC Microbiology*, 14(1).

Xiao, Z., Ehrlich, E., Yu, Y., Luo, K., Wang, T., Tian, C., Yu, X.F. (2006) 'Assembly of HIV-1 Vif-Cul5 E3 ubiquitin ligase through a novel zinc-binding domain-stabilized hydrophobic interface in Vif', *Virology*, 349(2), 290–299.

Yang, B., Gao, L., Li, L., Lu, Z., Fan, X., Patel, C.A., Pomerantz, R.J., DuBois, G.C.,

Zhang, H. (2003) 'Potent suppression of viral infectivity by the peptides that inhibit multimerization of human immunodeficiency virus type 1 (HIV-1) Vif proteins', *Journal of Biological Chemistry*, 278(8), 6596–6602.

Yang, S., Sun, Y., Zhang, H. (2001) 'The multimerization of human immunodeficiency virus type I Vif protein: A requirement for Vif function in the viral life cycle', *Journal of Biological Chemistry*, 276(7), 4889–4893.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, a M. (2000) 'Codon-substitution models for heterogeneous selection pressure at amino acid sites.', *Genetics*, 155(1), 431–449.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, W., Mao, P., Yu, X.F. (2003) 'Induction of APOBEC3G Ubiquitination and Degradation by an HIV-1 Vif-Cul5-SCF Complex', *Science*, 302(5647), 1056–1060.

Zhang, H., Yang, B., Pomerantz, R.J., Zhang, C., Arunachalam, S.C., Gao, L. (2003) 'The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA', *Nature*, 424(6944), 94–98.

CHAPTER SEVEN

General Conclusions, Limitations and Recommendations

7.1 General conclusions

Protocols developed in this study can be applied to amplify and sequence any host and HIV-1 genes of interest allowing cheaper in depth sequencing.

Our findings show that a highly significant number of chronically HIV-1 subtype C infected patients under Maraviroc-free treatment harbor CXCR4 utilizing viruses. The data is useful in the consideration of whether to include entry antagonists such as Maraviroc in alternative forms of treatment for patients failing second line treatment regimen in the study setting. The determination of co-receptor usage prior to initiation of therapy that includes of Maraviroc is suggested.

Variation in the CCR5 coding region were observed at higher frequencies compare to other studies conducted in South African populations at different geographic locations. This data may suggest that different populations in South Africa have different SNP frequencies. None of the polymorphisms identified in the study were located at the Maraviroc binding motif, therefore the subset of patients infected by R5 viruses may benefit from this drug.

We have shown that significant APOBEC3 variation exists among an ethnically diverse population of South Africa by providing extensive data for 4 different A3 genes that are known to restrict HIV infection, but have only been sparsely studied in African populations. This study provides a baseline for future studies that would functionally characterize SNPs identified in this population, in order to understand the role of novel and/or low frequency variants observed. *Ex vivo* and *in vivo* studies will increase our understanding of how these variants might have cumulatively impacted the epidemic in Northern South Africa.

This study also show that there is a high level of HIV-1 Vif diversity in the study area. This diversity may impact the expression and packaging of Vif proteins, and the infectivity of HIV. In addition, a significant correlation was observed between HIV-1 Vif variation and APOBEC3H haplotypes.

7.2 Study limitations

The limitations of this study include;

1. There was no control group, all the subjects were HIV infected and were mostly at the chronic stage of infection. It is unclear if HIV-1 negative subjects would present a different APOBEC3 and CCR5 profile. If this is the case, it could imply that these APOBEC3 and CCR5 genotypes either alone or in combination might influence HIV transmission. Although we did not quantify

the APOBEC3 and CCR5 mRNA levels in this study, reports comparing HIV-1 non-controllers versus long-term non-progressors (LTNP) have shown that LTNPs express a higher level of A3G and A3F proteins (Jin *et al*. 2005).

2. Another limitation of this study was assembling quasispecies and determining haplotypes from shorter NGS reads, this mostly for the HIV-1 Vif reads.

3. Single Nucleotide Polymorphisms were not analyzed based on ethnicity because of low numbers of individuals from some of the ethnic groups.

7.3 Recommendations

The following recommendations can be inferred from the study:

1. The determination of co-receptor usage prior to initiation of therapy consisting of Maraviroc is recommended for a better management of treatment.

2. Low CD+ counts (below 350 cells/ul) could be used as a proxy for the presence of CXCR4 using viruses.

3. Next Generation Sequencing (NGS) technologies due to their sensitivity are recommended in co-receptor determination to detect minority quasispecies that could easly be missed when Sanger based sequencing is used.

4. Studies comparing different genotyping tools such as NGS and Sanger sequencing are recommended to investigate the sensitivity of SNPs detection.

5. This study provides a baseline for future studies that would functionally characterize SNPs identified in this population, in order to understand the role of novel and/or low frequency variants observed. *Ex vivo* and *in vivo* studies will increase our understanding of how these variants might have cumulatively impacted the epidemic in Northern South Africa.

APPENDICES

Appendix I

Appendix II

Characterization of APOBEC3 variations in a South African population

Nontokozo D. Matume[1,3], Denis M. Tebit[1,3], Laurie R. Gray[1], Stephen D. Turner[2], David Rekosh[1,3], Pascal O. Bessong[3*], Marie-Louise Hammarskjöld[1,3], *

[1] Myles H. Thaler Center for AIDS and Human Retrovirus Research, Department of Microbiology, Immunology and Cancer Biology, University of Virginia, Charlottesville, USA

[2] Department of Public Health Sciences and Bioinformatics Core, University of Virginia School of Medicine, Charlottesville Virginia 22908

[3] HIV/AIDS & Global Health Research Programme, Department of Microbiology, University of Venda, Thohoyandou, South Africa*Corresponding authors

Marie-Louise Hammarskjold

University of Virginia, Charlottesville, VA, USA

Email: mh7g@virginia.edu

Tel: +1 434-982-1598

Pascal O. Bessong

University of Venda, Thohoyandou, South Africa

Email: bessong@univen.ac.za

Tel: +27 15 962 8301

ABSTRACT

The apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3 (APOBEC3) genes A3D, A3F, A3G and A3H have all been implicated in the restriction of human immunodeficiency virus type 1 (HIV-1) replication. Polymorphisms in these genes may impact viral replication and fitness and contribute to viral diversity. To characterize polymorphisms in the coding regions of these APOBEC3 genes in a population from the Limpopo Province of South Africa, APOBEC3 gene fragments were amplified from genomic DNA of 192 HIV-1 infected subjects and sequenced on an Illumina MiSeq platform. SNPs were confirmed and compared to SNPs in other populations reported in the 1000 Genome Phase III and HapMap databases. Hardy-Weinberg Equilibrium was calculated and haplotypes and Linkage Disequilibrium (LD) were inferred using the LDlink 3.0 web tool.

Known variants compared to the GRCh37 consensus genome sequence were detected at relatively high frequencies (>5%) in all of the Apobec genes. A3H showed the most variation, with several of the variants present in both alleles in all of the patients. Several minor allele variants (<5%) were also detected in A3D, A3F and A3G. In addition, novel R6K, L221R and T238I variants in A3D and I117I in A3F were observed. Most of the SNPs were in strong LD with minor deviation from the Hardy-Weinberg Equilibrium. Four, five, four, and three haplotypes were identified for A3D, A3F, A3G, and A3H respectively. In general, polymorphisms in APOBEC3D, 3F, 3G and 3H were higher in our South African cohort than previously reported among other African, European and Asian populations.

Key words: APOBEC3; Single nucleotide polymorphism; South Africa

INTRODUCTION

The genes for the apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like protein gene family (APOBEC3), a family of seven members (APOBEC3 A, B, C, D, F, G and H), are situated on human chromosome 22. The proteins encoded by these genes are cytidine deaminases that have been classified as restriction factors because of their role as innate immunity factors. They provide host cell defense against a diverse set of retroviruses, endogenous retroelements and DNA viruses, including human immunodeficiency virus (HIV) [1]–[3]. APOBEC proteins restrict HIV through deamination of cytosines in viral cDNA during reverse transcription, causing G-to-A hypermutations in the viral DNA product, which results in degradation and viral inhibition [4]. The Vif protein of HIV has evolved to counteract this restriction by binding to APOBEC proteins leading to proteasomal degradation.

One of the most studied APOBEC proteins and the first that was discovered to restrict HIV-1 replication is APOBEC3G. In the absence of the HIV-1 Vif protein, APOBEC3G is efficiently packaged into viral particles, causing restriction during reverse transcription. The gene was originally identified as an HIV restriction factor because its expression converted a T-cell line that could support the replication of an HIV lacking *vif* into one that had a non-permissive phenotype [1].

Three other members of the APOBEC family, APOBEC3D (A3D), APOBEC3F (A3F), and APOBEC3H (A3H) can also be packaged into HIV particles and inhibit viral replication, when stably expressed in human T-cell lines [5]. Endogenous A3D and A3F combine to generate the 5'-GA-to-AA mutation pattern observed in *vif*-negative HIV grown in the non-permissive T-cell line CEM2n [6], [7]. Of the seven different human haplotypes of APOBEC3H, only hapII, hapV and hapVII are stable at the protein level and capable of HIV restriction [8]–[12].

Several APOBEC3 (A3D, A3F, A3G and A3H) genes are known to possess common polymorphisms that render them defective with reduced antiviral activity and increased

sensitivity to HIV-1 Vif [5], [13]–[16]. The genetic associations between natural polymorphisms in APOBEC genes and the ability of the resulting proteins to restrict HIV and the contribution of polymorphisms to overall HIV diversity and disease progression have not received widespread attention. Polymorphisms in APOBEC genes could play a significant role in HIV-1 evolution and diversity, especially in African populations, where the prevalence of HIV-1 is still increasing.

African populations are characterized by a high level of genetic diversity owing to a large number of variable genes and alleles [17]–[20]. Patterns of genetic variation in the African population are influenced by a demographic history that includes changes in population size, admixture and locus-specific forces such as natural selection, recombination and mutation. Genetic studies of structural variation of genes across ethnically diverse populations have been conducted [21]. Many population genetic studies of African populations are based on analysis of genetic markers genotyped in a small number of populations in projects such as the 1000 Genomes Project (2010) and the International Haplotype Map (HapMap) Project [22]–[24]. Although these projects are valuable in their description of the overall human genetic diversity, they are limited in their coverage of African populations [25]. Thus, it is important to continue to add information about African populations that are underrepresented in human genomic studies, such as the South African population.

South Africa embodies a rich collection of ethnic backgrounds in addition to the more recent Caucasian immigrants. The major ethnic groups include the Bapedi, Basotho, Ndebele, Swati, Tsonga, Tswana, Xhosa, Venda and Zulu. The genetic substructure of these populations has been assessed by studying the Y-chromosome and autosomal DNA resulting into a cluster of three specific groups: Tswana/Sotho, Nguni and Venda [26], [27]. It is of clear interest to characterize the APOBEC3 gene polymorphisms existing in these various populations, since they may play a crucial role in the restriction and evolution of HIV-1.

In the current study, we characterized the genetic variability within the coding regions of A3D, A3F, A3G and A3H to document the level of diversity in samples obtained from HIV-1 positive individuals attending three HIV clinics in the Limpopo Province of Northern South Africa.

MATERIALS AND METHODS

Ethical considerations

The study protocol was approved by the Research Ethics Committee of the University of Venda (SMNS/13/MBY/01/0625) and the University of Virginia Institutional Review Board (IRB-HSR #16815). Signed informed consent was obtained from study subjects prior to blood draw and collection of demographic and clinical data. Personal identifiers were stripped prior to sample processing and data analysis.

Study population and DNA extraction

The study population comprised of HIV positive individuals who presented for routine care at the Bela-Bela Wellness clinic, Donald Fraser Hospital and Hospitals in Polokwane in the Limpopo province of Northern South Africa. DNA was extracted from peripheral blood mononuclear cells (PBMC) of 192 HIV infected individuals, using QIAamp DNA blood mini kit (Qiagen) according to the manufacturer's instructions.

Primer design

Primers to amplify the four APOBEC3 genes (A3D, A3F, A3G, and A3H) were designed using Geneious® 8.1.5 software (Biomatters, Inc.). A nested PCR strategy was used to amplify each APOBEC gene. The outer primer set was designed to flank and amplify a long gene fragment in the 1st polymerase chain reaction (PCR), while two sets of primers were designed to amplify two fragments of each gene in a nested PCR using the 1st round PCR product as the template (Table 1). The primer sets mapped in Ensembl genome browser to transcripts ENSG00000243811, ENSG00000128394, ENSG00000239713 and ENSG00000100298 for A3D, A3F, A3G and A3H respectively.

Table 1: List of APOBEC3 primers designed; primer name, sequence and product size are indicated.

Name	Sequence (5'-3')	Product size
A3D (12.1 kb)		
A3D Forward primer	AGGAAGCCTCGCTCTCTCA	12,069 bp
A3D Forward primer	CAGGCAGGGTCTTGATCTGT	
A3D Amplicon 1F	AAAAAGAGGGAGACTGGGACAAGCGTATCTAAGA	4,300 bp
A3D Amplicon 1R	GAGTGTGGGTGAGGGGGTGTAACCATGAG	
A3D Amplicon 2F	AGCTAGGAGAGGTCACCCTG	3,188 bp
A3D Amplicon 2F	CAGGAGGCTAGAAGAGACAGACCATGAGGC	
A3F (13.31 kb)		
A3F 1st round f	ACCAGAAAGAGGGTGAGAGACTGAGGAAGATAAAG	13,142 bp
A3F 1st round rv	AGCCATTTATTGCAGAAGCTATGGATAAAGCTGGT	
A3F Amplicon 1 f	ACCAGAAAGAGGGTGAGAGACTGAGGAAGATAAAG	4,918 bp
A3F Amplicon 1 rv	GGGTGAGGGGTGTAACCATG	
A3F Amplicon 2 f	TTCAGAAACCCGATGGAGGC	4,478 bp
A3F Amplicon 2 rv	AGCCATTTATTGCAGAAGCTATGGATAAAGCTGGT	
A3G (10.74 kb)		
A3G 1st round f	TGTTAACCAGAGGCTGCTCTTCCCAGG	11,852 bp
A3G 1st round rv	TCCCTGGGACTCAGCTCC	
A3G Amplicon 1 f	ATTTGTCCCCAGCTCTGTGG	3,231 bp
A3G Amplicon 1 rv	AGAGGACCTGGTCTGGAACA	
A3G Amplicon 2 f	CAAGGGAGGAAGCGTGGAG	3,908 bp
A3G Amplicon 2 rv	TGCATTGCTTTGCTGGTGTC	
A3H (6.8 kb)		
APOBEC3 H forward primer full length	TCTGTTGCACAGAAACACGATGG	3522bp
APOBEC3 H reverse	CAACTGACATGCCCCAGGG	

primer full length		
APOBEC3 H forward primer Exon2 (A3HfE2)	TCTGTTGCACAGAAACACGATGG	452bp
APOBEC3 H Reverse primer Exon 2(A3HrE2)	TTCCCGAAGTAGTGACTGAGC	
APOBEC3 H forward primer Exon 3 &4(A3HfE3/4	GCCACGCACTAGAAAGTTCAC	934bp
APOBEC3 H Reverse primer Exon 3&4(A3HrE3/4)	ACAGTGCCTCACCTTTATCC	

Polymerase Chain Reaction (PCR) to amplify A3D, A3F, A3G and A3H genes

The Takara (LA) PCR Kit Ver. 2.1 for long DNA fragments amplification (Clontech) was used to amplify the complete 12.16 kb A3D, 13.31 kb A3F, 10.74 kb of A3G, and 6.8 kb A3H genes in a 1st round PCR reaction using genomic patient DNA. The 1st round primary PCR products were then used as templates in “nested” PCR reactions to generate shorter PCR products/ All the PCR reactions contained: 1X PCR Mg^{2+} plus buffer, 400µM dNTPs, 0.2µM of each primer (Table 1) and 1.25 units of LA Taq high fidelity polymerase in a total volume of 20 µl. The following cycling conditions were used for all PCR reactions: Initial denaturation at 94°C for 1 min, 30 cycles of denaturation at 98°C for 10sec, annealing at temperatures varying from 53°C to 68°C for 15 min (depending on primers) and extension at 72°C for 10 min. Final amplicons were purified using AMpure XP beads (Beckman Coulter) and quantified using a Qubit 3.0 Fluorometer with the dsDNA HS kit (Invitrogen). Equimolar concentrations of the two shorter amplicons generated for each gene were pooled and normalized to 1ng using 10mM Tris elution buffer.

Fragmentation, tagmentation and addition of Illumina indices

The Tn5 transposase enzyme, which employs similar principles as the Illumina Nextera Kit, was used to fragment about 1-10ng of DNA amplicons to sizes ranging from 35bp to 700bp, tagged with sequencing adaptors. The reaction mixture contained: 4µl tagmentation buffer (5X TAPS-DMF), 1-5µl Tn5 transposase (1X-5X) and 1-10ng DNA, with an addition of nuclease free water to add up to a final volume of 20µl. The reaction was performed at 55°C for 5 minutes. The Tn5 transposase enzyme was produced and characterized in our laboratory, using published protocols [28]. Following this step, unique Illumina dual-index barcodes (index 1 (i7) and index 2 (i5)) were added to each sample in a short PCR of 12 cycles, followed by a second AMpure XP bead purification, generating 300-500bp indexed fragments for sequencing. Using the full complement of Nextera XT indices, up to 96 individual samples were pooled for each run.

Library normalization, pooling and sequencing

After purification, libraries were size verified using a bioanalyzer 2100 with a High Sensitive DNA assay kit (Agilent Genomics), quantified and normalized to a concentration of 4nM each. The normalized libraries were then pooled, and denatured into single strands. For good cluster generation, 1.8pM of the pooled library spiked with 25-30% PhiX was then loaded into the sequencing cartridge. Biological sample sheets were created in Basespace by labeling each sample with the appropriate index and setting up a sequencing run for the MiSeq. Each run generated approximately 25 million reads/sequences per sample. All of the individual patient sequences used in this study have been submitted to the NCBI Sequence Read Archive (Project number: PRJNA429751) and can be accessed using the following link; http://www.ncbi.nlm.nih.gov/bioproject/429751 (see Table 2)

Table 2: Supplementary data, indication patient ID, clinic where the samples were collected and genes sequenced per sample.

Patient ID	Origin (clinic)	Genes sequenced
BB01	Bela-Bela	A3G, A3H
BB02	Bela-Bela	A3D, A3F, A3G, A3H
BB03	Bela-Bela	A3D, A3F, A3G, A3H
BB04	Bela-Bela	A3F, A3G, A3H
BB05	Bela-Bela	A3G, A3H
BB07	Bela-Bela	A3F, A3G, A3H
BB08	Bela-Bela	A3D, A3F, A3H
BB09	Bela-Bela	A3D, A3F, A3H
BB10	Bela-Bela	A3D, A3F, A3G, A3H
BB11	Bela-Bela	A3D, A3F, A3G, A3H
BB12	Bela-Bela	A3D, A3F, A3G, A3H
BB13	Bela-Bela	A3, A3F, A3G, A3H
BB14	Bela-Bela	A3H
BB15	Bela-Bela	A3F, A3G, A3H
BB16	Bela-Bela	A3F, A3G, A3H
BB17	Bela-Bela	A3F, A3G, A3H
BB18	Bela-Bela	A3H
BB19	Bela-Bela	A3F, A3G, A3H
BB20	Bela-Bela	A3D, A3F, A3G, A3H
BB21	Bela-Bela	A3D, A3F, A3G, A3H
BB22	Bela-Bela	A3F, A3G, A3H
BB23	Bela-Bela	A3F, A3G, A3H
BB24	Bela-Bela	A3D, A3F, A3G, A3H
BB26	Bela-Bela	A3D, A3F, A3G, A3H
BB28	Bela-Bela	A3F, A3G, A3H

BB29	Bela-Bela	A3D, A3F, A3H
BB30	Bela-Bela	A3D, A3G, A3H
BB31	Bela-Bela	A3D, A3F, A3G, A3H
BB32	Bela-Bela	A3D, A3F, A3G, A3H
BB33	Bela-Bela	A3D, A3F, A3H
BBS04	Bela-Bela	A3D, A3F, A3G, A3H
BBS05	Bela-Bela	A3D, A3F, A3G, A3H
DF02	Donald Frazer	A3D, A3F, A3G, A3H
DF04	Donald Frazer	A3D, A3F, A3G, A3H
DF05	Donald Frazer	A3D, A3F, A3G A3H
DF06	Donald Frazer	A3D, A3F, A3G, A3H
DF07	Donald Frazer	A3D, A3F, A3G, A3H
DF08	Donald Frazer	A3D, A3F, A3G, A3H
DF09	Donald Frazer	A3D, A3F, A3G, A3H
DF10	Donald Frazer	A3D, A3F, A3G, A3H
DF100	Donald Frazer	A3D, A3F, A3G, A3H
DF101	Donald Frazer	A3D, A3F, A3G, A3H
DF102	Donald Frazer	A3D, A3F, A3G, A3H
DF103	Donald Frazer	A3D, A3F, A3G, A3H
DF104	Donald Frazer	A3D, A3F, A3G, A3H
DF105	Donald Frazer	A3D, A3F, A3G, A3H
DF106	Donald Frazer	A3D, A3F, A3G, A3H
DF107	Donald Frazer	A3D, A3F, A3G, A3H
DF108	Donald Frazer	A3D, A3F, A3G, A3H
DF109	Donald Frazer	A3D, A3F, A3G, A3H
DF11	Donald Frazer	A3D, A3F, A3G, A3H
DF110	Donald Frazer	A3D, A3F, A3G, A3H
DF111	Donald Frazer	A3D, A3F, A3G, A3H
DF12	Donald Frazer	A3D, A3F, A3G, A3H

DF13	Donald Frazer	A3D, A3F, A3G, A3H
DF14	Donald Frazer	A3D, A3F, A3G, A3H
DF15	Donald Frazer	A3D, A3F, A3G, A3H
DF16	Donald Frazer	A3D, A3F, A3G, A3H
DF17	Donald Frazer	A3D, A3F, A3G, A3H
DF18	Donald Frazer	A3D, A3F, A3G, A3H
DF19	Donald Frazer	A3D, A3F, A3G, A3H
DF20	Donald Frazer	A3D, A3F, A3G, A3H
DF21	Donald Frazer	A3D, A3F, A3G, A3H
DF22	Donald Frazer	A3D, A3F, A3G, A3H
DF23	Donald Frazer	A3D, A3F, A3G, A3H
DF24	Donald Frazer	A3D, A3F, A3G, A3H
DF25	Donald Frazer	A3D, A3F, A3G, A3H
DF26	Donald Frazer	A3D, A3F, A3G, A3H
DF27	Donald Frazer	A3D, A3F, A3G, A3H
DF28	Donald Frazer	A3D, A3F, A3G, A3H
DF29	Donald Frazer	A3D, A3F, A3G, A3H
DF30	Donald Frazer	A3D, A3F, A3G, A3H
DF31	Donald Frazer	A3D, A3F, A3G, A3H
DF33	Donald Frazer	A3D, A3F, A3G, A3H
DF34	Donald Frazer	A3D, A3F, A3G, A3H
DF36	Donald Frazer	A3D, A3F, A3G, A3H
DF37	Donald Frazer	A3D, A3F, A3G, A3H
DF38	Donald Frazer	A3D, A3F, A3G, A3H
DF39	Donald Frazer	A3D, A3F, A3G, A3H
DF40	Donald Frazer	A3D, A3F, A3G, A3H
DF41	Donald Frazer	A3D, A3F, A3G, A3H
DF42	Donald Frazer	A3D, A3F, A3G, A3H
DF43	Donald Frazer	A3D, A3F, A3G, A3H

DF44	Donald Frazer	A3D, A3F, A3G
DF45	Donald Frazer	A3D, A3F, A3G, A3H
DF46	Donald Frazer	A3D, A3F, A3G, A3H
DF47	Donald Frazer	A3D, A3F, A3G, A3H
DF48	Donald Frazer	A3D, A3F, A3G
DF49	Donald Frazer	A3D, A3F, A3G, A3H
DF50	Donald Frazer	A3D, A3F, A3G, A3H
DF51	Donald Frazer	A3D, A3F, A3G, A3H
DF52	Donald Frazer	A3D, A3F, A3G, A3H
DF53	Donald Frazer	A3D, A3F, A3G, A3H
DF54	Donald Frazer	A3D, A3F, A3G, A3H
DF55	Donald Frazer	A3D, A3F, A3G, A3H
DF56	Donald Frazer	A3D, A3F, A3G, A3H
DF57	Donald Frazer	A3D, A3F, A3G, A3H
DF58	Donald Frazer	A3D, A3F, A3G, A3H
DF59	Donald Frazer	A3D, A3F, A3G, A3H
DF60	Donald Frazer	A3D, A3F, A3G, A3H
DF61	Donald Frazer	A3D, A3F, A3G, A3H
DF62	Donald Frazer	A3D, A3F, A3G, A3H
DF63	Donald Frazer	A3F, A3G, A3H
DF64	Donald Frazer	A3D, A3F, A3G, A3H
DF65	Donald Frazer	A3D, A3F, A3G, A3H
DF66	Donald Frazer	A3D, A3F, A3G, A3H
DF67	Donald Frazer	A3D, A3F, A3G, A3H
DF68	Donald Frazer	A3D, A3F, A3G, A3H
DF69	Donald Frazer	A3D, A3F, A3G, A3H
DF70	Donald Frazer	A3F, A3G, A3H
DF71	Donald Frazer	A3D, A3F, A3G, A3H
DF72	Donald Frazer	A3D, A3F, A3G, A3H

DF73	Donald Frazer	A3D, A3F, A3G, A3H
DF74	Donald Frazer	A3D, A3F, A3G, A3H
DF75	Donald Frazer	A3D, A3F, A3G, A3H
DF76	Donald Frazer	A3D, A3F, A3G, A3H
DF77	Donald Frazer	A3D, A3F, A3G, A3H
DF78	Donald Frazer	A3D, A3F, A3H
DF79	Donald Frazer	A3D, A3F, A3G, A3H
DF80	Donald Frazer	A3D, A3F, A3G, A3H
DF81	Donald Frazer	A3D, A3F, A3G, A3H
DF82	Donald Frazer	A3D, A3F, A3G, A3H
DF83	Donald Frazer	A3D, A3F, A3G, A3H
DF84	Donald Frazer	A3D, A3F, A3G, A3H
DF85	Donald Frazer	A3D, A3F, A3G, A3H
DF86	Donald Frazer	A3D, A3G, A3H
DF87	Donald Frazer	A3D, A3F, A3G, A3H
DF88	Donald Frazer	A3D, A3F, A3G, A3H
DF89	Donald Frazer	A3D, A3F, A3G, A3H
DF90	Donald Frazer	A3D, A3F, A3G, A3H
DF91	Donald Frazer	A3D, A3F, A3G, A3H
DF92	Donald Frazer	A3D, A3F, A3G, A3H
DF94	Donald Frazer	A3D, A3F, A3G, A3H
DF95	Donald Frazer	A3D, A3F, A3G, A3H
DF96	Donald Frazer	A3D, A3F, A3G, A3H
DF97	Donald Frazer	A3D, A3F, A3G, A3H
DF98	Donald Frazer	A3D, A3F, A3G, A3H
LP1111	Polokwane	A3D, A3F, A3G, A3H
LP1112	Polokwane	A3D, A3F, A3G, A3H
LP1119	Polokwane	A3D, A3F, A3G, A3H
LP1120	Polokwane	A3D, A3F, A3G, A3H

LP1121	Polokwane	A3D, A3F, A3G, A3H
LP1122	Polokwane	A3D, A3F, A3G, A3H
LP1123	Polokwane	A3D, A3F, A3G, A3H
LP1125	Polokwane	A3D, A3F, A3G, A3H
LP1126	Polokwane	A3D, A3F, A3G, A3H
LP1129	Polokwane	A3D, A3F, A3G, A3H
LP1131	Polokwane	A3D, A3F, A3G, A3H
LP1135	Polokwane	A3D, A3F, A3G, A3H
LP1136	Polokwane	A3D, A3F, A3G, A3H
LP1137	Polokwane	A3D, A3F, A3G, A3H
LP1138	Polokwane	A3D, A3F, A3G, A3H
LP1142	Polokwane	A3D, A3F, A3G, A3H
LP1148	Polokwane	A3D, A3F, A3G, A3H
LP1150	Polokwane	A3D, A3F, A3G, A3H
LP1151	Polokwane	A3D, A3F, A3G, A3H
LP1152	Polokwane	A3D, A3F, A3G, A3H
LP1153	Polokwane	A3D, A3F, A3G, A3H
LP1154	Polokwane	A3D, A3F, A3G, A3H
LP1157	Polokwane	A3D, A3F, A3G, A3H
LP1159	Polokwane	A3D, A3F, A3G, A3H
LP1160	Polokwane	A3D, A3F, A3G, A3H
LP1162	Polokwane	A3D, A3F, A3G, A3H
LP1164	Polokwane	A3D, A3F, A3G, A3H
LP1166	Polokwane	A3D, A3F, A3G, A3H
LP1170	Polokwane	A3D, A3F, A3G, A3H
LP1171	Polokwane	A3D, A3F, A3G, A3H
LP1186	Polokwane	A3D, A3F, A3G, A3H
LP1187	Polokwane	A3D, A3F, A3G, A3H
LP1207	Polokwane	A3D, A3G, A3H

LP1225	Polokwane	A3D, A3F, A3G, A3H
LP1226	Polokwane	A3D, A3F, A3G, A3H
LP1229	Polokwane	A3D, A3F, A3G, A3H
LP1265	Polokwane	A3D, A3F, A3G, A3H
LP1390	Polokwane	A3D, A3F, A3G, A3H
LP1436	Polokwane	A3D, A3F, A3G, A3H
LP1452	Polokwane	A3D, A3F, A3G, A3H
LP1501	Polokwane	A3D, A3F, A3G, A3H

Demultiplexing and sequence quality control evaluation

Sequences were demultiplexed automatically on the MiSeq as part of the data processing steps and ends pairing. FASTQ files were generated for each sample representing the two paired-end reads. Sequence quality was validated using the Galaxy NGS platform: QC and manipulation and fastQC program.

Sequence filtering, trimming, mapping and variant calling

Sequencing data quality, including the duplication rate, percent GC, and read quality was assessed by quality control tools for high throughput sequencing data [29], [30]. After filtering low coverage samples, reads were aligned against the human genome with BWA-MEM [31]. Alignments were sorted, marked for duplicates, and indexed using SAMtools [32]. Variants were called using Freebayes, a Haplotype-based tool to detect variants using short-read sequencing data [33]. Variant calls were normalized and decomposed with vt, a unified representation of genetic variants, and functionally annotated using SnpEff, a program for annotating and predicting the effects of single nucleotide polymorphisms [34]. Comprehensive annotation and prioritization was performed using the GEMINI framework for Integrative Exploration of Genetic Variation and Genome Annotations [35]. All further data manipulation and analysis was performed using R, a Language and Environment for Statistical Computing [36].

Statistical analysis

Hardy-Weinberg Equilibrium (HWE)

All variant loci detected within the coding regions of these genes were tested for deviation from the Hardy-Weinberg Equilibrium (HWE) using an excel HWE calculator and chi-square test with $P<0.05$ showing non-consistence with HWE [37].

Pairwise linkage disequilibrium (LD) and Haplotype assignment

Pairwise linkage disequilibrium (LD) analysis between the SNPs in each gene was performed to test if they were in LD and haplotypes were assigned using the LDLink 3.0 web tool, using the LDmatrix, LDpair and LDhap modules (https://analysistools.nci.nih.gov/LDlink/?tab=home). This tool investigates patterns of linkage disequilibrium across a variety of ancestral population groups, in this case the African population from the 1000 Genome project phase 3 (version 5). Linkage disequilibrium was measured by calculated D prime (D'), R squared (R^2) and goodness-of-fit (chi-square and p-value) with variant Rs number assigned by dbSNP used as input. The D' is an indicator of allelic segregation for two genetic variants. It ranges from 0 to 1 with higher values indicating tight linkage of alleles, whereas a low D' value or a value of 0 indicates no linkage of alleles. R squared (R^2) is a measure of correlation of alleles for two genetic variants. R^2 values range from 0 to 1 with higher values indicating a higher degree of correlation. An R^2 value of 0 indicates that alleles are independent, whereas an R^2 value of 1 indicates that an allele of one variant perfectly predicts an allele of another variant. Haplotypes were determined using the LDhap module, which calculates population specific haplotypes frequencies of all haplotypes observed for a list of query variants. Publicly available reference haplotype from the 1000 Genome project phase 3 (version 5) are used by LDlink to calculate population-specific measures of linkage disequilibrium [38]. Variant of minor allele frequencies <5% and those not assigned in the dbSNP (novel SNPs) were excluded from LD determination and haplotype assignment.

RESULTS

Single nucleotide polymorphisms (SNPs), detection of indels and verification

There is limited availability of APOBEC3 gene sequences from African populations, and when sequencing has been performed, it has generally been limited to A3G [39]. In this study, we applied next generation sequencing to determine variation in the coding exons of the APOBEC genes A3D, A3F, A3G and A3H in DNA from 192 HIV positive individuals residing in the Limpopo province of northern South Africa. The proteins expressed from these genes have all been shown to be capable of HIV restriction [5].

The A3D gene is 12.1kb long (Table 1) and has seven exons with exon 5 shown to display the most variation. Good quality A3D sequences after targeted DNA implication of the exons were successfully obtained for 165/192 subjects. In the DNA from these 165 individuals, 8 nonsynonymous and 2 synonymous changes were identified when compared to the GRCh37 build of the human genome (Table 3). Of the 165 subjects analyzed, 50.3% (83/165) were identified with nonsynonymous or synonymous changes in many positions in the coding region of the A3D gene, while no changes were detected in the remaining 49.7% (82/165). There were no insertions or deletions observed in A3D in the sequenced samples. Variant R248K was observed in 18.8% (31/165) of the patients, with 17.6% being heterozygotes. The R248K variant together with T316T was in strong linkage disequilibrium (LD) (D' =1, R^2= 0.122). Minor allele frequencies (MAF<5%) were observed in the case of L221L, C224Y, T238I and C320Y, therefore their LD profiles could not be calculated. Although no variants deviated from the Hardy-Weinberg Equilibrium (HWE) (P-value <0.05), R248K showed a tendency towards deviation with P-value =0.09. In addition, R6K, and T238I, variants that have not been reported elsewhere, were observed as heterozygotes in 9.7 and 4.8 % respectively (Table 3).

The A3F gene is 13.3kb long (Table 1) and contains seven exons, with the most variation observed in exon 4. The A3F exons were successfully amplified and sequenced from a total of 171/192 subjects. Synonymous or nonsynonymous changes were observed in 95.9%

(164/171) of the subjects while 4.1% (7/171) had no change relative to the GRCh37 human genome build. In the 171 samples successfully sequenced, there were seven nonsynonymous changes (R48P, A78V, I87L, P97L, A108S, V231I and Y307C) and seven synonymous changes (I117I, S118S, R143R, Y196Y, S229S, S327S and E245E). A108S and A78V were the most frequent nonsynonymous changes in A3F, found in 64.3% and 39.1% of the subjects respectively (Table 3). No insertions or deletions were observed for A3F in the sequenced samples. All A3F variants were in strong LD (D'= 1, R^2 range from 0.001- 0.899). A few variants (A108S, R143R and E245E) deviated from the HWE (P-value <0.05). The synonymous I117I mutation has not been reported previously (Table 3).

The A3G gene was the first one that was described as encoding an HIV restriction factor and it remains the most studied among Africans. The gene is 10.7 kb and has 8 exons. We successfully amplified A3G from 171/192 subjects. The most variation has been observed in exons 3, 6 and 7. A total of four nonsynonymous (H186R, R256H, Q275E and G363R) and four synonymous changes (S60S, A109A, F119F and L371L) were observed in A3G with the most frequent being H186R (53.8%) and Q275E (32.7%), Table 3. In total, nonsynonymous or synonymous changes were observed in 95.9% (164/171), whereas 4.1% (7/171) had no changes relative to the reference GRCh37 human genome. There were no insertions or deletions observed in this gene. Variants S60S, A109A, F119F H186R, R256H, and Q275E were in strong to partial LD (D'= 1, R2 range from 0-0.068. None of the variants deviated from the HWE (P-value > 0.05) (Table 3).

A3H is the shortest, but most polymorphic of the APOBEC3 genes we analyzed. It is 6.8kb in length (Table 1) and contains 5 exons, with the most variation in exons 1, 2 and 3. We observed nonsynonymous or synonymous changes in all the study subjects (175/175). We found 6 nonsynonymous (N15Δ, R18L, G105R, K121E, E140K and E178D) and 1 synonymous (T43T) changes. The N15Δ deletion was the only deletion observed and it occurred in 107 of 175 subjects (61.1%) either in a homozygous (57) or heterozygous (50) form. The G105R, K121E, E140K and E178D variants occurred mostly as homozygous forms in 90.3-100% of all

subjects (Table 3). No insertions were found. All variants in this gene are in strong LD (D'=1, R^2 range from 0.203- 1). The N15Δ and R18L variants deviated from the HWE (P-value >0.05), (Table 3).

Table 3: APOBEC 3D, 3F, 3G and 3H nonsynonymous and synonymous changes, genotype frequencies and Hardy Weinberg Equilibrium calculations from the study population.

APOBEC 3D Nonsynonymous changes (n= 165)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
R6K (NI)	A**G**A→A**A**A Transition	17G/G 17G/A	1	149 (90.3) 16 (9.7)	P-value = 0.51 X^2 = 0.43
R97C (rs75858538)	**C**GC→**T**GC Transition	289 C/C 289 C/T 289 T/T	1	148 (89.7) 16 (9.7) 1 (0.6)	P-value = 0.44 X^2 = 0.59
L221R (NI)	C**T**G→C**G**G Transition	662T/T 662T/G	5	162 (98.2) 3 (1.8)	P-value = 0.9 X^2 = 0.01
C224Y (rs772893975)	T**G**T→T**A**T Transition	671G/G 671G/A	5	159 (96.4) 6 (3.6)	P-value = 0.812 X^2 = 0.056
T238A (rs201709403) **T238I** (NI)	**A**CA→**G**CA Transition **A**CA→A**T**A Transition	712A/A 712A/G 713A/T	5	142 (86.1) 15 (9.1) 8 (4.8)	P-value = 0.218 X^2 = 1.5
R248K (rs61748819)	A**G**G→A**A**G	743 G/G	5	134 (81.2)	P-value = 0.09

	Transition	743 G/A 743 A/A		29 (17.6) 2 (1.2)	$X^2 = 0.76$
C320Y (rs61999342)	T**G**C→T**A**C Transition	959 G/G 959 G/A	6	164 (99.4) 1 (0.6)	P-value = 0.94 $X^2 = 0.006$
APOBEC3D Synonymous changes					
L221L (rs769426665)	CT**G** →CT**C** Transition	663G/G 663G/C	5	162 (98.2) 3 (1.8)	P-value = 0.97 $X^2 = 0.02$
T316T (rs184448269)	AC**C**→AC**T** Transition	948 C/C 948 C/T	6	158 (95.7) 7 (4.3)	P-value =0.84 $X^2 = 0.04$
No mutation				**82 (49.7)**	
APOBEC 3F Nonsynonymous changes (n= 171)					
R48P (rs35053197)	C**G**T→C**C**C Transversion	143 G/G 143 G/C	2	156 (91.2) 15 (8.8)	P-value = 0.54 $X^2 = 0.36$
A78V (rs5750728)	G**C**C→G**T**C Transition	233 C/C 233 C/T 233 T/T	4	108 (63.2) 59 (36.8) 4 (2.3)	P-value = 0.9 $X^2 = 0.04$
I87L (rs146543452)	**A**TC→**C**TC Transition	260 A/A 260 A/C	4	162 (94.7) 9 (5.3)	P-value = 0.72 $X^2 = 0.12$
P97L (rs201939303)	C**C**G→C**T**G Transition	290 C/C 290 C/T	4	168 (98.2) 3 (1.8)	P-value = 0.9 $X^2 = 0.04$
A108S (rs2020390)	**G**CT→**T**CT Transition	322 G/G 322 G/T		61 (35.7) 92 (53.8)	P-value = 0.05 $X^2 = 3.77$

		322 T/T	4	18 (10.5)	
V231I (rs2076101)	**G**TC→**A**TC Transition	898 G/G 898 G/A 898 A/A	5	139 (81.3) 30 (17.5) 2 (1.2)	P-value = 0.99 X^2 = 0.07
Y307C (rs12157816)	T**A**C→T**G**C Transition	920 A/A 920 A/G	6	156 (91.2) 15 (8.8)	P-value = 0.55 X^2 = 0.36
APOBEC3F Synonymous changes					
I117I (NI)	AT**C**→AT**T** Transition	351 C/C 351 C/T	4	169 (98.8) 2 (1.2)	P-value =0.94 X^2 = 0.003
S118S (rs35928287)	TC**C**→TC**T** Transition	354 C/C 354 C/T	4	125 (73.1) 46 (26.9)	P-value =0.04 X^2 = 4.69
R143R (rs4821862)	CG**C**→CG**T** Transition	429 C/C 429 C/T 429 T/T	4	23 (13.5) 97 (56.7) 51 (29.8)	P-value = 0.03 X^2 = 4.69
Y196Y (rs765418322)	TA**T**→TA**C** Transition	588 T/T 588 T/C 588 C/C	4	118 (69) 49 (28.7) 4 (2.3)	P-value = 0.7 X^2 = 0.17
S229S (rs549550231)	TC**A**→TC**G** Transition	894 A/A 894 A/G	5	169 (98.8) 2 (1.2)	P-value = 0.94 X^2 = 0.005
E245E (rs113109079)	GA**G**→GA**A** Transition	942 G/G 942 G/A 942 A/A	5	161 (94.2) 9 (5.2) 1 (0.6)	P-value = 0.04 X^2 = 4.08
S327S (rs35895636)	TC**C**→TC**T** Transition	981 C/C 981 C/T 981 T/T	5	143 (83.6) 25 (14.6) 3 (1.8)	P-value = 0.14 X^2 = 2.189

No mutation				**7 (4.1)**	
APOBEC 3G Nonsynonymous changes (n= 171)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
H186R (rs8177832)	C**A**C→C**G**C Transition	557 A/A 557 A/G 557 G/G	4	79 (46.2) 74 (43.3) 18 (10.5)	P-value = 0.89 $X^2 = 0.02$
R256H (rs17000736)	C**G**C→C**A**C Transition	767 G/G 767 G/A	6	167(97.7) 4 (2.3)	P-value = 0.97 $X^2 = 0.00$
Q275E (rs17496046)	**C**AG→**G**AG Transversion	823 C/C 823 C/G 823 G/G	6	115 (67.3) 50 (29.2) 6 (3.5)	P-value = 0.84 $X^2 = 0.03$
G363R (rs148267053)	**G**GA→**A**GA Transition	1087 G/G 1087 G/A	7	154 (90.1) 17 (9.9)	P-value = 0.5 $X^2 = 0.3$
APOBEC3G Synonymous changes					
S60S (rs112603901)	TC**C**→TC**T** Transition	180 C/C 180 C/T	3	152 (88.9) 19 (11.1)	P-value =0.59 $X^2 = 0.4$
A109A (rs375760983)	GC**C**→GC**T** Transition	327 C/C 327 C/T	3	170 (99.4) 1 (0.6)	P-value =0.97 $X^2 = 0.00$
F119F (rs5757465)	TT**T**→TT**C** Transition	357 T/T 357 T/C	3	170 (99.4) 1 (0.6)	P-value =0.97 $X^2 = 0.00$
L371L (rs11545130)	**C**TG→**T**TG Transition	1111 C/C 1111 C/T	7	164 (95.9) 7 (4.1)	P-value =0.8 $X^2 = 0.005$

No mutation				**7 (4.1)**	
APOBEC 3H Nonsynonymous changes (n= 175)					
CDS codon position change & rs numbers	**Type of change**	**Genotypes**	**Exon**	**Frequencies (%)**	**Hardy Weinberg equilibrium**
N15Δ (rs139292)	**-CAA** Deletion	45 CAA/CAA 45 CAA/ Δ 45 Δ / Δ	1	68 (38.9) 50 (28.6) 57 (32.5)	P-value = 0.00 X^2 = 31.8
R18L (rs139293)	C**G**C→C**T**C Transvertion	53 G/G 53 G/T 53 T/T	1	156 (89.1) 15 (8.6) 4 (2.3)	P-value = 0.00 X^2 = 15.9
G105R (rs139297)	**G**GC→**C**GC Transvertion	313 G/G 313 G/C 313 C/C	2	0 4 (2.3) 171 (97.7)	P-value = 0.87 X^2 = 0.022
K121E (rs139298)	**A**AG→**G**AG Substitution	361 A/A 361 A/G	2	0 175 (100)	P-value = 0.59 X^2 = 0.3
E140K (rs139300)	**G**AG→**A**AG Transvertion	419 G/G 419 A/A	2	0 175 (100)	P-value = 0.47 X^2 = 0.3
E178D (rs139302)	**G**AG→GA**C** Transvertion	534 G/G 534 G/C 534 C/C	3	0 17 (9.7) 158 (90.3)	P-value = 0.497 X^2 = 0.45
APOBEC3H Synonymous changes					
T43T (rs139294)	AC**G**→AC**C** Transvertion	129 G/G 129 C/C	1	31 (17.7) 144 (82.3)	P-value =0.97 X^2 = 0.00
No mutation				**0**	

NI= Not Identified;

Nucleotide in bold: Codon specific change

In order to better understand the A3 genetic changes observed in each subject, all clusters of variation per gene were assigned into haplotypes as described in materials and methods and their frequencies calculated. These haplotypes were classified as either confirmed or unconfirmed based on the number of heterozygous variants. This classification was necessary due to the fact that NGS reads were short and we could not confirm if SNPs occurred in the same allele (Table 4). Nonsynonymous variants were considered and their genotypes (homozygous or heterozygous) were indicated. Low frequency variants (MAF<5%) were excluded from the haplotype assignment. Comparisons were made to the GRCh37 human genome whose combinations are represented as haplotypes in A3D, A3F and A3G (Table 4). We identified four haplotypes for A3D, five for A3F and four for A3G (Table 4). The A3D, A3F and A3G haplotypes were comprised of 3, 6 and 3 variants respectively. It is worth noting that only haplotypes for A3G and A3H have been described previously [12], [15], [40], [41], (Table 4). The seven identified haplotypes of A3H were recently described as having an impact on the genetic diversity of HIV-1 Vifs in the global pandemic [12], [15], [16]. The seven known A3H haplotypes (I-VII) are combinations of 5 nucleotide changes located on exons 2, 3 and 4. Haplotypes II, V, and VII have been termed stable, because of the observed relatively long half-lives of the encoded proteins, enabling them to restrict HIV-1. The other four haplotypes (haplotypes I, III, IV, VI) have been termed unstable, since the encoded protein half-lives have been shown to be short, resulting in complete loss of the ability to restrict HIV [12], [39] In our subjects, we identified 3 haplotypes for A3H: the stable haplotype II (15N, 18R, 105R, 121E 178D), haplotype III (15Δ , 18R, 105R, 121E, 178D) and haplotype IV (15 Δ, 18L, 105R, 121E, 178D) (Table 4) [11], [12], [39] The prevalence of these haplotypes was: 30.8% haplotype II; 56.6% haplotype III and 8.6% confirmed and 4% unconfirmed haplotype IV (Table 4). Thus the

majority of the subjects in our patient population are not expressing ApoBec3H proteins that can restrict HIV.

Table 4: Haplotype formation and the frequencies with respect to the A3D, A3F, A3G and A3H

Confirmed APOBEC3D Haplotypes		
Variation	**Frequency (%)**	**Haplotype**
97R, 238T, 248R	82 (49.7)	Haplotype i
97C (hom), 238T, 248R **97C (het)**, 238T, 248R	1 (0.6) 16 (9.7)	Haplotype ii
97R, **238A (het)**, 248R	11 (6.7)	Haplotype iIi
97R, 238T, **248K (hom)** 97R, 238T, **248K (het)**	2 (1.2) 29 (17.6)	Haplotype iv
Minor variant frequency <5%	8 (4.8))	Not assigned
Others[a]	16 (9.7)	Not assigned
Unconfirmed APOBEC3D Haplotypes **None**		
Confirmed APOBEC3F Haplotypes		
48R, 78A, 87I, 108A, 231V, 307Y	7 (4.1)	Haplotype i
48R**, 78V (hom)**, 87I, **108S (hom)**, **231I (hom)**, 307Y 48R**, 78V (het)**, 87I, **108S (hom)**, **231I (hom)**, 307Y	1 (0.6) 1 (0.6)	Haplotype ii
48R, 78A, 87I, **108S (hom)**, 231V, 307Y 48R, 78A, 87I, **108S (het)**, 231V, 307Y	2 (1.2) 37 (21.6)	Haplotype iii

48R, 78A, 87I, 108A, 231V, **307C (het)**	4 (2.3)	Haplotype iv
48R, **78V (hom)**, 87I, **108S (hom)**, 231V, 30Y	1 (0.6)	Haplotype vi
Minor variant frequency <5%	7 (4.1)	Not assigned
Others[a]	47 (27.5)	Not assigned
Unconfirmed APOBEC3F Haplotypes		
48R**, 78V (het)**, 87I, **108S (het)**, **231I (het)**, 307Y	30 (17.5)	Haplotype ii
48P (het), 78A, 87I, **108S (het)**, 231V, 307Y	9 (5.3)	Haplotype v
48R, **78V (het)**, 87I, **108S (het)**, 231V, 307Y	25 (14.6)	Haplotype vi
Confirmed APOBEC3G Haplotypes		
186H, 275Q, 363G	7 (4.1)	Haplotype i
186R (hom), 275Q, 363G **186R (het)**, 275Q, 363G	20 (11.7) 42 (24.6)	Haplotype ii
186H, **275E (hom)**, 363G 186H, **275E (het)**, 363G	7 (4.1) 15 (8.8)	Haplotype iii
186H, 275Q, **363R (het)**	12 (7)	Haplotype iv
Minor variant frequency <5%	12 (7)	Not assigned
Others[a]	56 (32.7)	Not assigned
Unconfirmed APOBEC3G Haplotypes **None**		
Confirmed APOBEC3H Haplotypes[b]		
15N, 18R, **105R (hom)**, **121E (hom)**, **178D (hom)**	54 (30.8)	Haplotype ii

15Δ (hom), 18R, **105R (hom)**, **121E (hom)**, **178E (hom)** **15Δ (het)**, 18R, **105R (hom)**, **121E (hom)**, **178D (hom)**	50 (28.6) 49 (28)	Haplotype iii
15Δ, **(hom), 18L (hom)**, **105R (hom)**, **121E (hom)**, **178D (hom)** **15Δ**, **(hom), 18L (het)**, **105R (hom)**, **121E (hom)**, **178D (hom)**	7 (4) 8 (4.6)	Haplotype iv
Unconfirmed APOBEC3H Haplotypes[b]		
15Δ, **(het), 18L (het)**, **105R (hom)**, **121E (hom)**, **178D (hom)**	7 (4)	Haplotype iv

[a] Refers to the haplotypes with synonymous changes and those of novel SNPs (not reported on the dpSNP)

[b]A3H haplotypes were determined using previous classification from references [11, 12, 37]

hom=homozygous; het=heterozygous

Bold defines variants that are different from those listed in haplotype I in each gene.

Haplotypes are unconfirmed in our population due to more than 1 heterozygous SNPs in the cluster

Allele frequencies and their comparison with other populations

We next compared the nonsynonymous and synonymous variant frequencies in the South African population in our study to previously reported variant frequencies in the following populations: African (AFR), East Asian (EAS), European (EUR), Ad Mixed American (AMR), and South Asian (SAS), as reported in the 1000 Genome Project phase III, the HapMap project (NCBI), the dsSNP database and the Ensembl genome browser (Table 5). Our data show that with the exception of, V231I, (A3F) and F119F (A3G), all the variants identified showed a higher frequency than previously reported for the overall AFR population (Table 5). In a previous study by Duggal and colleagues comparing A3 variations between Africans, Asian and Europeans, nonsynonymous variation in A3D (R97C, R248K); A3F (A108S, V231I, Y307C); A3G (H186R,

E275Q) and A3H (15Δ, R18L, R105G, E121E/K, E178D) were reported [39]. Our data suggest that several variants occur more frequently in our South African population than in the "African" population previously studied [13]. This included R97C and R248K in A3D; A108S in A3F; H186R, Q275E in A3G and N15Δ, R18L, G105R, K121E and E178D in A3H (Table 5). Of particular interest was the A3F A108S variant, which occurs in about 80% of Asians and has been reported in only 30% of Africans [13]. We found this variant in 64.3% of our subjects, more than twice the prevalence previously reported among Africans by Duggal and colleagues [13].

Overall, the EAS, EUR, AMR and SAS populations showed a higher level of A3 conservation than our study population. For example, the A3D sequences in these populations were more closely related to the reference GRCh37 human genome (98-100%). In the case of A3F and A3G, the percent identity was between 92-100%, with the exception of the A3F variants A78V, A108S, V231I, R143R and the A3G variant F119F (Table 5). The percentage identity among the A3H variants were highly variable, with the N15**Δ** variant clearly present in higher frequency in our population compared to the others (Table 5).

Table 5: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population and total populations from East Asian (**EAS**), European (**EUR**), African (**AFR**), Ad Mixed American (**AMR**) and South Asian (**SAS**).

APOBEC 3D Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
R6K (NI)	G (**R**) A (**K**)	90.3 9.7	NA	NA	NA	NA	NA
R97C (rs758585538)	C (**R**) T (**C**)	89.7 10.3	100 0	100 0	96.6 3.4	100 0	100 0
L221R (NI)	T (**L**) G (**R**)	98.2 1.8	NA	NA	NA	NA	NA
C224Y (rs772893975)	G (**C**)	96.4	100	100	100	100	100

	A (**Y**)	3.6	0	0	0	0	0
T238A (rs201709403)	A (**T**) G (**A**)	86.1 9.1	100 0	100 0	100 0.001	100 0	100 0
T238I (NI)	A (**T**) T (**I**)	86.1 4.8	NA	NA	NA	NA	NA
R248K (rs61748819)	G (**R**) A (**K**)	81.2 18.8	100 0	100 0	89. 11	98.9 4	100 0
C320Y (rs61999342)	G (**C**) A (**Y**)	99.4 0.6	NA	NA	NA	NA	NA
APOBEC3D synonymous allele frequencies (%)							
L221L (rs769426665)	G (**L**) C (**L**)	98.2 1.8	100 0	100 0	100 0	100 0	100 0
T316T (rs184448269)	C (**T**) T (**T**)	95.7 4.3	100 0	100 0	98.8 1.2	99.6 4.3	100 0
APOBEC 3F Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
R48P (rs35053197)	G (**R**) C (**P**)	91.2 8.8	100 0	NA	97 3	99 1	99 1
A78V (rs5750728)	C (**A**) T (**V**)	63.2 36.8	29 79.3	58.8 49.1	79.8 22	37.6 62.4	39.1 60.8
I87L (rs114704208)	A (**I**) C (**L**)	94.7 5.3	100 0	100 0	99 1	100 0	100 0
P97L (rs201939303)	C (**P**) T (**L**)	98.2 1.8	NA	NA	NA	NA	NA
A108S (rs2020390)	G (**A**) T (**S**)	35.7 64.3	29 71	51 49	68 32	37 63	40 60

V231I (2076101)	G (**V**) A (**I**)	81.3 18.7	71 29	51 49	81 19	38 62	39 61
Y307C (rs12157816)	A (**Y**) G (**C**)	91.2 8.8	100 0	98 2	97 3	98 2	100 0
APOBEC3F synonymous allele frequencies (%)							
I117I (NI)	C (**I**) T (**I**)	98.8 1.2	NA	NA	NA	NA	NA
S118S (rs35928287)	C (**S**) T (**S**)	73.1 26.9	100 0	100 0	97 3	100 0	100 0
R143R (rs4821862)	C (**R**) T (**R**)	13.5 86.5	29 71	51 49	45 55	36 64	39 61
Y196Y (rs765418322)	T (**Y**) C (**Y**)	69 31	100 0	100 0	100 0	100 0	100 0
S229S (549550231)	A (**S**) T (**S**)	98.8 1.2	NA	NA	NA	NA	NA
E245E (113109079)	G (**E**) A (**E**)	94.2 5.8	100 0	100 0	99 1	100 0	100 0
S327S (35895636)	C (**S**) T (**S**)	83.6 16.4	100 0	100 0	99 2	100 0	100 0
APOBEC 3G Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
H186R (rs8177832)	A (**H**) G (**R**)	46.2 53.8	92.7 7.2	97.2 2.98	56.9 43	92.5 7.5	99.1 0.82
R256H (rs17000736)	G (**R**) A (**H**)	97.7 2.3	100 0	100 0	98.6 1.44	100 0	100 0
Q275E (rs17496046)	C (**Q**) G (**E**)	67.3 32.7	97.3 2.68	94.6 5.3	87.5 12.5	96 4	98.7 1.3

G363R (rs148267053)	G (**G**) A (**R**)	90.1 9.9	100 0	100 0	98.6 1.4	99.9 1	100 0
APOBEC3G synonymous allele frequencies (%)							
S60S (rs112603901)	C (**S**) T (**S**)	88.9 11.1	100 0	100 0	99.7 0.7	100 0	100 0
A109A (rs375760983)	C (**A**) T (**A**)	99.4 0.6	100 0	100 0	NA	NA	NA
F119F (rs5757465)	T (**F**) C (**F**)	99.4 0.6	77.6 22.4	55.3 44.7	97.1 28.7	60.2 39.7	55.5 44.5
L371L (rs11545130)	C (**L**) T (**L**)	95.9 4.1	97 3	100 0	100 0	100 0	100 0
APOBEC 3H Nonsynonymous allele frequencies (%)							
Rs numbers	**Allele**	**SA**	**EAS**	**EUR**	**AFR**	**AMR**	**SAS**
N15Δ (rs139292)	CAA(**N**) Δ	38.9 61.1	69 31	72 28	74 26	66 34	60 40
R18L (rs139293)	G (**R**) T (**L**)	89.1 10.9	84.1 15.9	70.7 29.3	93.3 0.6	75.8 24.2	69.4 30.6
G105R (rs139297)	C (**G**) G (**R**)	0 100	33.1 66.9	46.3 53.7	87.5 12.5	38.3 61.7	42.9 57.1
K121E (rs139298)	C (**K**) G (**E**)	0 100	33.1 66.9	46.3 53.7	84.7 15.3	34.6 65.4	43.1 56.9
E140K (rs139300)	G (**E**) A (**K**)	0 100	0 100	0 100	0 100	0 100	0 100
E178D (rs139302)	G (**E**) C (**D**)	0 100	83.3 66.6	45.4 54.6	82.7 17.3	37.7 62.3	43.9 56.1
APOBEC3H synonymous allele frequencies (%)							
T43T (rs139294)	G (**T**)	17.7	33.3	45.4	82.7	37.7	43.8

	C (**T**)	82.3	66.6	54.6	17.3	62.3	56.2

Note: NI= Not Identified

(NA)= No data available

The term “Africans” has been loosely used to describe datasets generated from different parts of the African continent. To provide a more accurate comparison, we compared the variants detected in our study with previous datasets derived from other more specific studies in African populations or involving people of African descent (Table 6). These included Americans of African Ancestry in USA (ASW); African Caribbeans in Barbados (ACB); Gambians in the Western Gambia (GWD); Esan in Nigeria (ESN); Luhya in Webuye, Kenya (LWK); Mende in Sierra Leone (MSL) and Yoruba in Ibadan, Nigeria (YRI). We noticed significantly higher level of single nucleotide changes in our population for the following variants: R48P, A108S, R143R, S327S in A3F; Q275E, G363R in A3G) and G105R, K121E, E178D in A3H (Table 6). However, in most cases, the variant allele frequencies in A3D, A3F, A3G and A3H were closer (Table 6). Notably, the variant frequencies of R97C and R248K in A3D are almost close to frequencies in ASW and ESN and MSL respectively, whereas the frequency of R48P in A3F is closer to that reported in, ESN, LWK and YRI. The frequency of I87L was close to that observed in GWD, LWK and MSL. The frequency of V231I was close to that observed in all Africans, whereas the frequency of Y307C was close to the frequencies observed in ACB and YRI. The frequency of E245E was close to that observed in ASW, GWD, MSL and YRI. For A3G variants, the frequency observed for H186R is similar to that observed in ESN and MSL. The frequency of R256H and L371L were similar among all Africans. In the case of A3H, information about the N15Δ variant was not available from any of the previous African studies, while the R18L variant was similar in prevalence among all Africans (Table 6).

Table 6: Comparison of A3D, A3F, A3G and A3H allele frequencies between a South African population and other Africa population, which comprises of the African Carribeans in Barbados (**ACB**), Americans of African Ancestry in USA (**ASW**), Esan in Nigeria (**ESN**), Gambian in the

Western Gambia (**GWD**), Luhya in Webuye, Kenya (**LWK**), Mende in Seirra Leone (**MSL**), Yoruba in ibadan, Nigeria (**YRI**).

APOBEC 3D Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
R6K (NI)	G (**R**) A (**K**)	90.3 9.7	NA	NA	NA	NA	NA	NA	NA
R97C (rs758585538)	C (**R**) T (**C**)	89.7 10.3	97 3	94 6	97 3	98 2	94 6	98 2	97 3
L221R (NI)	T (**L**) G (**R**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
C224Y (rs772893975)	G (**C**) A (**Y**)	96.4 3.6	NA	NA	NA	NA	NA	NA	NA
T238A (rs201709403)	A (**T**) G (**A**)	86.1 9.1	100 0	99 1	100 0	100 0	100 0	100 0	100 0
T238I (NI)	A (**T**) T (**I**)	86.1 4.8	NA	NA	NA	NA	NA	NA	NA
R248K (rs61748819)	G (**R**) A (**K**)	81.2 18.8	91 9	96 4	86 14	92 8	87 13	86 14	87 13
C320Y (rs61999542)	G (**C**) A (**Y**)	99.4 0.6	NA	NA	NA	NA	NA	NA	NA
APOBEC3D synonymous allele frequencies (%)									
L221L (rs769426665)	G (**L**) C (**L**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
T316T (rs184448269)	C (**T**) T (**T**)	95.7 4.3	99 1	99 1	98 2	99 1	99 1	98 2	99 1
APOBEC 3F Nonsynonymous allele frequencies (%)									

Rs numbers	Allele	SA	ACB	ASW	ESN	GWD	LWK	MSL	YRI
R48P (rs35053197)	G (**R**) C (**P**)	91.2 8.8	98 2	98 2	95 5	99 1	95 5	98 2	95 5
A78V (rs5750728)	C (**A**) T (**V**)	63.2 36.8	NA	NA	NA	NA	NA	NA	NA
I87L (rs146543452)	A (**I**) C (**L**)	99.4 0.6	100 0	100 0	100 0	98 2	99 1	99 1	100 0
P97L (rs201939303)	C (**P**) T (**L**)	98.2 1.8	NA	NA	NA	NA	NA	NA	NA
A108S (rs2020390)	G (**A**) T (**S**)	35.7 64.3	67 33	66 34	69 31	62 38	65 35	76 24	70 30
V231I (2076101)	G (**V**) A (**I**)	81.3 18.7	79 21	73 27	84 16	78 22	80 20	85 15	87 13
Y307C (rs12157816)	A (**Y**) G (**C**)	91.2 8.8	95 5	98 2	96 4	98 2	98 2	96 4	95 5
APOBEC3F synonymous allele frequencies (%)									
I117I (NI)	C (**I**) T (**I**)	98.8 1.2	NA	NA	NA	NA	NA	NA	NA
S118S (rs35928287)	C (**S**) T (**S**)	73.1 26.9	NA	NA	NA	NA	NA	NA	NA
R143R (rs4821862)	C (**R**) T (**R**)	13.5 86.5	45 55	45 55	39 61	50 50	45 55	46 54	47 53
Y196Y (rs765418322)	T (**Y**) C (**Y**)	69 31	NA	NA	NA	NA	NA	NA	NA
S229S (549550231)	A (**S**) T (**S**)	98.8 1.2	NA	NA	NA	NA	NA	NA	NA

E245E (113109079)	G (**E**) A (**E**)	94.2 5.8	99 1	98 2	100 0	98 2	99 1	98 2	98 2
S327S (35895636)	C (**S**) T (**S**)	83.6 16.4	98 2	99 1	96 4	100 0	97 3	100 0	98 2
APOBEC 3G Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
H186R (rs8177832)	A (**H**) G (**R**)	46.2 53.8	56 44	75 25	49 51	57 43	68 32	49 51	52 48
R256H (rs17000736)	G (**R**) A (**H**)	97.7 2.3	98 2	98 2	99 1	97 3	98 2	99 1	99 1
Q275E (rs17496046)	C (**Q**) G (**E**)	67.3 32.7	90 10	91 9	86 14	87 13	83 17	91 9	87 13
G363R (rs148267053)	G (**G**) A (**R**)	90.1 9.9	98 2	99 1	100 0	98 2	99 1	98 2	98 2
APOBEC3G synonymous allele frequencies (%)									
S60S (rs112603901)	C (**S**) T (**S**)	88.9 11.1	99 1	98 2	100 0	100 0	99 1	100 0	100 0
A109A (rs375760983)	C (**A**) T (**A**)	99.4 0.6	NA	NA	NA	NA	NA	NA	NA
F119F (rs5757465)	T (**F**) C (**F**)	99.4 0.6	93 7	89 11	100 0	98 2	98 2	99 1	99 1
L371L (rs11545130)	C (**L**) T (**L**)	95.9 4.1	98 2	98 2	97 3	98 2	95 5	98 2	95 5
APOBEC 3H Nonsynonymous allele frequencies (%)									
Rs numbers	**Allele**	**SA**	**ACB**	**ASW**	**ESN**	**GWD**	**LWK**	**MSL**	**YRI**
N15Δ (rs139292)	CAA (**N**) Δ	38.9 61.1	NA	NA	NA	NA	NA	NA	NA

R18L (rs139293)	G (**R**) T (**L**)	89.1 10.9	93 7	87 13	93 7	94 6	94 6	93 7	96 4
G105R (rs139297)	C (**G**) G (**R**)	0 100	85 15	75 25	90 10	87 13	91 9	89 11	91 9
K121E (rs139298/99)	A (**K**) G (**E**)	0 100	85 15	75 25	90 10	87 13	91 9	89 11	91 9
E140K (rs139300)	G (**E**) A (**K**)	0 100	100 0	100 0	100 0	100 0	100 0	100 0	100 0
E178D (rs139302)	G (**E**) C (**D**)	0 100	17 83	29 71	11 89	16 84	16 84	13 87	11 89
APOBEC3H synonymous allele frequencies (%)									
T43T (rs139294)	G (**T**) C (**T**)	17.7 82.3	22 78	30 70	16 84	16 84	13 87	17 83	13 87

Note: NI= Not identified

(NA)= No data available

DISCUSSION

In this study we characterized SNPs and indels within the coding exons of several human APOBEC3 genes (A3D, A3F, A3G and A3H) to document the level of diversity in these genes in a diverse South African population residing in the Limpopo Province in Northern South Africa. We observed a high level of A3 diversity and a higher prevalence of certain variants than has previously been observed in other African populations. Interestingly, some of these variants have previously been linked to HIV disease progression [14], [39], [42]. The use of next generation sequencing also allowed the identification of genetic variants that were not previously identified using methods such as TaqMan assays, restriction fragment length polymorphism (RFLP) or Sanger sequencing [39].

Common variants in APOBEC3 genes have been intensively studied and many have been found to have differential effects on antiviral activity [7], [13], [14], [39], [42]. For example, the variants R97C and R248K in A3D were shown to moderately decrease antiviral activity [13]. In contrast, the A3F variants A108S, V231I and Y307C have been reported to have potent antiviral activity against HIV-1 ΔVif strains [43], [44]. Among the A3F variants, A108S was the most interesting in our study, as its prevalence was twice that reported previously for the "African" population (Table 6), but very similar to reports from EAS, EUR, AMR and SAS populations (Table 5). It will be of clear interest to perform additional studies to determine the prevalence of this variant in other regions of South Africa. SNPs in A3G can alter its antiviral activity and sometimes enhance the rate of HIV-1 disease progression, as reported in a cohort of HIV-1 subtype C infected South African women and a US based cohort of African Americans [14], [39]. In particular, the H186R variant has previously been associated with more rapid decline in CD4+ cells and accelerated disease progression [14], [39], [42]. Our study show that this variant is similar in prevalence in our population as in some other African populations (Table 6), but significantly higher than reported in a previous study performed in Durban, South Africa [39]. This difference could be explained either by differences in the ethnicity of the study population or the genotyping method used, since the previous study used TaqMan genotyping rather than the NGS methodology used in our study.

Recent studies have shown A3H as the most polymorphic member of the A3 family. The A3H variants (15Δ, R18L, G105R, K121E, E178D) which make up 7 different haplotypes have been functionally described in other studies, showing varying protein expression and stability [8], [11], [16], [45]–[48]. Data from the 1000 genome project suggest that stable A3H haplotypes (II, V and VII) predominate in Africa while unstable haplotypes (I, III, IV, VI) are more prevalent in Asia [15], Interestingly, the unstable A3H haplotypes III and IV (which encode complete loss-of-function proteins) were unexpectedly dominant among our study population. This can be attributed mainly to the high prevalence (69.2%) of the deletion at amino acid residue 15 (Tables 3, 4, 5 and 6). This prevalence is inversely proportional to that reported in previous

studies of Africans in which stable A3H haplotypes were dominant (56%) [15]. Data from two recent studies illustrate that stable A3H haplotypes may function as contemporary HIV-1 restriction factors, contributing to limiting viral replication and rates of transmission [12], [15]. It is unclear what role, if any, the unstable A3H haplotype III and IV may play in the high prevalence and transmission of HIV-1 in Limpopo.

Because HIV-1 Vif acts as an antagonist to APOBEC proteins including A3H, we speculate that the distribution of stable versus unstable A3H haplotypes in our study might also influence Vif variation in HIV in our study population. Studies performed in primary CD4+ lymphocytes have shown that HIV-1 Vif variants with certain amino acid residues (F39 and H48), known as hyper Vifs, are better capable of neutralizing stable A3H genotypes, implying that HIV-1 Vif might adapt to the A3H haplotype in a particular population [15]. We are presently analyzing HIV-1 Vif sequences from our study subjects in order to determine a possible correlation between the A3H haplotypes and HIV-1 Vif genetic variation in this rural area of South Africa.

One limitation of our study was that all the subjects were HIV infected and were mostly at the chronic stage of infection. It is possible that HIV-1 negative subjects would present a different A3 profile. If this is the case, it could imply that these A3 genotypes either alone or in combination might influence HIV transmission. This is thus worthy of further studies. Although we did not quantify the A3 mRNA levels in this study, reports comparing HIV-1 non-controllers versus long-term non-progressors (LTNP) have shown that LTNPs express a higher level of A3G and A3F proteins [49] .

In conclusion, we have shown that significant A3 variation exists among HIV patients in an ethnically diverse population in Northern South Africa, by providing extensive data for 4 different A3 genes that are known to restrict HIV infection, but have previously only been sparsely studied in African populations. Our NGS results provide a baseline for future studies

that could functionally characterize the SNPs identified in the ApoBec proteins in this population and specifically analyze how they affect restriction of HIV replication and Vif function. Such studies will serve to increase our understanding of how the ApoBec landscape might have shaped the HIV epidemic in Northern South Africa.

CONFLICT OF INTEREST

The authors declare no conflict of interest

ACKNOWLEDGEMENT

Research reported in this publication was supported by the Myles H. Thaler Research Endowment at the University of Virginia and the South African Medical Research Council (RCDI) through funding received from the South African National Treasury and the South African National Research Foundation (GUN109312, GUN86037). The contents are solely the responsibility of the authors and do not necessarily represent the official views of the University of Virginia, the South African Medical Research Council or the National Research Foundation.

Nontokozo D. Matume was supported by the Research Capacity Development Initiative of the Medical Research Council (RCDI project number: 57009), and the Fogarty International Center/NIH (D43TW006578) and research funds from the Myles H. Thaler Center for AIDS and Human Retrovirus Research at the University of Virginia.

Denis M. Tebit was supported by funds from the Myles H. Thaler Center for AIDS and Human Retrovirus Research at the University of Virginia, and also received partial support through a Carnegie African Diaspora Fellowship Award.

David Rekosh was partially supported by funds from the Myles H. Thaler Professorship at the University of Virginia.

Marie-Louise Hammarskjold was partially supported by funds from the Charles H. Ross Jr. Professorship at the University of Virginia.

The authors are grateful to the study participants; Jing Huang at the Myles Thaler Center for Human Retrovirus Research at the University of Virginia, USA for assisting with NGS, and Elizabeth Mashu Etta of the HIV/ AIDS & Global Health Research Programme, University of Venda for assisting with sample collection and processing.

REFERENCES

1. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature. 2002;418(6898):646–50.

2. Chiu Y-L, Greene WC. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. Annu Rev Immunol. 2008;26(1):317–53.

3. Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. Vol. 4, Nature Reviews Immunology. 2004. p. 868–77.

4. Zhang H, Yang B, Pomerantz RJ, Zhang C, Arunachalam SC, Gao L. The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. Nature. 2003;424(6944):94–8.

5. Hultquist JF, Lengyel JA, Refsland EW, LaRue RS, Lackey L, Brown WL, et al. Human and Rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H Demonstrate a Conserved Capacity To Restrict Vif-Deficient HIV-1. J Virol. 2011;85(21):11220–34.

6. Refsland EW, Hultquist JF, Harris RS. Endogenous origins of HIV-1 G-to-A hypermutation and restriction in the nonpermissive T cell line CEM2n. PLoS Pathog. 2012;8(7):39.

7. An P, Penugonda S, Thorball CW, Bartha I, Goedert JJ, Donfield S, et al. Role of APOBEC3F Gene Variation in HIV-1 Disease Progression and Pneumocystis Pneumonia. PLoS Genet. 2016;12(3).

8. Harari A, Ooms M, Mulder LCF, Simon V. Polymorphisms and Splice Variants Influence the Antiretroviral Activity of Human APOBEC3H. J Virol. 2009;83(1):295–303.

9. Dang Y, Wang X, Esselman WJ, Zheng Y-H. Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. J Virol. 2006;80(21):10522–33.

10. OhAinle M, Kerns JA, Li MMH, Malik HS, Emerman M. Antiretroelement Activity of APOBEC3H Was Lost Twice in Recent Human Evolution. Cell Host Microbe. 2008;4(3):249–59.

11. Wang X, Abudu A, Son S, Dang Y, Venta PJ, Zheng Y-H. Analysis of human APOBEC3H haplotypes and anti-human immunodeficiency virus type 1 activity. J Virol. 2011;85(7):3142–52.

12. Ooms M, Brayton B, Letko M, Maio SM, Pilcher CD, Hecht FM, et al. HIV-1 Vif adaptation to human APOBEC3H haplotypes. Cell Host Microbe. 2013;14(4):411–21.

13. Duggal NK, Fu W, Akey JM, Emerman M. Identification and antiviral activity of common polymorphisms in the APOBEC3 locus in human populations. Virology. 2013;443(2):329–37.

14. An P, Bleiber G, Duggal P, Nelson G, May M, Mangeat B, et al. APOBEC3G Genetic Variants and Their Influence on the Progression to AIDS. J Virol. 2004;78(20):11070–6.

15. Refsland EW, Hultquist JF, Luengas EM, Ikeda T, Shaban NM, Law EK, et al. Natural Polymorphisms in Human APOBEC3H and HIV-1 Vif Combine in Primary T Lymphocytes to Affect Viral G-to-A Mutation Levels and Infectivity. PLoS Genet. 2014;10(11).

16. Ooms M, Majdak S, Seibert CW, Harari A, Simon V. The Localization of APOBEC3H Variants in HIV-1 Virions Determines Their Antiviral Activity. J Virol. 2010;84(16):7961–9.

17. Cavalli-Sforza LL. Genes, People and Languages. Sci Am. 1991;265(5):104–10. 18. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, et al. The

Distribution of Human Genetic Diversity: A Comparison of Mitochondrial, Autosomal, and Y-Chromosome Data. Am J Hum Genet. 2000;66(3):979–88.

19. Tishkoff SA, Williams SM. Genetic analysis of African populations: human evolution and complex disease. Nat Rev Genet. 2002;3(8):611–21.

20. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008;451(7181):998–1003.

21. Conrad DF, Hurles ME. The population genetics of structural variation. Nat Genet. 2007;39(7S):S30–6.

22. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Vol. 3, Nature Reviews Genetics. 2002. p. 370–9.

23. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, et al. Genetic structure of human populations. Science. 2002;298(5602):2381–5.

24. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science (80-). 2008;319(5866):1100–4.

25. Altshuler D, Lander E, Ambrogio L. A map of human genome variation from population scale sequencing. Nature. 2010;476(7319):1061–73.

26. Lane AB, Soodyall H, Arndt S, Ratshikhopha ME, Jonker E, Freeman C, et al. Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. Am J Phys Anthropol. 2002;119(2):175–85.

27. Mitchell P. Genetics and southern African prehistory: An archaeological view. Vol. 88, Journal of Anthropological Sciences. 2010. p. 73–92.

28. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171–81.

29. Andrews S. FastQC: A quality control tool for high throughput sequence data. Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. 2010. p. http://www.bioinformatics.babraham.ac.uk/projects/.

30. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.

31. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available: http://arxiv.org/abs/13033997. 2013;

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

33. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015;31(13):2202–4.

34. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.

35. Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. PLoS Comput Biol. 2013;9(7).

36. R Development Core Team. R: A Language and Environment for Statistical Computing. Vol. 0, R Foundation for Statistical Computing Vienna Austria. 2010. p. {ISBN} 3-900051-07-0.

37. Court MH MH. Court's (2005–2008) online calculator. Tuft University Web site. 2012.

38. Machiela MJ, Chanock SJ. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555–7.

39. Reddy K, Winkler CA, Werner L, Mlisana K, Abdool Karim SS, Ndung'U T. Apobec3g expression is dysregulated in primary hiv-1 infection and polymorphic variants influence cd4+ t-cell counts and plasma viral load. AIDS. 2010;24(2):195–204.

40. Feng Y, Chelico L. Intensity of deoxycytidine deamination of HIV-1 proviral DNA by the retroviral restriction factor APOBEC3G is mediated by the noncatalytic domain. J Biol Chem. 2011;286(13):11415–26.

41. Mhandire K, Duri K, Mhandire D, , Musarurwa C, Stray-Pedersen B, Dandara C. Evaluating the contribution of APOBEC3G haplotypes on influencing HIV infection in a Zimbabwean paediatric population [Internet]. Vol. 106, South African Medical Journal. 2016. p. S119–23.

42. Compaore TR, Soubeiga ST, Ouattara AK, Obiri-Yeboah D, Tchelougou D, Maiga M, et al. APOBEC3G variants and protection against HIV-1 infection in Burkina Faso. PLoS One. 2016;11(1).

43. Mulder LCF, Ooms M, Majdak S, Smedresman J, Linscheid C, Harari A, et al. Moderate influence of human APOBEC3F on HIV-1 replication in primary lymphocytes. J Virol. 2010;84(18):9613–7.

44. Duggal NK, Malik HS, Emerman M. The Breadth of Antiviral Activity of Apobec3DE in Chimpanzees Has Been Driven by Positive Selection. J Virol . 2011;85(21):11361–71.

45. Tan L, Sarkis PTN, Wang T, Tian C, Yu X-F. Sole copy of Z2-type human cytidine deaminase APOBEC3H has inhibitory activity against retrotransposons and HIV-1.

FASEB J. 2009;23(1):279–87.

46. Li MMH, Wu LI, Emerman M. The range of human APOBEC3H sensitivity to lentiviral Vif proteins. J Virol. 2010;84(1):88–95.

47. Zhen A, Wang T, Zhao K, Xiong Y, Yu X-F. A single amino acid difference in human APOBEC3H variants determines HIV-1 Vif sensitivity. J Virol. 2010;84(4):1902–11.

48. Zhen A, Du J, Zhou X, Xiong Y, Yu XF. Reduced APOBEC3H variant anti-viral activities are associated with altered RNA binding activities. PLoS One. 2012;7(7).

49. Jin, X., Brooks, A., Chen, H., Bennett, R., Reichman R& SH. APOBEC3G/CEM15 (hA3G) mRNA levels associate inversely with human immunodeficiency virus viremia. J viriology. 2005;79:11513–11516.

Accession numbers of all individual APOBEC3 sequences submitted to NCBI Sequence Read Archive (project number: PRJNA429751).