

Characterising the Prevalence and Mode of CXCR4 Usage in HIV-1 Group M Subtype C



Saleema Crous

South African National Bioinformatics Institute
University of the Western Cape

Supervisor: Prof. Simon A. Travers

Co-supervisor: Prof. Alan Christoffels

A thesis submitted in fulfillment of the requirements for the award of degree of Magister Scientiae (M.Sc.) in Bioinformatics at the South African National Bioinformatics Institute (SANBI), University of the Western Cape

November 2013

For my Mother, for all you've taught me.
To Montaser, for encouraging me and believing in me.



UNIVERSITY *of the*
WESTERN CAPE

Acknowledgements

My sincere gratitude goes to my supervisor. Simon, I can't thank you well enough for all you have done.

I would like to thank Conor Meehan for providing the scripts for implementing the 11/25 and 11/24/25 charge rules, Ram Krishna Shrestha for assistance with Python scripting as well as Alexander Thielen for running the larger dataset through geno2pheno in batch mode.

I would especially like to thank my family. Without your love, support and patience, this achievement would not be possible.

This work was supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa (grant # 64751).

Abstract

Determination of CXCR4-usage patterns is essential in establishing suitability of CCR5 antagonist prescription in HIV-1 infected individuals to prevent treatment failure. Previous studies have suggested a switch to CXCR4-usage to be far less common in subtype C, yet recent studies have reported between 30 - 50% CXCR4-usage in this subtype. However, CXCR4-usage in subtype C is poorly characterised. Furthermore, the reliability of available genotypic algorithms is unknown for subtype C sequences.

In this study, a comparative analysis of the predictive ability of several subtype B-modeled genotyping algorithms in subtype C tropism determination was undertaken. A total of 731 HIV-1 subtype C V3 sequences with phenotypically determined coreceptor tropism were collated from several sources. Datasets of 349 CCR5, 25 CXCR4 exclusive and 31 R5X4 (Dual) sequences were submitted to 11 various tropism prediction tools. The best performing tool was used to determine the tropism of 12,121 subtype C V3 sequences with unknown phenotypes, in order to characterise the prevalence and method of CXCR4 usage in HIV-1 subtype C.

We determined that geno2pheno with a false positive rate of 5% is the best approach for predicting CXCR4-usage in subtype C sequences with an accuracy of 94% (89% sensitivity and 99% specificity). Contrary to what has been reported for subtype B, the optimal approaches for prediction of CXCR4-usage in sequence from viruses that use CXCR4 exclusively, also perform best at predicting CXCR4-use in dual-tropic viral variants. Furthermore, we find that a switch to CXCR4 usage is seen in subtype C for well over 20 years and has occurred consistently over time. At 5%, the frequency of CXCR4-usage in subtype C database records is lower than previous reports for both subtype C and B.

The Geno2pheno coreceptor tool may be used as a reliable genotypic predictor in clinical settings to establish the viability of CCR5-antagonist therapies using drugs such as Maraviroc and provides a rapid and cost effective alternative to phenotypic testing in resource limited areas. A

switch to CXCR4-usage in subtype C is constant but lower when compared to subtype B, a finding which may have broad implications for the design of intervention and treatment strategies for HIV-1 subtype C.



Declaration

I declare that *Characterising the Prevalence and Mode of CXCR4 Usage in HIV-1 Group M Subtype C* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.



Full name

UNIVERSITY of the
WESTERN CAPE

Date

Signed

Contents

1	Introduction	1
1.1	General Background	1
1.2	Origins of HIV	4
1.3	Classification	5
1.4	HIV Replication Cycle	7
1.5	Viral Genome Organisation	10
1.6	Structure of gp120	13
1.7	HIV-1 Cell Entry	14
1.7.1	CD4 Binding	14
1.7.2	Coreceptor Usage	14
1.7.3	Cell Membrane Fusion	15
1.8	HIV-1 Tropism Nomenclature	16
1.9	Tropism Determination	17
1.9.1	Phenotypic methods	17
1.9.2	Genotypic methods	18
1.10	CXCR4-usage in HIV-1 subtype C	19
1.11	Thesis Rationale	22
1.12	Thesis Outline	23
2	Methodology	25
2.1	Dataset Curation: Appraising the Performance of Coreceptor Genotyping Tools at Accurately Predicting Coreceptor Usage in HIV-1 Group M Subtype C.	25
2.2	Handling of Ambiguous Nucleotide Bases	26
2.3	Genotypic Algorithm Evaluation	27
2.3.1	Web PSSM matrices	28
2.3.2	Geno2pheno	28
2.3.3	Wetcat package	29
2.3.4	Charge rules	29
2.3.5	Raymond method	30
2.4	Determining Sensitivity and Specificity of Genotypic Algorithms	30

CONTENTS

2.5	Determining Accuracy of Genotypic Algorithms	32
2.6	Determining Effect of Dual Tropic Viruses on Prediction of CXCR4-usage	32
2.7	Dataset Curation: Prevalence of CXCR4-usage in Subtype C	33
2.8	Multiple Sequence Alignment Using RAMICS	34
2.9	Coreceptor Tropism Prediction	34
2.10	Exploring Prevalence and Patterns of CXCR4-usage in Subtype C	35
2.11	Maximum Likelihood Phylogeny Estimation	36
3	Results	37
3.1	Dataset Compilation: Appraising the Performance of Coreceptor Genotyping Tools at Accurately Predicting Coreceptor Usage in HIV-1 Group M Subtype C.	37
3.2	Handling of Ambiguous Nucleotide Bases	38
3.3	Sensitivity and Specificity of Genotypic Algorithms	39
3.4	Genotypic Algorithm Evaluation	41
3.5	Accuracy of Genotypic Algorithms	42
3.6	Effect of Dual Tropic Viruses on Prediction of CXCR4-usage	43
3.7	Dataset Compilation: Prevalence and Mode of CXCR4-usage in Subtype C	45
3.8	Multiple Sequence Alignment Using RAMICS	45
3.9	CXCR4-usage Patterns in HIV-1 Subtype C Sequences	46
3.10	Prevalence of CXCR4-usage in HIV-1 Subtype C sequences	50
3.11	Maximum Likelihood Phylogeny Estimation	53
4	Discussion	56
4.1	Ability to Account for Ambiguous Nucleotide Positions	56
4.2	Effect of Dual Tropic Viruses on Prediction of CXCR4-usage	57
4.3	Genotypic Algorithm Evaluation	58
4.4	Multiple Sequence Alignment Using RAMICS	61
4.5	Prevalence and Patterns of CXCR4-usage in HIV-1 Subtype C	62
4.6	Study Limitations	66
5	Conclusions	69
6	Appendix A	72
7	Appendix B	75
8	Appendix C	78

List of Figures

1.1	Global occurrence of the Human Immunodeficiency Virus (2010)	2
1.2	Phylogenetic clustering of HIV-1 groups M, N, O and P	6
1.3	Life cycle of the Human Immunodeficiency Virus	9
1.4	HIV-1 genome map	11
3.1	Performance of genotyping algorithms in predicting CXCR4-usage. . .	40
3.2	Ability of each approach at predicting CXCR4-usage in dual-tropic viral sequences.	44
3.3	Prevalence of CXCR4-usage over time.	47
3.4	Prevalence of CXCR4-usage over time for each of the countries with the highest number of subtype C sequences.	48
3.5	Number of subtype C sequences over time	51
3.6	Proportion of CXCR4-using sequences from countries with more than 200 subtype C sequences each.	53
3.7	Distribution of CCR5 and CXCR4-using sequences, indicating bootstrap values greater than or equal to 70	54
4.1	Comparison of Muscle and RAMICS multiple sequence alignments . .	63

List of Tables

3.1	Number of CCR5, CXCR4 and dual tropic sequences obtained from each source.	38
3.2	Performance of genotyping approaches at predicting CXCR4-usage in HIV-1 subtype C viral sequences.	41
3.3	Accuracy of genotyping approaches at correctly predicting coreceptor tropism.	43
3.4	Number of predicted R5 and X4 sequences for countries with more than 200 subtype C sequences each.	52
6.1	Uncorrected numbers of TP, TN, FP, FN predicted by each of the approaches.	73
7.1	Number of HIV-1 subtype C sequences recorded per country	76

Chapter 1

Introduction

1.1 General Background

No single factor has had a more devastating impact on modern society than the spread of the Human Immunodeficiency Virus (HIV). The number of HIV infections has reached epidemic proportions, with an estimated 34 million people living with HIV globally by the end of 2011 (UNAIDS, 2013). It is estimated that 2.5 million people were newly infected with this virus in 2011 (UNAIDS, 2013). HIV is the etiological agent responsible for the development of Acquired Immune Deficiency Syndrome (AIDS), which has been attributed to over 25 million deaths globally since it was first discovered 30 years ago (UNAIDS, 2013). In 2011 alone, it was estimated that 1.7 million adults and children died from this disease (UNAIDS, 2013).

The highest prevalence of HIV infections are in developing countries, with the highest incidences occurring in Africa, Central Asia, Eastern Europe and the Caribbean (Figure 1.1) (UNAIDS, 2013). However, the effects of this virus are most severe in Sub-Saharan Africa, where 69% of all people living with HIV reside, and the number of people living with HIV in the region was estimated to be 23.5 million (UNAIDS,

1. INTRODUCTION

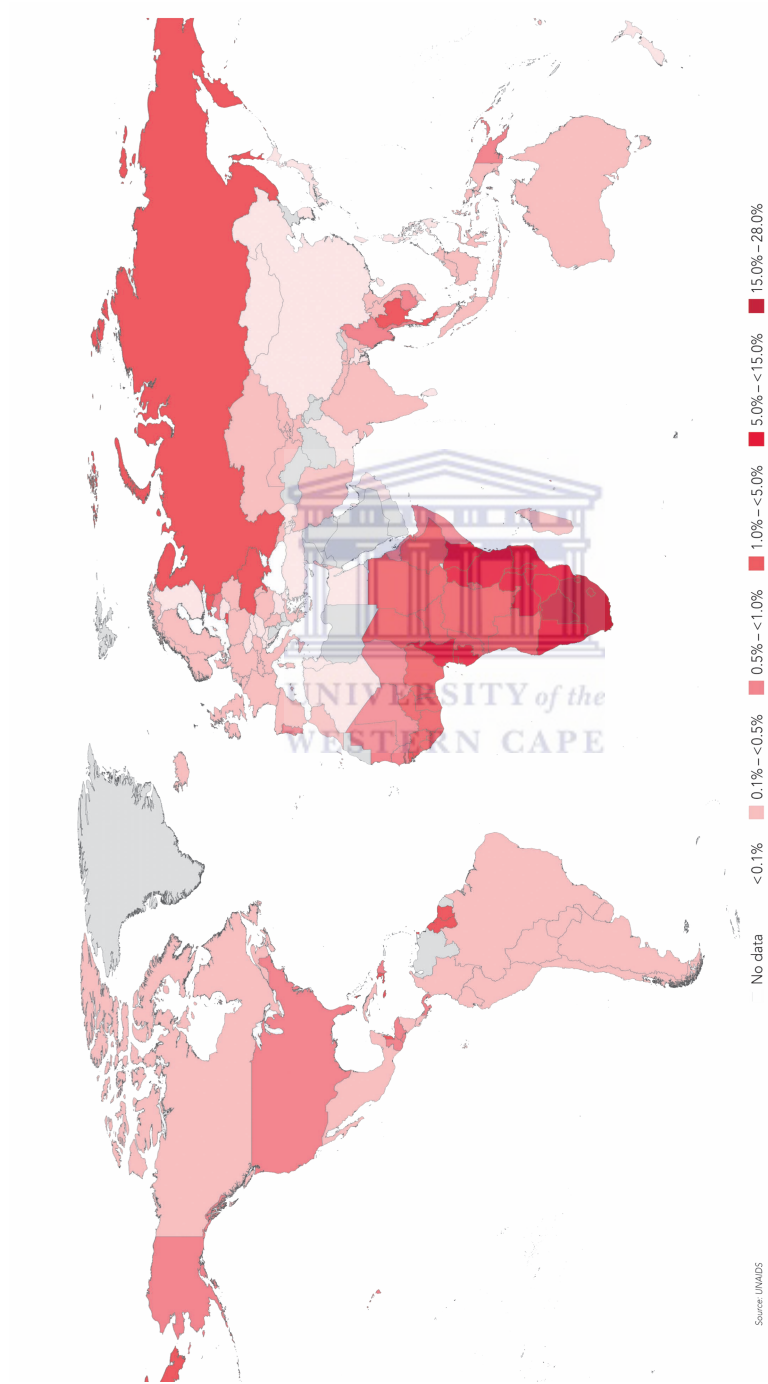


Figure 1.1: Global prevalence of the Human Immunodeficiency Virus (2010), with the highest prevalence observed in Africa and Central Asia. Source: http://unaids.org/documents/20101123_2010_hiv_prevalence_map_em.pdf

1. INTRODUCTION

2013). Of the countries in this region, South Africa has an estimated 5.6 million people living with HIV - more than any other country in the world. In spite of these alarming figures and the dire need to curb the spread of this deadly virus, our knowledge of basic HIV etiology, such as mode of coreceptor usage, is incomplete, limiting our ability to effectively design suitable treatments. As a result, the death toll associated with this disease continues to rise globally, particularly in Africa, where AIDS is the most common cause of death (UNAIDS, 2013).

The first documented reports of AIDS occurred in 1981, when young, previously healthy homosexual males in the United States presented with a series of rare diseases and opportunistic infections (CDC, 1981, 1982a,b, 2006). The interconnected nature of these cases soon became evident, and in 1982 the term AIDS was coined by the Centre for Disease Control and Prevention (CDC) to describe this disease which was spread primarily through sexual contact, blood products and through breast feeding.

Although little was known about the etiology of the disease, it was characterized by a decrease in the CD4+ helper/inducer subset of T lymphocyte cells, leading to a suppressed immune response, and which in turn resulted in an array of opportunistic infections (CDC, 1982a). In 1983, a link was established between the AIDS seen in macaque monkeys and that seen in humans, and it was suggested that an infectious agent such as a virus might be the cause of the disease in humans as it was in the primates (Hunt et al., 1983). Later in that same year, the retrovirus termed Human Immunodeficiency Virus (HIV) was established to be the definitive etiologic agent of AIDS (Barré-Sinoussi et al., 1983; Gallo et al., 1983).

1. INTRODUCTION

1.2 Origins of HIV

Human infection by HIV is a consequence of cross species transmission of simian immunodeficiency virus (SIV) from non-human primates (Robertson et al., 1995; Sharp and Hahn, 2011; Hahn et al., 2000; Lemey et al., 2003). Based on the SIV strain that it is derived from, HIV has been classified into two groups: HIV-1 and HIV-2, each of which have arisen through independent zoonotic events (Gao et al., 1999; Tebit and Arts, 2011; Wertheim and Worobey, 2009).

Human Immunodeficiency Virus type 2 (HIV-2) is believed to have arisen in Guinea Bissau in the early 1930s (Tebit and Arts, 2011; Lemey et al., 2003), through a sooty mangabey transmission event (SIVsmm). This strain of HIV, which is less easily transmitted is now primarily restricted to western Africa (Lemey et al., 2003), and is responsible for roughly 1% of all HIV infections (Lemey et al., 2003; Marlink et al., 1994).

By contrast, Human Immunodeficiency Virus type 1 (HIV-1) infections have become pandemic, with 99% of the global HIV infections caused by this viral strain (Lemey et al., 2003; Marlink et al., 1994). Studies indicate that HIV-1 arose in the West Central African forests of Cameroon and the Democratic Republic of Congo in the early 1900s (Tebit and Arts, 2011; Santiago et al., 2002; Gao et al., 1999; Keele et al., 2006). The origin of this highly virulent strain of HIV has been traced back to SIVcpz infected chimpanzees (*Pan troglodytes troglodytes*) (Gao et al., 1999).

1. INTRODUCTION

1.3 Classification

The Human Immunodeficiency Virus is characterized by extensive genetic heterogeneity, driven by several factors, including recombination events during replication (Temin, 1993), lack of proof-reading ability of the reverse transcriptase (RT) (Roberts et al., 1988; Preston et al., 1988; Temin, 1993), the high replication rate of HIV-1 *in vivo* (Ho et al., 1995), and host selective immune pressures (Michael, 1999; Preston and Dougherty, 1996; Preston et al., 1988; Ho et al., 1995; Wei et al., 1995; Roberts et al., 1988). As a result of this variability, HIV-1 viruses have been divided into several distinct genetic groups: group M (main), group O (outlier), and group N (non-M/ non-O). Group P (Pending identification of further human cases) was recently described and subsequently confirmed as the fourth group, having being found in two unrelated patients of Cameroonian origin (Plantier et al., 2009; Vallari et al., 2011). Phylogenetic evidence suggests that this HIV-1 group is most closely related to a strain of SIV derived from gorillas (SIVgor) than it is to any of the other groups (Figure 1.2) (Plantier et al., 2009; Vallari et al., 2011).

Of these, Group M contributes to the greatest phylogenetic diversity among all groups and is responsible for over 95% of global HIV isolates (Vidal et al., 2000; Hahn et al., 2000; Ward et al., 2013). As determined by the analysis of complete viral genomes, group M is divided into nine distinct subtypes (A-D, F-H, J and K), with each phylogenetically associated clade denoting a distinct lineage of HIV and having a different geographical distribution (Robertson et al., 2000). Within a single subtype, viral isolates may display nucleotide distances of up to 35% (Subbarao and Schochetman, 1996). Furthermore, depending on the gene analysed genetic variation between

1. INTRODUCTION

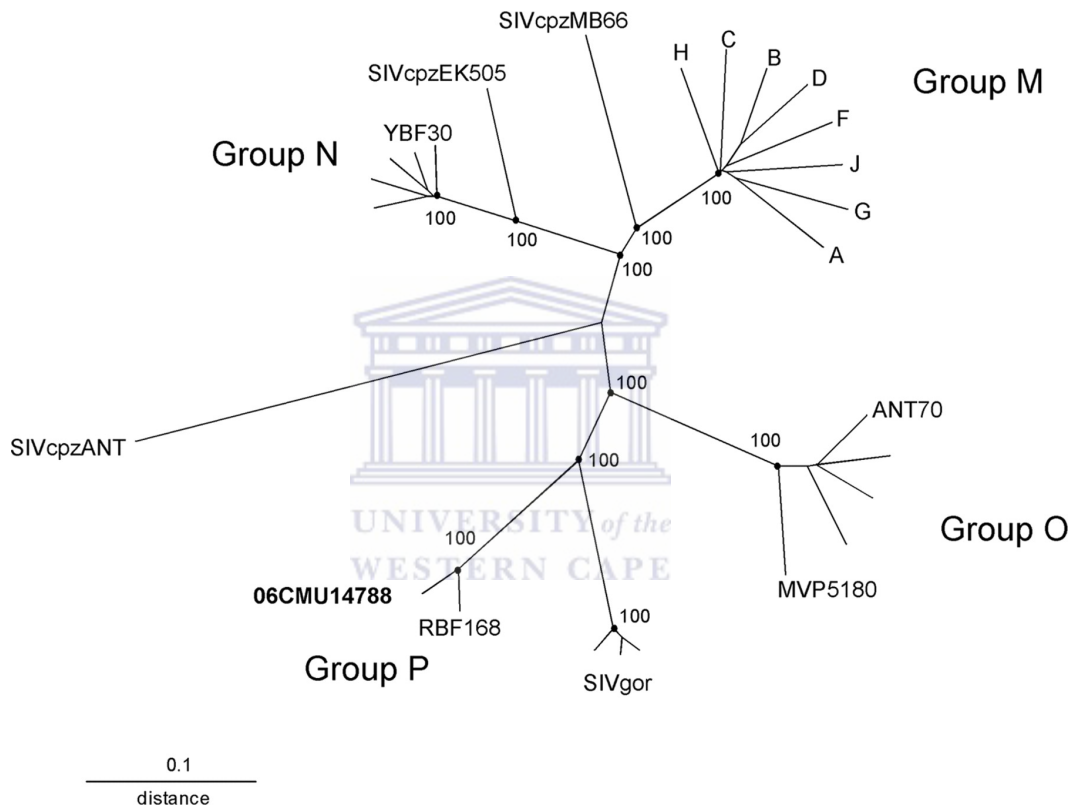


Figure 1.2: Phylogenetic clustering of HIV-1 groups M, N, O and P derived from nucleotide alignment of genome sequences, showing group P clustering closely together on a branch shared with the SIVgor strains. Source: <http://jvi.asm.org/content/85/3/1403/F2.expansion.html>

1. INTRODUCTION

subtypes can range from 30 - 40% (Rotta and Almeida, 2011; Moore et al., 2001). Genetic recombination between multiple subtypes of HIV circulating within an individual has given rise to a further 58 unique circulating recombinant forms (CRF) of the virus within Group M (Han et al., 2013).

Subtype C is the most prevalent subtype, accounting for almost 60% of all HIV cases, and representing 22 million infections globally (UNAIDS, 2013). It is the most rapidly spreading form of HIV-1, with the highest number of infections occurring in developing countries such as those in East and Southern Africa, as well as India, Nepal and parts of China (Goudsmit, 1997). Despite the major role this subtype plays in shaping the epidemic, the vast majority of HIV research is centered on HIV-1 subtype B, which is primarily found in Europe, the Americas and Oceania where the majority of HIV research is financed and conducted (Pagán and Holguín, 2013). However, worldwide subtype B accounts for approximately 12% of all HIV infections (Goudsmit, 1997).

1.4 HIV Replication Cycle

HIV-1 infects cells of the CD4+ T-cell and macrophage lineages and takes between 48 and 72 hours to complete a single replication cycle (McDougal et al., 1986; Mohammadi et al., 2013). The HIV replication cycle can be divided into an early and a late phase, based on the order of events taking place (Freed, 2001).

The early phase of the HIV life cycle begins with the binding of the viral docking glycoprotein gp120 to the primary host cell receptor, CD4, triggering a series of

1. INTRODUCTION

conformational changes. These structural changes allow chemokine receptor binding to take place and subsequently, fusion of the viral and cellular membranes (Figure 1.3 Steps 1-3) (Sattentau and Moore, 1991; Bergeron et al., 1992). This entry process is described in more detail in section 1.7 of this chapter.

Once membrane fusion has taken place, the HIV core containing the viral genome is released into the cytoplasm of the target cell, along with viral proteins (Freed, 2001; Weber, 2001; Coiras et al., 2009). The two single stranded viral RNA molecules are then retrotranscribed into linear double stranded proviral cDNA by the viral enzyme reverse transcriptase, and is subsequently known as the reverse transcription complex (RTC) (Figure 1.3 Step 4) (Freed, 2001; Weber, 2001; Coiras et al., 2009). Prior to nuclear integration, uncoating of the capsid takes place, followed by the formation of the RTC, (McDougal et al., 1986; Temin, 1993). A large nucleoprotein complex or pre-integration complex (PIC) consisting of viral and cellular proteins surrounding the viral genetic material is assembled (Miller et al., 1997; Coiras et al., 2009). The PIC is translocated from the host cell periphery towards the nucleus via microtubules and actin filaments, where nuclear import takes place through a nuclear pore complex (Miller et al., 1997; Ross and Cullen, 1998). The viral DNA is then integrated into the host cell genome by integrase (IN), and transfer of the modified provirus DNA into the host genome takes place (Figure 1.3 Step 5), completing the early phase of the replication cycle (Chiu and Davies, 2004).

Although the host cellular machinery is used by the HIV proviral insert for replication, the synthesis of viral RNA and proteins is strictly controlled by viral regulatory proteins in a process known as transcription (Wu and March, 2003) (Figure 1.3 Step

1. INTRODUCTION

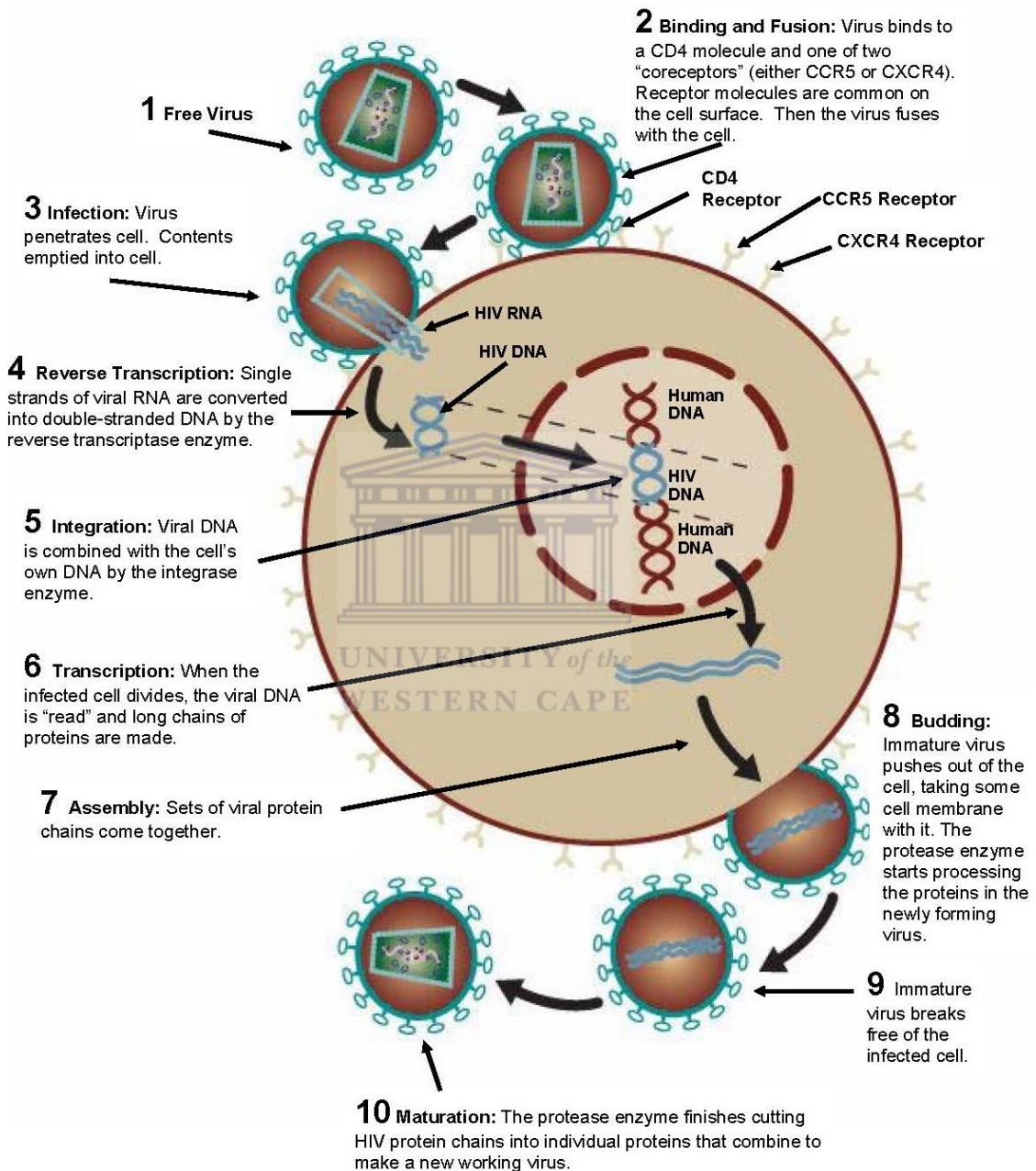


Figure 1.3: Life cycle of the Human Immunodeficiency Virus. Source: <http://stanford.edu/~rabriggs/hiv/lifecycle.jpg>

1. INTRODUCTION

6). This Viral transcriptome is able to encode for all necessary viral proteins, with the early viral proteins, Tat and Rev, regulating the expression of the late viral proteins, the structural and accessory proteins (Wu and March, 2003). Newly produced viral proteins and the RNA genome assemble in the cytoplasm at the cell membrane (Figure 1.3 Step 7), where processed Env is expressed and new virus particles will form (Briggs et al., 2009). The final step of the virus life cycle includes budding from infected cells (Figure 1.3 Step 8 and 9), followed by viral protease processing of Gag and Gag-Pol precursors to form mature infectious particles (Briggs et al., 2009) (Figure 1.3 Step 10).

1.5 Viral Genome Organisation

HIV and SIV are members of the *Lentivirus* genus of the Retroviridae family. Retroviridae are enveloped viruses that are composed of two identical positive-sense RNA strands and are characterised by the ability to reverse transcribe RNA into DNA during viral replication (Freed, 2001). The HIV-1 genome is approximately 9.8 kilobases and, when integrated as a double stranded DNA (provirus), is flanked by long terminal repeats (LTR) generated during reverse transcription (Freed, 2001).

The HIV genome consists of several unique genes encoding regulatory proteins. These include *Tat* and *Rev*, as well as the accessory proteins *Nef*, *Vif*, *Vpr*, and *Vpu* (Figure 1.4). These genes are important regulators of viral replication, transcription and assembly (Wei et al., 1995; Weber, 2001; Doehle et al.). In addition to gene products, the retroviral genome also contains structural regulatory motifs, such as the trans-active response element (TAR) and *Rev* response element (RRE), which are required

1. INTRODUCTION

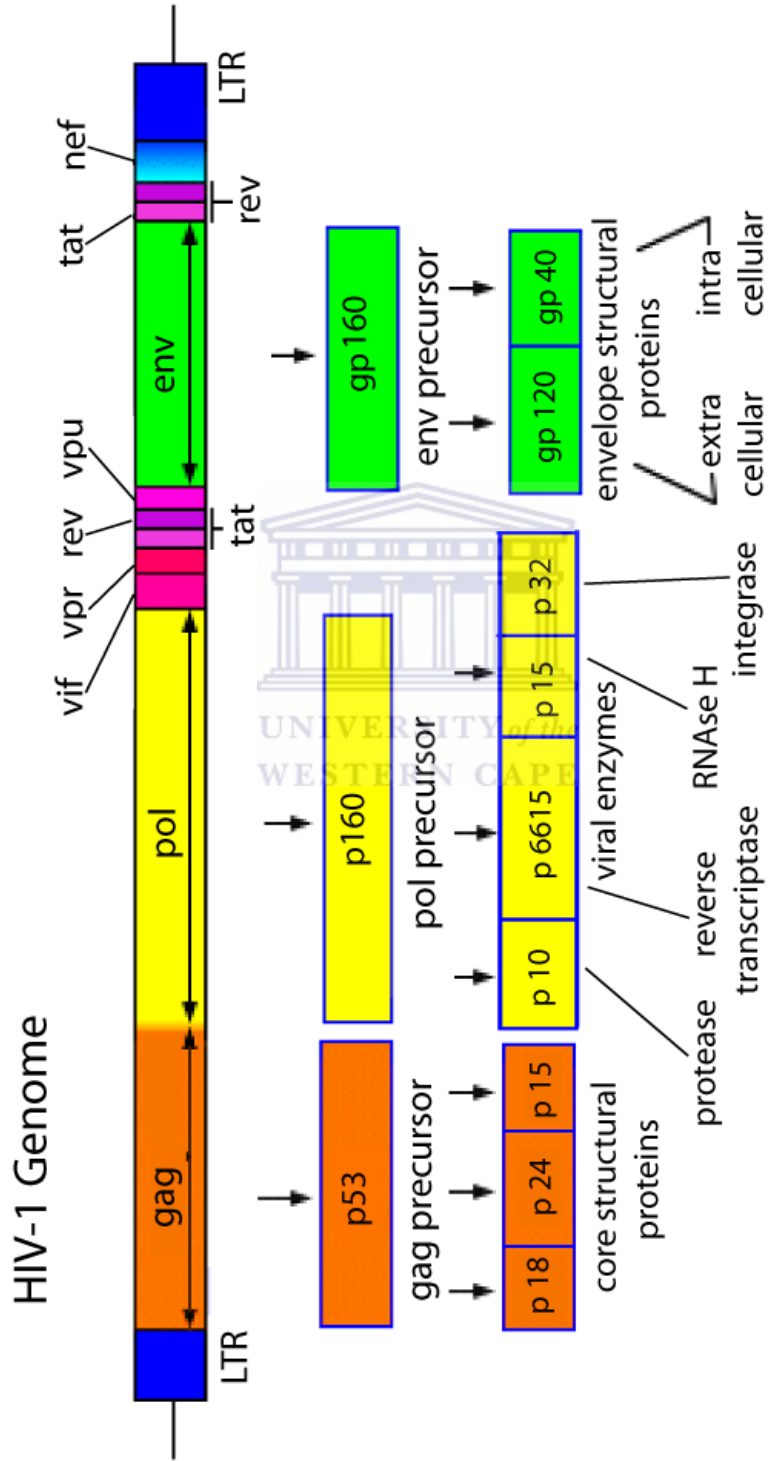


Figure 1.4: HIV-1 and associated genes. Adapted from: <http://yale.edu/bio243/HIV/genome.html>.

1. INTRODUCTION

for transcription and export of non-spliced and partially spliced viral mRNAs to the cytoplasm (Wei et al., 1995; Weber, 2001).

The genomic RNA contains three open reading frames and encodes the group-specific antigen (*gag*), polymerase (*pol*), and envelope (*env*) genes (Figure 1.4). These three primary structural gene products are initially synthesized as polyprotein precursors, which are then processed by viral or cellular proteases into viral proteins (Wei et al., 1995; Weber, 2001). The *gag* gene encodes the Gag polyprotein precursor p55 that is cleaved by the viral protease into matrix (MA), capsid (CA), nucleocapsid (NC) and Vpr-binding proteins (McDougal et al., 1986). The viral enzymes protease (PR), integrase (IN), RNase H and reverse transcriptase (RT) are encoded by the *pol* gene and are initially produced as a Gag-Pol precursor and processed by the viral protease (McDougal et al., 1986). The *env* gene encodes the polyprotein precursor gp160, which is produced intracellularly through enzymatic addition of complex carbohydrates to the synthesised protein (Ross and Cullen, 1998) (Figure 1.4). The gp160 molecule is proteolytically cleaved by host cellular enzymes, forming the gp120 and gp41 proteins (Chan et al., 1997). The gp120 protein lies on the external surface of the viral particle, while the gp41 protein is located internally and is attached to it across the membrane, together forming a non-covalent transmembrane complex. Fusion of the cellular and viral membranes is mediated by the gp41 component of the transmembrane complex, resulting in the formation of a pore which allows virion genetic material to pass into the cell.

1. INTRODUCTION

1.6 Structure of gp120

Three gp120s bound as heterodimers to a transmembrane glycoprotein gp41, form a trimer complex on the surface of a HIV virion - called the envelope spike (Liu et al., 2008). To enable cell entry by HIV, the gp120 glycoprotein, must first be recognized by, and bind to a CD4 receptor on the target host cell (Dalglish et al., 1984; Maddon et al., 1986; McDougal et al., 1986). This binding induces a conformational change in the gp120/gp41 trimer complex (Sattentau and Moore, 1991; Liu et al., 2008) thereby enabling binding of a chemokine receptor. *In vivo*, these coreceptors may be either CCR5 or CXCR4, and in some instances, both coreceptors can be used for cell entry (Dragic et al., 1996).

Gp120 amino acid sequences consist of five relatively conserved regions (C1 – C5) alternating with five variable regions (V1 – V5) (Starcich et al., 1986; Chan et al., 1997). The conserved regions form the protein core, which has an inner and an outer domain formed by secondary folding structures, while the variable regions are known to form loops which are anchored to the core via cysteine disulphide bonds (Kwong et al., 1998; Liu et al., 2008). Attached to the outer domain of the gp120 core as well as within the loops, are a series of complex, host derived sugars, forming a glycan shield around the protein (Quiñones-Kochs et al., 2002; Polzer et al., 2002). Apart from maintaining *env* structure, the heavily glycosylated nature of gp120 is important in the occlusion of neutralising antibody epitopes and receptor binding sites (Kwong et al., 1998; Wyatt et al., 1998; Quiñones-Kochs et al., 2002; Pollakis et al., 2001).

1. INTRODUCTION

1.7 HIV-1 Cell Entry

The entry of an HIV particle into the target cell involves an intricate series of sequential interactions, the primary stages of which are: (i) attachment of the viral gp120 to the CD4 receptor, (ii) binding of the gp120 to either of the co-receptors CCR5 or CXCR4 and (iii) fusion of the viral and cellular membranes.

1.7.1 CD4 Binding

The CD4 glycoprotein, expressed on the surface of macrophages, T-cells, monocytes, and dendritic cells, is the primary receptor involved in HIV infection and host cell entry (Chan et al., 1997). Cellular entry is brought about by strong electrostatic and molecular interactions between several gp120s and several CD4 proteins. This binding induces a conformational change in the gp120/gp41 trimer complex (Sattentau and Moore, 1991; Liu et al., 2008), subsequently leading to the exposure of the coreceptor binding sites. Thereafter, a gp41 fusion peptide is inserted into the host cell membrane (Buzon et al., 2010).

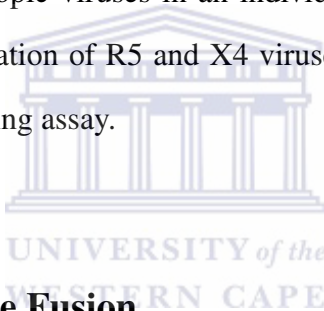
1.7.2 Coreceptor Usage

CCR5-tropic viruses are associated with primary transmission and can persist throughout infection (Dragic et al., 1996; Alkhatib, 2009). In as many as 50% of HIV-1 subtype B infections, a switch to CXCR4-usage has been observed and this switch is generally regarded as an indicator of disease progression (Koot et al., 1993; Hazenberg et al., 2003; Levine and Sodora, 2006). Early studies of HIV-1 subtype C suggested that a switch to CXCR4-usage was less common in subtype C compared to subtype B (Abebe et al., 1999; Pollakis et al., 2004), however more recent studies have suggested

1. INTRODUCTION

that between 30-50% of subtype C infected individuals exhibit a change to CXCR4-usage during disease progression (Connell et al., 2008; Kassaye et al., 2009; Michler et al., 2008; Cilliers et al., 2003; Papathanasopoulos et al., 2002; Johnston et al., 2003).

Dual-tropic viruses (R5X4) capable of using either CCR5 or CXCR4 for host cell entry have been described (Berger et al., 1998; Alkhatib, 2009) as have dual-tropic viruses that, while capable of using either receptor for cell entry, exhibit preferential use of either CCR5 (dual-R) or CXCR4 (dual-X) (Huang et al., 2007, 2009). Detecting the presence of dual-tropic viruses in an individuals viral population is difficult however, as a mixed population of R5 and X4 viruses will be identified as dual in a population-based phenotyping assay.



1.7.3 Cell Membrane Fusion

Once coreceptor binding has taken place, a second conformational shift occurs in the gp120/gp41 trimer complex, enabling the fusion of viral and cellular membranes (Sattentau and Moore, 1991; Bergeron et al., 1992). The viral and cellular membranes are brought near each other following the formation of a six-helix bundle of the gp41 ectodomain (Weissenhorn et al., 1997; Chan et al., 1997; Chan and Kim, 1998). An intermediate bridging state is created when the gp41 fusion peptide is exposed to the cell membrane, enabling for fusion to occur and the subsequent transfer of viral genetic material to the host cell (Buzon et al., 2010).

1. INTRODUCTION

1.8 HIV-1 Tropism Nomenclature

Although several chemokine receptors are found to enable cell entry in a laboratory setting, *in vivo*, HIV-1 viruses are capable of using only the CCR5 or CXCR4 receptors (Doms and Trono, 2000). The following nomenclature, based on potential viral coreceptor usage patterns, is currently used to describe viral tropism (Berger et al., 1998, 1999):

- **CCR5-tropic:** Viruses or virus populations that are only able to use the CCR5 chemokine co-receptor to infect CD4+ cells. They are typically referred to as R5 viruses.
- **CXCR4-tropic:** Viruses or virus populations that are only able to use the CXCR4 chemokine co-receptor to infect CD4+ cells. They are typically referred to as X4 viruses.
- **Dual (D)-tropic:** Viruses or virus populations that are able to use either the CCR5 or CXCR4 co-receptors to infect CD4+ cells.

The term CXCR4-using (as used throughout this thesis) is often used to broadly refer to all HIV viruses capable of using the CXCR4 coreceptor for cell entry, and includes both viral strains using the CXCR4 coreceptor exclusively as well as dual tropic viruses (Kiselyeva et al., 2007). The assays commonly used in a clinical setting are not capable of distinguishing between dual tropic viruses and mixed viral populations containing both R5- and X4-using strains (Huang et al., 2007, 2009).

1. INTRODUCTION

1.9 Tropism Determination

Determining the coreceptor usage profile of an individual's viral population has been used as an indicator of disease progression (Connor et al., 1997; Verhofstede et al., 2012; Zhang et al., 1998). In more recent years it has been used as an approach for detecting resistance to CCR5 antagonists (Cilliers et al., 2003; Westby and van der Ryst, 2005; Doranz et al., 1997). The approval of the first CCR5 antagonist, maraviroc (Selzenti®), has sparked the need for HIV coreceptor determination in the clinical setting (Poveda et al., 2006, 2009; Collins and iBase, 2007). Treatment with maraviroc essentially mimics the CCR5 $\Delta 32$ phenotype in that it blocks the CCR5 receptor, making it unavailable for binding (Baba et al., 1999). This 32bp deletion mutation in the CCR5 receptor – seen in about 4-15% of Europeans (Hummel et al., 2005) – confers immunity to homozygous carriers and a delay in disease progression to individuals with a heterozygous deletion, as HIV is unable to use this receptor to gain cell entry (Huang et al., 1996; Baba et al., 1999; Lederman et al., 2006). Since CCR5 antagonists are ineffective against CXCR4-using viral populations, it is essential for HIV-1 coreceptor to be determined before the onset of treatment.

1.9.1 Phenotypic methods

To date, phenotypic assays are the most effective means of elucidating the coreceptor tropism of a viral population (Fouchier et al., 1992). Monogram Biosciences Trofile™ assay (Whitcomb et al., 2007) which is based on recombinant virus technology, has been the most widely used diagnostic test, given that it was the only assay which provided tropism information in the maraviroc clinical trials (Poveda et al.,

1. INTRODUCTION

2010). Phenotypic approaches such as this however, are expensive, laborious, time consuming and unavailable for routine use in all laboratories, especially for use in developing countries (Prosperi et al., 2010; Sierra et al., 2007). Thus, bioinformatics approaches based on viral genotyping have been suggested to be a viable alternative for routine coreceptor tropism testing (McGovern et al., 2010).

1.9.2 Genotypic methods

Genotypic tropism testing is currently accepted as a standard for tropism determination according to the European, British and German/Austrian guidelines for tropism testing, and is routinely used in the clinical setting (Vandekerckhove et al., 2010, 2011a,b). While many amino acid positions throughout gp120 have been suggested to influence coreceptor affinity and tropism (Rizzuto and Sodroski, 2000; Rizzuto et al., 1998; Boyd et al., 1993; Bergeron et al., 1992; Ross and Cullen, 1998; Hoffman et al., 2002; Nabatov et al., 2004), the V3 loop appears to be the strongest determinant of coreceptor tropism with amino acid mutations affecting V3 net charge, charge at positions 11, 24 and 25 and glycan binding patterns all implicated in causing a switch from CCR5- to CXCR4-usage (Clevestig et al., 2006; Pollakis et al., 2001; Polzer et al., 2002; Fouchier et al., 1992; Cardozo et al., 2007; Resch et al., 2001).

Early genotypic algorithms predicted the coreceptor tropism of HIV-1 V3 sequences using the properties of the amino acids at positions 11 and 25 while later algorithms account for various properties of the entire V3 loop (Fouchier et al., 1992; Cardozo et al., 2007; Jensen et al., 2003, 2006; Sing et al., 2007a; Pillai et al., 2003; Cormier

1. INTRODUCTION

and Dragic, 2002; Poveda et al., 2010). With the exception of C-PSSM (Jensen et al., 2006) and the Raymond combined 11/25 and net charge rules (Raymond et al., 2010), all of these approaches have been optimised for coreceptor tropism prediction in subtype B and show varying levels of sensitivity at predicting CXCR4-usage in subtype B (Garrido et al., 2008; Poveda et al., 2010).

1.10 CXCR4-usage in HIV-1 subtype C

Despite HIV-1 subtype C accounting for almost 60% of worldwide HIV infections (Requejo, 2006), little is known about the prevalence and patterns of CXCR4-usage in this subtype. Earlier studies have suggested that a switch to CXCR4-usage was rare or never occurred in subtype C (Abebe et al., 1999; Cecilia et al., 2000). However, an increasing number of recent studies have shown that a switch from CCR5- to CXCR4-usage can and does occur (Cilliers et al., 2003; Pollakis et al., 2004). Cilliers and colleagues demonstrated that CCR5 and CXCR4 receptors are both used by HIV-1 subtype C (Cilliers et al., 2003). Pollakis and colleagues likewise reported that a switch from CCR5- to CXCR4-usage does occur in subtype C, and does so in a manner similar to that of subtype B (Pollakis et al., 2004). A recent study by Esbjörnsson and colleagues reported a frequency of 15% CXCR4-usage in subtype C sequences (Esbjörnsson et al., 2010). In addition, a study by Connell and colleagues reported that of the 20 South African AIDS patients they examined (19 of which had subtype C infections), 30% of primary isolates were CXCR4-using, indicating an increase in frequency of CXCR4-usage in subtype C over time (Connell et al., 2008). However, these studies were conducted on a relatively small number of subtype C sequences while the

1. INTRODUCTION

geographical range of sampling was limited.

Furthermore, the genetic determinants of the switch in coreceptor use are less-well understood than in subtype B. Conflicting reports have been published with some suggesting that these determinants are the same for subtype C as subtype B (Raymond et al., 2010), while others have presented evidence to the contrary (Jensen et al., 2006). Jensen and colleagues developed the only subtype C specific genotyping tool with a reported sensitivity of 75% (Jensen et al., 2006) while others evaluated the ability of this and other algorithms trained on subtype B data at correctly predicting CXCR4-use in subtype C sequence data (Raymond et al., 2010). They found that the most appropriate approach for predicting CXCR4-usage in subtype C were C-PSSM and their combined 11/25 and net charge rule (Raymond et al., 2010). When specificity was considered, however, Raymond and colleagues approach was significantly better than C-PSSM (96.4% versus 81.8%). The dataset used in this study, however, did not represent the entire spectrum of HIV-1 subtype C diversity in that it had a limited number of phenotyped sequences (55 R5 and 15 X4 sequences) collected from only two countries (Malawi and France).

Although maraviroc (Selzenti®) is in use worldwide, Food and Drug Administration (FDA) regulations have made tropism testing compulsory prior to its prescription. In resource limited settings, phenotypic methods are often too costly and require specialized facilities. Genotypic testing thus represents a viable alternative for use in a clinical setting. This holds particularly true for a country such as South Africa where the number of people receiving treatment is currently in excess of 2 million, and genotyping represents the only real alternative. Validation and improving the accu-

1. INTRODUCTION

racy of the available genotypic algorithms thus becomes imperative in the fight against HIV/AIDS. The relevance of this study is further increased by the recent approval of maraviroc/celsentri (Selzenti®) for use in South Africa in January 2014.



1. INTRODUCTION

1.11 Thesis Rationale

The overall objectives of this project were to:

1- Analyse the predictive ability of available genotypic algorithms in their prediction of coreceptor tropism in HIV-1 subtype C.

A large dataset consisting of all available HIV-1 subtype C envelope V3 loop sequences with phenotypically verified coreceptor tropism were used to evaluate the performance of currently available genotyping tools.

2- Evaluate the effect of the conflicting signal from dual-tropic viruses on genotypic algorithm evaluation.

CXCR4-using viruses were separated into CXCR4-exclusive and dual-tropic viruses, and the possible conflicting signal from the dual-tropic viruses on the ability of the genotypic approaches to correctly predict coreceptor phenotype was examined.

3- Determine prevalence and mode of CXCR4-usage in HIV-1 Subtype C.

A second larger dataset of subtype C V3 loop sequences with experimentally undetermined coreceptor tropism was genotypically characterized, and used to determine the prevalence of CXCR4-usage over time as well as to characterise the emergence of CXCR4-usage in HIV-1 subtype C.

1. INTRODUCTION

1.12 Thesis Outline

Chapter 1 *Introduction*: This chapter forms a general introduction to the Human Immunodeficiency Virus. It briefly describes aspects of the virus' biology related to coreceptor tropism. This chapter also provides an outline of the objectives of this study, as well as the rationale for undertaking it.

Chapter 2 *Methodology*: In this chapter, the various methodologies employed in this research, as well as the rationales behind them are explained. It details the way in which genotypic tools were evaluated as well as how the effect of potential conflicting signal from dual-tropic viruses was determined. It also deals with the methodologies used to determine the prevalence and mode of CXCR4-usage in HIV-1 Subtype C.

Chapter 3 *Results*: The results of the experiments carried out are described and analyzed in this chapter. The way in which genotypic algorithms handle ambiguous base pairs in sequences is presented. Genotypic algorithms are compared to determine which is most suitable for coreceptor tropism prediction in subtype C, in terms of sensitivity, specificity as well as overall accuracy of these tools. The ability of these algorithms to handle dual tropic viral sequences is uncovered. Finally, the prevalence of CXCR4-usage in the subtype C viral population over time is determined.

Chapter 4 *Discussion*: This findings of this study are discussed in the context of similar works done on coreceptor usage in HIV subtypes, particularly subtype C. Limitations of this study are discussed here too.

1. INTRODUCTION

Chapter 5 Conclusion: This is the final chapter of the thesis and draws conclusions from the analysis of the results seen in Chapter 3 and the observations made in the preceding chapter.



Chapter 2

Methodology

2.1 Dataset Curation: Appraising the Performance of Coreceptor Genotyping Tools at Accurately Predicting Coreceptor Usage in HIV-1 Group M Subtype C.

To evaluate the performance of each of the available genotypic tools currently used to predict subtype B tropism, and determine which of these performed best on subtype C sequences, a dataset of 731 sequences with known coreceptor tropism was collated. First, a comprehensive search of the Los Alamos National Laboratory (LANL) HIV Sequence Database (hiv.lanl.gov) was undertaken and 604 sequences using the CCR5 coreceptor only, 53 using the CXCR4 coreceptor only and 43 dual tropic sequences (R5X4) were retrieved. A search of published literature resulted in the addition of 31 more sequences, of which 22 were denoted as CCR5 and 9 as R5X4 sequences (Jensen

2. METHODOLOGY

et al., 2006; Raymond et al., 2010).

The inclusion of multiple sequences from an individual at the same time point could bias the estimation of the accuracy of a genotypic tool. To avoid this potential bias in results, multiple samples from the same individuals were excluded with a single representative sequence randomly selected for each of these individuals.

Multiple sequence alignments of each of the V3 nucleotide sequences were then produced manually using MacClade 4.08 software (Maddison and Maddison, 1992). To ensure alignment consistency, all sequences were aligned to the LANL retrieved HXB2 reference sequence (Korber et al., 1998). Comparing sample sequences to the reference sequence not only allows reference to be made to specifically numbered nucleotide bases relative to the reference sequence, but also facilitates identification of inserted and deleted bases.

2.2 Handling of Ambiguous Nucleotide Bases

Several of the sequences in the dataset contained degenerate base symbols, representing multiple possible alternatives for a single base position within a codon. The presence of these ambiguous nucleotide calls in a sequence can affect the accuracy of genotyping approaches as each of these algorithms is based on the presence or absence of positively or negatively charged amino acids at specific locations within the genome (Sing et al., 2007b). Geno2pheno is the only one of the tools tested that is capable of accounting for ambiguous nucleotides in its genotypic predictions (Sing et al., 2007b). In this instance, the sequences containing ambiguous base symbols were included with

2. METHODOLOGY

all other sequences in the analysis with the geno2pheno web-tool.

The exclusion of these sequences with ambiguous base positions from the analyses with the remaining genotypic tools could artificially influence sensitivity, specificity and accuracy calculations. Thus, if a tested genotyping approach was not designed to account for ambiguous nucleotide positions, all possible combinations of amino acid sequences were output for the sequences containing ambiguous positions, using a script written in Python (RK Shrestha), and each of these was submitted for genotypic testing. The genotypic call of the translated sequences was compared. A worst-case scenario approach similar to that of Sing et al. (Sing et al., 2007b) was employed whereby if one of these translated sequences was predicted as CXCR4-using, the genotyping call for the original sequence was taken to be X4.

2.3 Genotypic Algorithm Evaluation

Viral sequences were separated into three distinct categories based upon their experimentally verified viral phenotype: CCR5-using (R5), CXCR4-using (X4) and dual-tropic (R5X4). Dual-tropic and CXCR4-tropic viruses were studied both separately and together (as CXCR4-using) in order to determine the affect of the conflicting signal of dual-tropic viruses on the sensitivity of each tool.

The coreceptor tropism of every V3 sequence in each of the categories was predicted using a number of genotyping methods. These comprised the Position Specific Scoring Matrix (PSSM) tools, including $PSSM_{X4R5}$ and $PSSM_{SINSI}$ (Jensen et al., 2003) as well as the subtype C PSSM tool (Jensen et al., 2006); geno2pheno (Sing

2. METHODOLOGY

et al., 2007b) and four variants of the wetcat package, including C4.5, C4.5 with p8-p12, PART and SVM (Pillai et al., 2003). Tropism was also predicted using the 11/25 (Fouchier et al., 1992) and 11/24/25 (Cardozo et al., 2007) charge rules. Raymond and colleagues recently proposed a combination of the 11/25 and charge rules for prediction of CXCR4-use in subtype C sequences (Raymond et al., 2010), and this method was evaluated too.

2.3.1 Web PSSM matrices

CCR5, CXCR4 and R5X4 sequences were submitted for genotypic testing to each of the Web PSSM tools as separate sub-datasets, using default settings. The genotypic call for each sequence was determined based on the percentile given for each sequence, where a percentile score of 0.96 and above were considered as CXCR4-using while those below this value were considered as R5. When more than one optimal alignment to the HXB2 reference sequence could be determined by the Web PSSM tools for each of the submitted sequences, multiple output sequences along with their results were generated. These sequences were only considered for further analysis when genotypic predictions made by the matrix were the same for all alignment variations.

2.3.2 Geno2pheno

Sequences in each coreceptor dataset were split into groups of 50 or fewer to meet the maximum handling capability of the geno2pheno web based tool. In selecting how conservative the detection of CXCR4-usage should be, cut-offs of 5%, 10% and 20% were used in this study to determine coreceptor usage with geno2pheno. For clarity purposes, each of the geno2pheno false positive rates used is described as an individual

2. METHODOLOGY

approach throughout this thesis.

2.3.3 Wetcat package

Four of the five classifiers in the wetcat package were evaluated in this study. These included C4.5, C4.5 with positions 8 and 12 only, PART and SVM. According to on-line instructions, all V3 sequences were aligned to the 40 base pair consensus sequence provided on the website, which was designed on training data obtained from the Los Alamos Database. Due to the unconventional length of the consensus V3 sequence, the fifth wetcat classifier, the Charge Rule, was not used in this study. Particular care was taken in aligning test sequences to position 12 in the wetcat alignment, which corresponds to position 11 according to public consensus and is known to be a major determinant in coreceptor determination. The number of correctly and incorrectly predicted sequences was determined for each dataset.

2.3.4 Charge rules

The 11/25 and 11/24/25 charge rules were assessed using a script written in Python (C Meehan), and sequences were predicted as CXCR4-using or CCR5-using based on the principles relevant to each method. Both of these rules use sequence features of the V3 loop exclusively to predict coreceptor usage (Sander et al., 2007). For the 11/25 rule, sequences were determined to be X4-using if a positive charge was found at either position 11 and/or 25 of the V3 loop (Fouchier et al., 1992). The positively charged amino acids are Arginine (R), Lysine (K), and Histidine (H). Similarly, if a positive charge was present at the 11, 24 and/or 25 positions of the V3 loop, the virus was predicted to be CXCR4-using according to the 11/24/25 rule (Cardozo et al., 2007).

2. METHODOLOGY

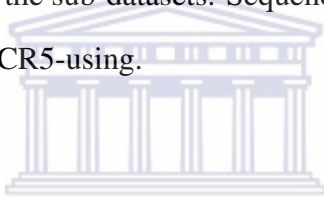
2.3.5 Raymond method

Raymond and colleagues recently proposed a combination of the 11/25 and charge rules for prediction of CXCR4-use in subtype C sequences (Raymond et al., 2010).

One of the following criteria is required for predicting CXCR4 coreceptor usage:

- (1) a R or K at position 11 and/or a K at position 25,
- (2) a R at position 25 and a net charge of greater than or equal to +5, or
- (3) a net charge of greater than or equal to +6 (Raymond et al., 2010).

A script implementing these rules was written in Python (RK Shrestha) and used to assess sequences in each of the sub-datasets. Sequences that did not satisfy any of the criteria were classified as CCR5-using.



2.4 Determining Sensitivity and Specificity of Genotypic Algorithms

Both the ability of an algorithm to correctly predict CXCR4-usage and its ability to correctly predict CCR5-usage in an HIV-infected individual are important in determining the best genotypic approach. Sensitivity corresponds to the ability of the approach to correctly predict CXCR4-use, while specificity corresponds to the ability to correctly predict CCR5-usage. The perfect approach would therefore have 100% sensitivity and 100% specificity and would be able to correctly distinguish between R5 and X4-using viruses.

To calculate each of these statistical values, the number of True Positive (TP), False

2. METHODOLOGY

Positive (FP), True Negative (TN) and False Negative (FN) sequences were recorded, where:

- **True positive** refers to the number of CXCR4-using sequences correctly predicted as CXCR4-using.
- **False positive** refers to the number of CCR5-using sequences incorrectly predicted as CXCR4-using.
- **True negative** refers to the number of CCR5-using sequences correctly predicted as CCR5-using sequences.
- **False negative** refers to the number of CXCR4-using sequences incorrectly predicted as CCR5-using sequences.

The sensitivity of each approach for CXCR4 prediction was calculated as the number of predicted X4 viruses in the CXCR4-using dataset divided by the total number of sequences in the CXCR4-using dataset. This can be expressed as:

$$\frac{TP}{TP + FN}$$

The specificity of each approach for CXCR4 prediction was calculated as the number of predicted R5 viruses in the CCR5-using dataset divided by the total number of sequences in the CCR5-using dataset. This can be expressed as:

$$\frac{TN}{TN + FP}$$

The same statistical approach was used to calculate the sensitivity and specificity of each genotyping method on the CXCR4-exclusive and dual-tropic datasets.

2. METHODOLOGY

2.5 Determining Accuracy of Genotypic Algorithms

Although sensitivity and specificity are both useful measures in determining the predictive ability of an algorithm, neither is a true measure on its own of how good an algorithm is at predicting tropism, and a measure taking both values into account is required. Therefore, an overall accuracy score for each of the approaches used was calculated using:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

where, TP, TN, FP and FN are defined as above. For the CXCR4-exclusive dataset the TP and FN values were calculated only for sequences phenotypically determined to exclusively use CXCR4. For each calculation, the TP and FN values were normalised relative to the TN and FP values to account for the disproportionate number of sequences representing the positive (CXCR4-using or CXCR4-exclusive) and negative (CCR5) datasets (see Appendix 1 for the uncorrected values used to calculate accuracy). This was done by multiplying each of the TP and FN values by the difference in ratio between the total number of CCR5 (TN + FP) sequences and the total number of CXCR4-using sequences (TP + FN).

2.6 Determining Effect of Dual Tropic Viruses on Prediction of CXCR4-usage

Dual tropic viruses can enter host cells using either CCR5 or CXCR4 chemokine receptors and, in some instances, display preferential use for one of these. This may

2. METHODOLOGY

result in mis-prediction of some X4-capable viruses as R5 using. In order to determine the affect of the conflicting signal of dual-tropic viruses on sensitivity estimates, dual-tropic and CXCR4-tropic viruses were studied both separately and together (as CXCR4-using). For this purpose, viral sequences were separated into three distinct categories (R5, X4 and R5X4) based upon their experimentally verified viral phenotype and tested as separate sub-datasets on each of the genotypic tools.

2.7 Dataset Curation: Prevalence of CXCR4-usage in Subtype C

Once the most accurate genotyping tool was established for the prediction of coreceptor tropism in subtype C (as described above), this tool was then used to determine the genotype of all the available subtype C V3 loop sequences. To do this, a comprehensive search of the LANL database and literature for every available HIV-1 subtype C V3 loop sequences was conducted. This second, larger dataset, comprising of a total of 17,353 individual sequences, was largely composed of sequences with unknown phenotypes. Epidemiological and demographic data including subtype, phenotype (where available), sampling year, country of origin, as well as sequence accession number and name was recorded for each sequence. Multiple sequences for the same patient, including clonal sequences from longitudinal studies, were excluded, retaining one, randomly chosen, representative sequence for each of these individuals.

2. METHODOLOGY

2.8 Multiple Sequence Alignment Using RAMICS

Because of the large number of sequences contained in this dataset, an alternative to the MacClade software previously used to align multiple sequences for genotypic testing was used. While MacClade requires the alignment of multiple sequence to be undertaken manually, alignment in RAMICS (Rapid Amplicon Mapping in Codon Space) is automated. Furthermore, MacClade was only able to handle a maximum of 500 sequences at a time, while RAMICS was able to simultaneously align sequences from all individuals to the LANL sourced HXB2 reference sequence. RAMICS is a novel tool employing hidden Markov models to align multiple sequences in codon space. In doing so, it is able to take into consideration both the nucleotide and the amino acid for every position in the reference sequence. It also takes into account and compares the likelihood of insertions, deletions and mutations at each position of the reference sequence (Wright et al., currently under review in *Nucleic Acids Research*).

2.9 Coreceptor Tropism Prediction

Coreceptor tropism predictions were done using geno2pheno as it was previously seen to be the best approach for subtype C coreceptor genotyping (see results). Due to the large number of sequences it contained, the dataset was run in batch mode through geno2pheno by Alexander Thielen. On return of the predicted dataset, the coreceptor tropism of each sequence was then determined based on the significance level previously established to be the most suitable cutoff.

2. METHODOLOGY

2.10 Exploring Prevalence and Patterns of CXCR4-usage in Subtype C

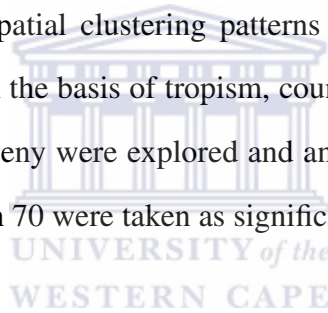
To investigate CXCR4 coreceptor usage in subtype C, we explored the genotyping results from the large dataset containing all available subtype C V3 loop sequences. In order to determine if the epidemic was evolving over time, a direct comparison was performed and the dataset was split into an early (samples between 1984 and 1997) and a late group (samples between 1998 and 2010) (Connell et al., 2008). CXCR4-usage was plotted for each year to determine the prevalence of CXCR4-usage over time and to characterise the emergence of CXCR4-usage. To determine CXCR4-usage patterns of subtype C in earlier sampled sequences, the first identification of CXCR4-usage in HIV-1 subtype C sequence database records was dated.

The geographical occurrence of subtype C and, more specifically CXCR4-using sequences, were also determined by comparing the number of subtype C and CXCR4-using sequences for each country. By comparing the dates for the emergence of CXCR4-usage in each country, the spread of CXCR4-usage within countries could be compared and the CXCR4-usage patterns in subtype C could be established. In this way, potential expansion of CXCR4-usage in subtype C could be identified. The earliest identification of CXCR4-usage was further analysed for countries with the highest numbers of subtype C sequences. Furthermore, by analyzing the geographical data of all available subtype C V3 loop sequences over the years, we attempted to explore any geographic disparities in the emergence of CXCR4-usage in subtype C over the entire period of time for which records are available.

2. METHODOLOGY

2.11 Maximum Likelihood Phylogeny Estimation

To facilitate downstream analysis, each sequence descriptor in the dataset was appended with its coreceptor prediction. Phylogenetic relationships of all available subtype C V3 sequences, with genotypically determined coreceptor tropism, were inferred by the maximum likelihood approach implemented in RAxML v7.0.4 MPI algorithm (Stamatakis, 2006). The GTRGamma model of nucleotide substitution was employed with 100 replicate bootstrap support. To root the tree, the HXB2 reference sequence was used as the outgroup. FigTree v1.3.0. (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to explore the spatial clustering patterns within the resulting phylogeny. Clustering was explored on the basis of tropism, country of origin and sampling year. Supports within the phylogeny were explored and any relationships supported with a bootstrap value greater than 70 were taken as significant.



Chapter 3

Results

3.1 Dataset Compilation: Appraising the Performance of Coreceptor Genotyping Tools at Accurately Predicting Coreceptor Usage in HIV-1 Group M Subtype C.

A total of 731 HIV-1 group M subtype C V3 sequences with experimentally verified coreceptor tropism were initially sourced from the Los Alamos National Laboratory (LANL) HIV Sequence Database (hiv.lanl.gov) and published literature. Of these, 604 sequences using the CCR5 coreceptor only, 53 using the CXCR4 coreceptor only and 43 dual tropic sequences (R5X4) were retrieved from LANL. The remaining 31 sequences were sourced from published literature, of which 22 were denoted as CCR5 and 9 as R5X4 sequences (Jensen et al., 2006; Raymond et al., 2010).

3. RESULTS

Table 3.1: Number of CCR5, CXCR4 and dual tropic sequences obtained from each source.

Source	CCR5-using	CXCR4-using	Dual tropic
LANL	327	25	22
Raymond <i>et al.</i>	3	0	3
Coetzer <i>et al.</i>	19	0	6
Total:	349	25	31

Number of phenotypically determined HIV-1 subtype C sequences obtained from each source.

Furthermore, multiple sequences from individuals were removed, and only one randomly selected representative sequence for each individual was retained, reducing the total number of sequences to 405 (Table 3.1). The final analysis dataset contained V3 loop sequences from 349 CCR5-using and 56 CXCR4-using viruses. Sequences from CXCR4-using viruses were further separated into R5X4 (dual-tropic) and CXCR4-exclusive viruses with 31 and 25 sequences, respectively, comprising these datasets.

3.2 Handling of Ambiguous Nucleotide Bases

The coreceptor usage of every sequence in each of the datasets was predicted using all of the genotyping approaches. Twenty-three of the sequences tested contained at least one ambiguous nucleotide position, and none contained more than four ambiguous base pairs. Of these 23 sequences, 18 were CCR5, one CXCR4 and four dual tropic. Geno2pheno is the only one of the tools tested that is capable of accounting for ambiguous positions in its genotypic predictions (Sing *et al.*, 2007b). To assess all of the other approaches, nucleotide sequences were translated into all the possible combinations of amino acid sequences and if one or more of these translated sequences was

3. RESULTS

predicted as CXCR4-using, the genotyping call for the original sequence was taken as X4. However, for none of the sequences was this approach necessary, as the resolved sequences for each of the ambiguous-containing sequences were predicted to have the same coreceptor usage for each of the methods. Additional sequences generated through the translational process were excluded from the study so as to prevent inflation of sequence numbers. For each of the 23 sequences, all possible translations of the sequence had the same coreceptor tropism prediction for each method. Thus, in this data, ambiguous positions did not affect the genotypic predictions.

3.3 Sensitivity and Specificity of Genotypic Algorithms

The sensitivity of each of the tested approaches at predicting X4 viruses in the CXCR4-using dataset (dual tropic and CXCR4-exclusive combined) varied widely from 40-97% (Figure 3.1 and Table 3.2). The method by Raymond and colleagues had the highest sensitivity at 97% while geno2pheno (FPR₂₀) and C-PSSM both exhibited high sensitivities of greater than 90%. Two variants of the wetcat package, C4.5 and C4.5 with p8-p12, performed most poorly with sensitivities of 40% each.

The specificity of each approach was also calculated, where specificity corresponded to the number of CCR5-tropic viruses correctly predicted as R5 divided by the total number of CCR5-using viral sequences evaluated. All approaches performed well with three having 100% specificity, eight having specificity greater than 90% and geno2pheno FPR₂₀ and the Raymond method exhibiting lower specificity of 86% and 76% respectively (Figure 3.1 and Table 3.2).

3. RESULTS

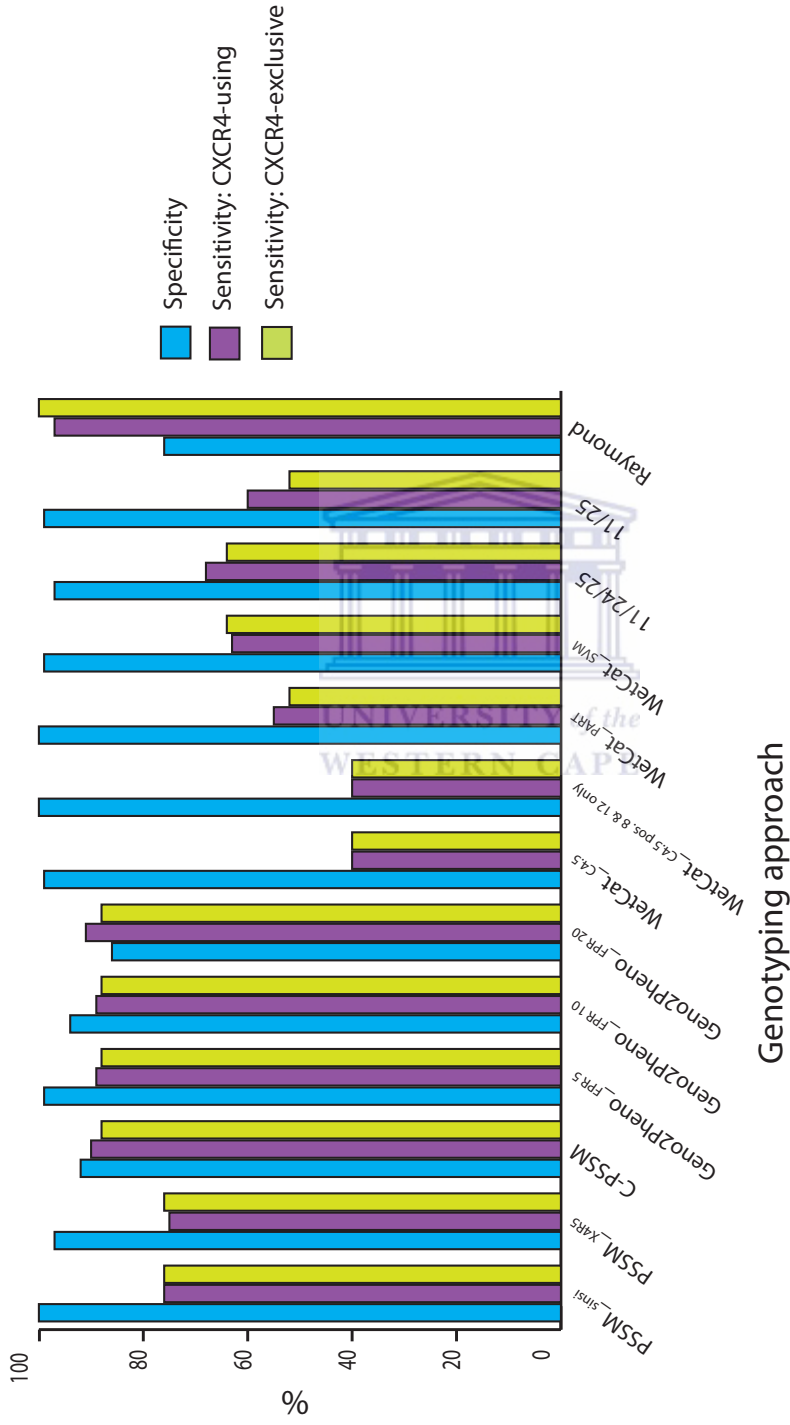


Figure 3.1: Performance of each of the genotyping algorithms in predicting CXCR4-usage. Sensitivity for both the CXCR4-using and CXCR4-exclusive datasets was calculated as the number of viral sequences predicted as CXCR4-using divided by the total number of CXCR4-using or CXCR4-exclusive sequences tested. Specificity corresponds to the number of CCR5-using viruses predicted as R5 divided by the total number of CCR5-using viral sequences evaluated.

3. RESULTS

Table 3.2: Performance of genotyping approaches at predicting CXCR4-usage in viral sequences from individuals infected with HIV-1 group M subtype C.

Method	CXCR4-using sensitivity (%)	Specificity
PSSM _{sinsi}	76	100
PSSM _{X4R5}	75	97
C-PSSM	90	92
Geno2Pheno_FPR ₅	89	99
Geno2Pheno_FPR ₁₀	89	94
Geno2Pheno_FPR ₂₀	91	86
WetCat_C4.5	40	99
WetCat_C4.5 _{pos.8&12}	40	100
WetCat_PART	53	100
WetCat_SVM	63	99
11/24/25	68	97
11/25	60	99
Raymond Approach	97	76

Sensitivity corresponds to the ability of the approach to predict CXCR4-use, while specificity corresponds to the ability to correctly predict CCR5-use.

3.4 Genotypic Algorithm Evaluation

Sensitivity and specificity for the Raymond method were estimated at 97% and 76% respectively for their approach in this study. Compared to the other approaches tested, however, Raymonds method is not the optimal approach. While it does show the highest sensitivity, it also has the lowest specificity of all the approaches tested (Table 3.2). For the other approaches it is found that specificity increases by as much as 24% for three of the approaches relative to the Raymond study, and 23% for a further four approaches.

The sensitivities of the three geno2pheno approaches tested here showed little difference, with geno2pheno FPR₅ and FPR₁₀ both having sensitivities of 89% and dif-

3. RESULTS

fering by only 2% more than FPR_{20} . However, the specificities of these approaches showed a greater difference, with specificities differing by up to 13% between approaches.

Web C-PSSM was the only one of the genotypic algorithms tested that was designed based on a subtype C sequence training dataset. Although not the method with the highest overall sensitivity, the C-PSSM approach out-performed both other PSSM approaches at correctly predicting CCR5 usage. A difference of 15% and 14% in specificity was recorded between it and $PSSM_{X4R5}$ and $PSSM_{sinsi}$ respectively in this study.

Overall, compared to the other approaches tested, the wetcat package of tools performed most poorly in predicting CXCR4-usage in subtype C. There was no difference in the way two variants of the wetcat package, C4.5 and C4.5 with p8-p12 performed, with each having the lowest ability to correctly predict CXCR4-usage.

3.5 Accuracy of Genotypic Algorithms

While some methods are extremely sensitive at correctly predicting CXCR4-use, the optimum approach for clinical implementation also needs to be highly specific in correctly identifying viruses that do not use the CXCR4 receptor. Thus, an accuracy score was calculated for each of the approaches tested that takes into account an approaches sensitivity and specificity (Table 3.3). For the CXCR4-using dataset, it was found that three of the 13 approaches tested have an accuracy of 90% or greater at predicting coreceptor usage in HIV-1 group M subtype C viral sequences with $geno2pheno FPR_5$ being the most accurate of all approaches tested with an accuracy of 94% (89% sen-

3. RESULTS

Table 3.3: Accuracy of genotyping approaches at correctly predicting coreceptor tropism.

Method	CXCR4-using accuracy	CXCR4-exclusive accuracy	R5X4 accuracy
PSSM _{sinsi}	88	88	88
PSSM _{X4R5}	86	87	86
C-PSSM	91	90	91
Geno2Pheno_FPR ₅	94	93	94
Geno2Pheno_FPR ₁₀	92	91	92
Geno2Pheno_FPR ₂₀	88	87	90
WetCat_C4.5	70	70	70
WetCat_C4.5 _{pos.8&12}	70	70	70
WetCat_PART	77	76	79
WetCat_SVM	81	82	81
11/24/25	81	81	82
11/25	79	76	82
Raymond Approach	86	88	85

Accuracy scores are presented for a combined dataset containing CXCR4-using viruses (both CXCR4-exclusive and dual-tropic viruses) as well as separately for the CXCR4-exclusive and dual-tropic viral sequences.

sitivity and 99% specificity, Table 3.3). Two variants of the wetcat package, C4.5 and C4.5 with p8-p12, both perform poorest with accuracy scores of 70% (Table 3.3).

3.6 Effect of Dual Tropic Viruses on Prediction of CXCR4-usage

The CXCR4-using viruses were separated into CXCR4-exclusive and dual-tropic viral sequences and the accuracy of each of the approaches at correctly predicting coreceptor tropism was calculated (Table 3.3). When dual-tropic sequences are excluded, the accuracy of three of the approaches increases minimally, with four methods showing

3. RESULTS

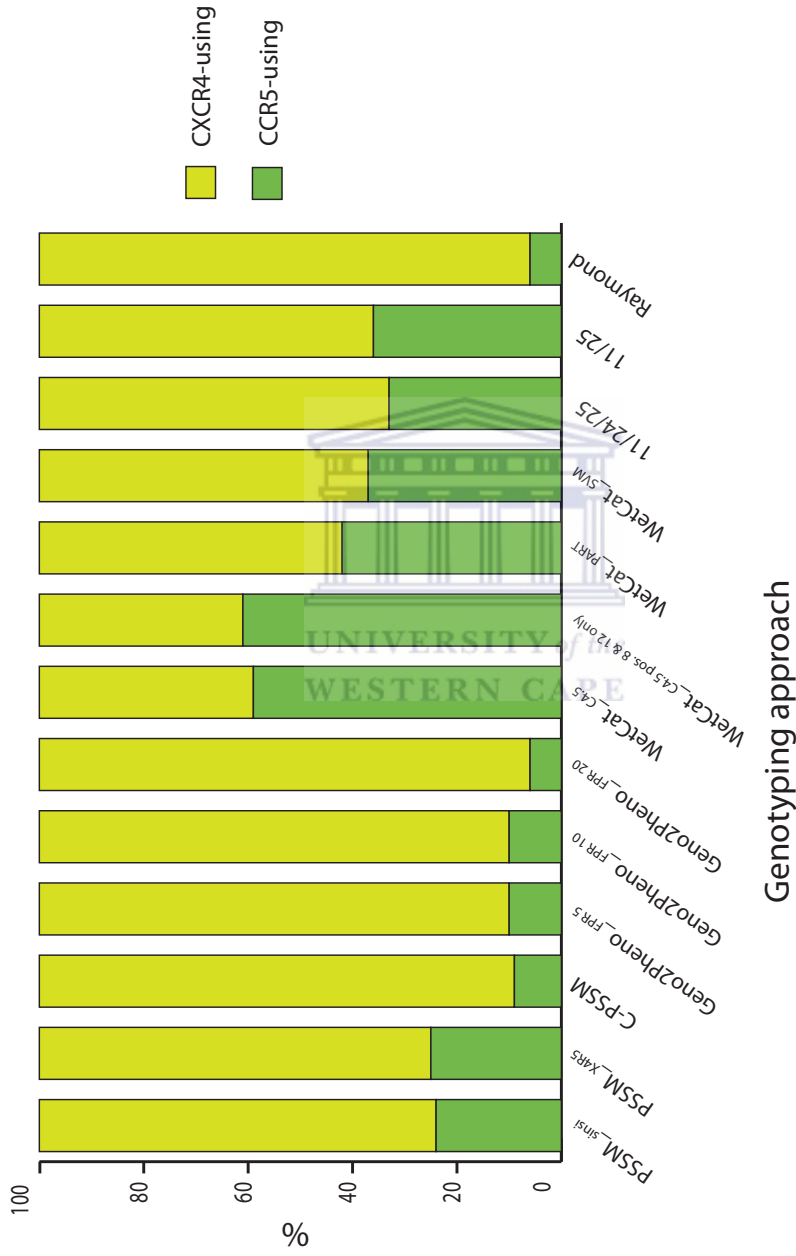


Figure 3.2: Ability of each approach at predicting CXCR4-usage in dual-tropic viral sequences. The percentage of dual-tropic sequences predicted as CCR5-using and CXCR4-using is shown with dark and light shaded areas of each bar corresponding to the percentage of sequences predicted as CCR5-using and CXCR4-using respectively.

3. RESULTS

no change in accuracy and six showing a slight decrease of 1% in accuracy (Table 3.3). Similarly, when the dual-tropic viruses were studied separately there was minimal effect on the accuracy of each of the approaches when compared to CXCR4-exclusive viruses (Figure 3.2). There was significant variability in the ability of the approaches to accurately predict CXCR4-usage in dual-tropic viruses, ranging from 40% (wetcat C4.5 with p8-p12) to 94% (geno2pheno FPR₂₀) of sequences from dual-tropic viruses predicted as CXCR4-using (Figure 3.2). It appears that, in subtype C at least, the ability of approaches to predict CXCR4- usage in dual tropic viruses directly correlates with their ability to predict CXCR4-usage in CXCR4-exclusive viruses.

3.7 Dataset Compilation: Prevalence and Mode of CXCR4-usage in Subtype C

A total of 17,353 HIV-1 group M subtype C V3 sequences were initially retrieved from the LANL sequence database and published literature. This dataset was largely composed of viral sequences with experimentally undetermined phenotypes. The exclusion of multiple sequences for the same patient reduced the dataset to a final total of 12,121 sequences, which was used to determine the prevalence of CXCR4-usage in subtype C. This was done using geno2pheno (FPR₅), which was previously determined to be the most accurate approach for HIV-1 subtype C tropism determination.

3.8 Multiple Sequence Alignment Using RAMICS

The RAMICS tool for multiple sequence alignment was found to be robust, accurate and generated biologically relevant multiple sequence alignments rapidly. RAMICS

3. RESULTS

was able to handle a considerably large number of sequences simultaneously - in this instance, a total number of 12,121 sequences were aligned to the HXB2 reference sequence in 35.2 seconds. This approach to sequence alignment ensured consistency in handling insertions, deletions and mutations throughout the alignments and was able to account for variances in length.

3.9 CXCR4-usage Patterns in HIV-1 Subtype C Sequences

From its first emergence in database records in 1988, the presence of sampled subtype C CXCR4-using sequences is not seen consistently over the years, up until a decade later (Figure 3.3). From 1998 onwards the presence of X4 sequences appears to stabilize between 4 – 9% per year (Figure 3.3). The largest number of CXCR4-using sequences were identified in sequences generated from samples collected in 1997, at 16% of the total number of sequences for that year (Figure 3.3). These 31 sequences were all from India (Figure 3.4e). In total, CXCR4-using sequences comprised less than 5% of the entire subtype C dataset, while CCR5 sequences make up a significantly larger proportion of the 12,121 subtype C sequences in this dataset at over 95%.

Sequenced subtype C viral sequences first appear in the database records in 1984 (Figure 3.5). For this year, eight sequences were recorded - all of which were predicted as CCR5-using. A notable increase in the number of recorded subtype C sequences is seen in 1988 with a total of 121 sequences recorded for this year - 2 of which were CXCR4-using from Ethiopia. This figure peaks in 2005, when the largest number of recorded subtype C sequences was 1,397. Of these, 96% were CCR5-using sequences while 4% were CXCR4-using sequences.

3. RESULTS

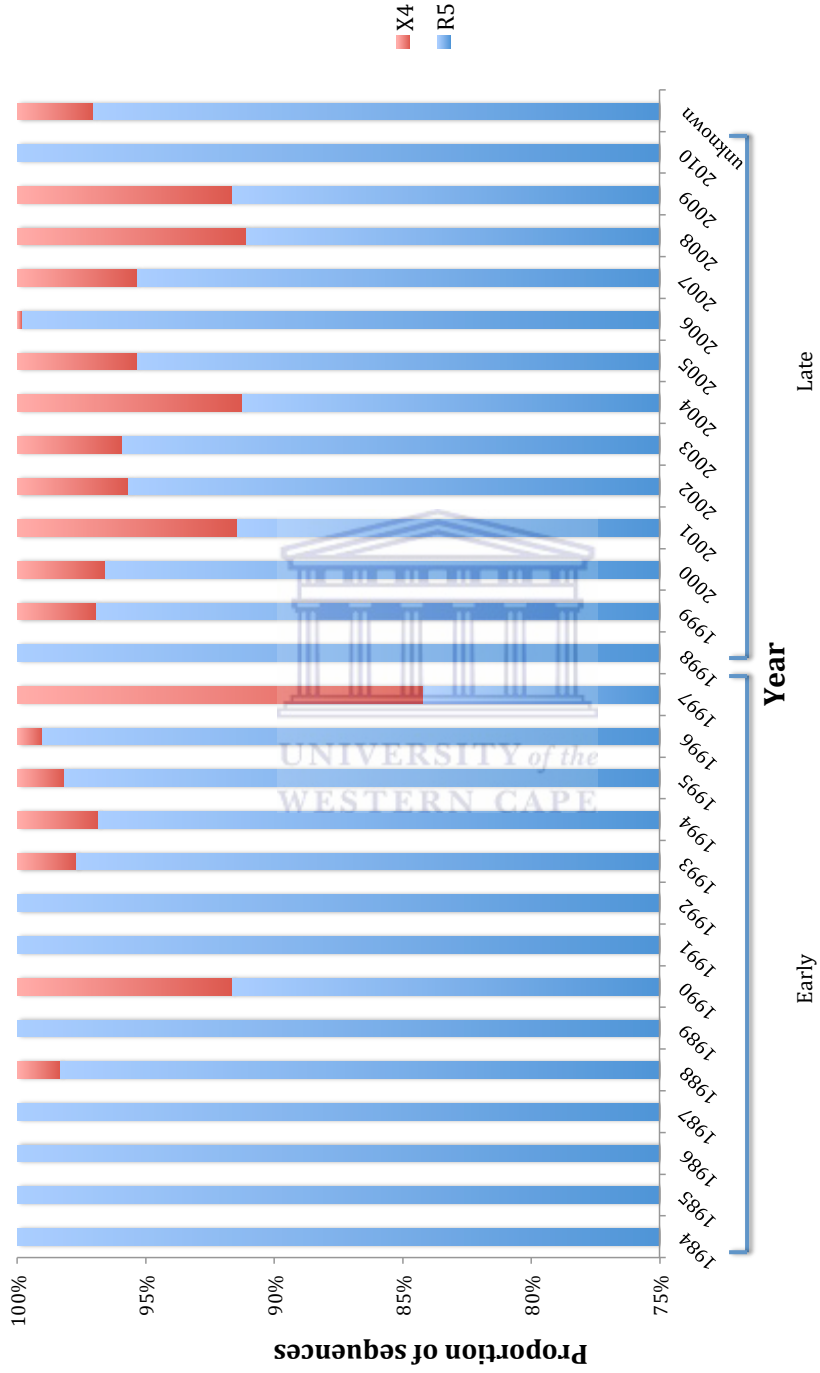


Figure 3.3: Prevalence of CXCR4-usage over time.

3. RESULTS

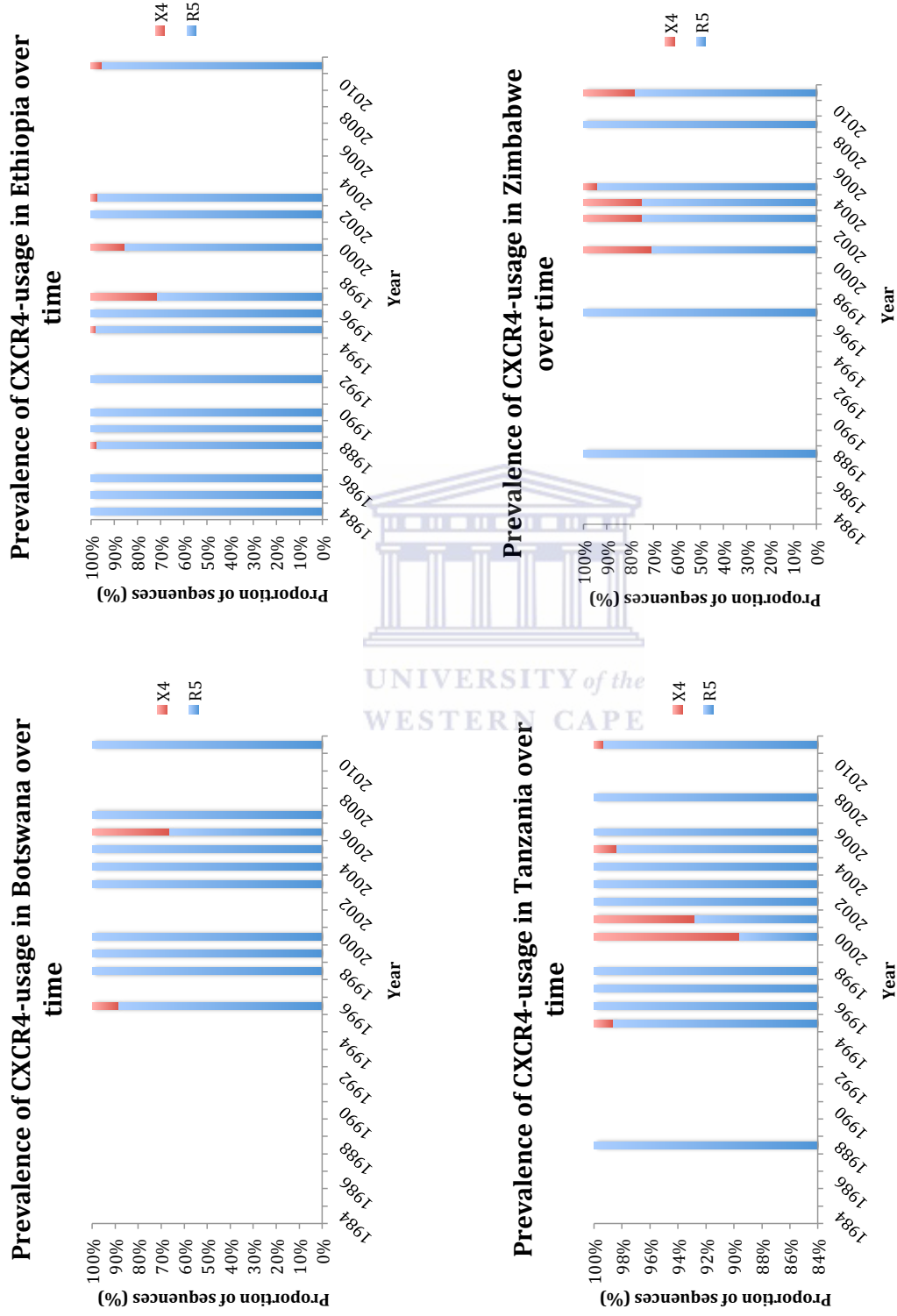


Figure 3.4: Prevalence of CXCR4-usage over time for each of the countries with the highest number of subtype C sequences.

3. RESULTS

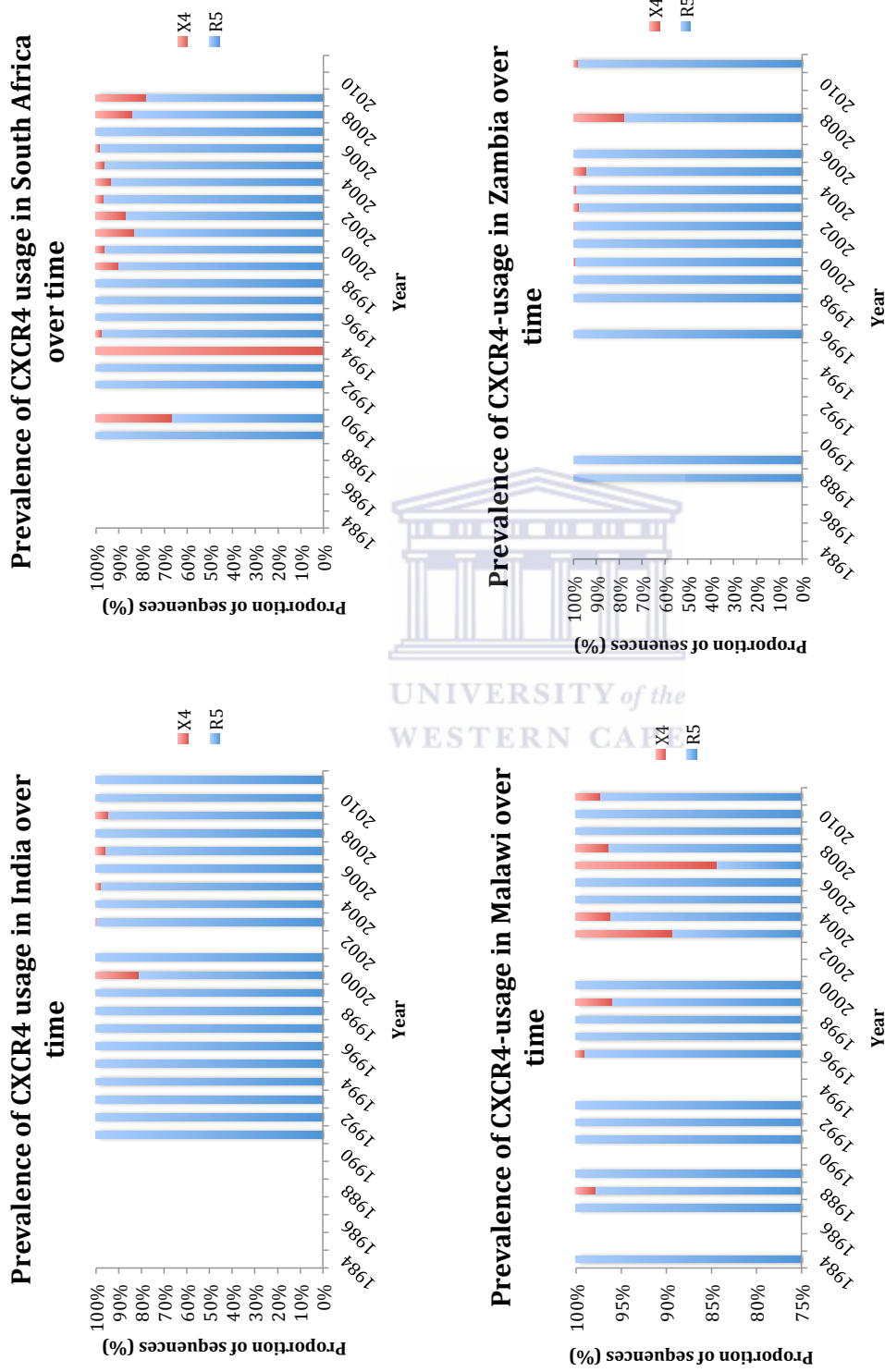


Figure 3.4 (Continued): Prevalence of CXCR4-usage over time for each of the countries with the highest number of subtype C sequences

3. RESULTS

The earliest appearance of CXCR4-using sequences in the LANL database dates to 1988 when three sequences were recorded - two of which were from Ethiopia, while one came from Malawi (Figure 3.4). The earliest observations of sequenced CXCR4-using sequences in the database records do not appear to be consistent for all countries, with CXCR4-usage first appearing in South African records in 1990, Tanzania in 1995, while India and Zambia subtype C CXCR4-using sequences first appear a decade later in the year 2000 and for Zimbabwe in 2001.

3.10 Prevalence of CXCR4-usage in HIV-1 Subtype C sequences



Of the 77 countries from which the 12,121 sequences derive, 60 have fewer than 100 subtype C sequences each in the LANL database, while nine countries have between 100 and 200 subtype C sequences each. To better understand the patterns of CXCR4-usage, larger groups of subtype C sequences were studied in more detail. Thus, countries with more than 200 representative subtype C sequences were selected for further analysis, including CXCR4-usage over time. Of the eight countries that have above 200 subtype C sequences (Table 3.4), excluding India, which has a total of 1,201 subtype C sequences, all are African countries, and include: Botswana, Ethiopia, Tanzania, Zimbabwe, South Africa, Malawi and Zambia. The largest number of subtype C sequences recorded came from Zambia, with a total of 2,804 unique sequences (See Appendix B).

3. RESULTS

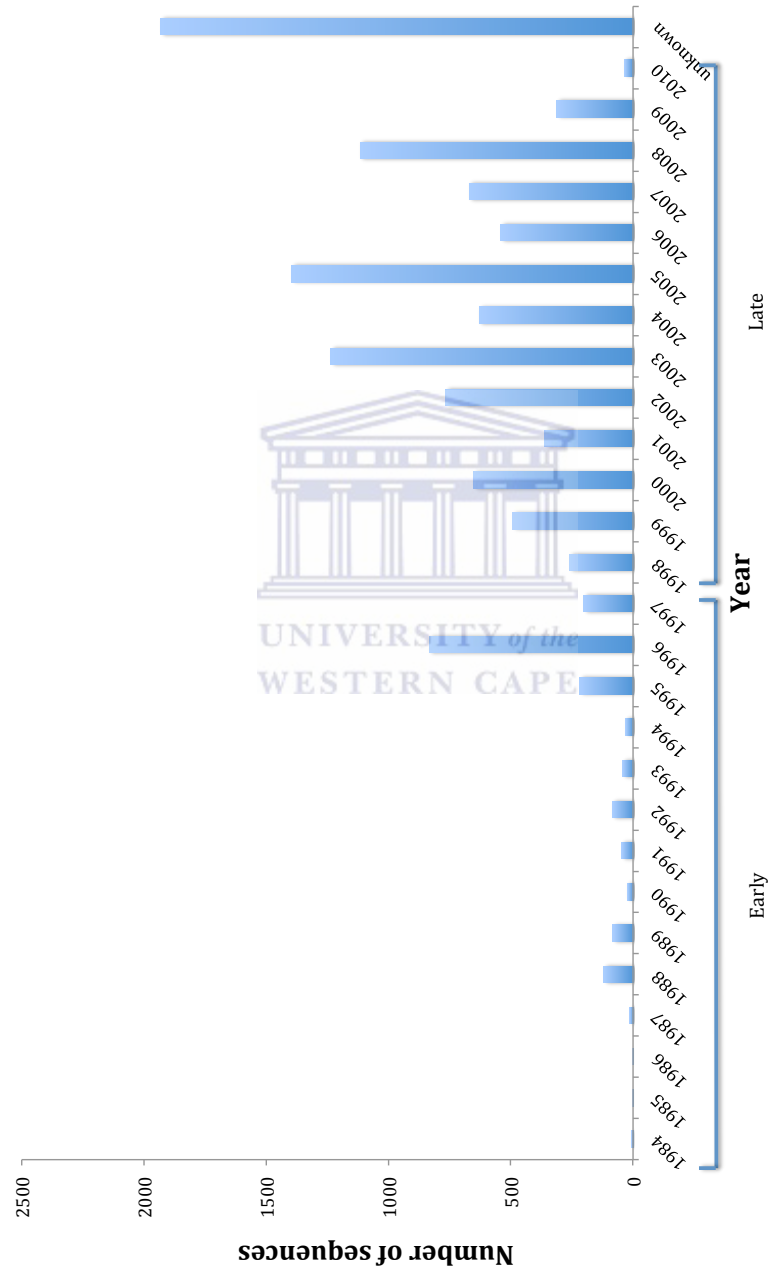


Figure 3.5: Number of subtype C sequences over time

3. RESULTS

Table 3.4: Number of predicted R5 and X4 sequences for countries with more than 200 subtype C sequences each.

Country of origin	Total number of sequences	Predicted CCR5-using	Predicted CXCR4-using
Botswana	225	221	4
Ethiopia	467	426	41
Tanzania	479	464	15
Zimbabwe	612	564	48
India	1201	1191	10
South Africa	2265	2151	114
Malawi	2462	2371	91
Zambia	2804	2698	106

In this study, the proportion of CXCR4-using sequences is not seen to be consistent for each country. While Malawi and Zambia were the countries with highest number of subtype C sequences recorded (each with over 2,000 subtype C sequences seen) (Table 3.4), the proportion of CXCR4-using sequences for these countries was relatively low at around 4% of the total subtype C viral population (Figure 3.6). In contrast to this, the highest proportion of CXCR4-using viral sequences were seen in Ethiopia and Zimbabwe, with 9% and 8% respectively (Figure 3.6) despite each of these countries showing fewer than 500 subtype C sequences each in the LANL database. Although it was one of the countries with the greatest number of subtype C sequences listed, at 1%, India had the lowest proportion of observed CXCR4-using sequences recorded.

3. RESULTS

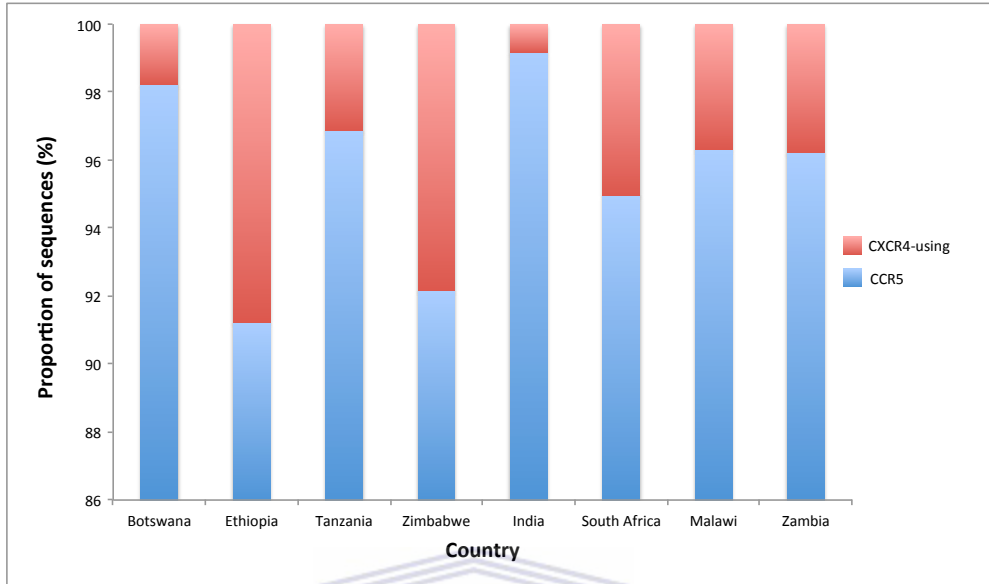


Figure 3.6: Proportion of CXCR4-using sequences from countries with more than 200 subtype C sequences each.

3.11 Maximum Likelihood Phylogeny Estimation

The GTRGamma substitution model was used since it was found by to be the most appropriate substitution model according to ModelTest (Posada and Crandall, 1998). Phylogenetic analysis based on 12,121 subtype C *env* V3 loop sequences showed that CXCR4-using sequences were spread throughout the phylogeny. This observation was indicative of the convergent evolution in the development of CXCR4-usage in HIV-1 Group M subtype C (Figure 3.7). No significant clustering patterns of CXCR4-using sequences are found, indicating that CXCR4 receptor usage has arisen independently in different subtype C populations, as opposed to being established by founder effect. CXCR4-using sequences are spread throughout the tree, suggesting that X4 usage is not being transmitted, but rather evolving within individuals.

3. RESULTS

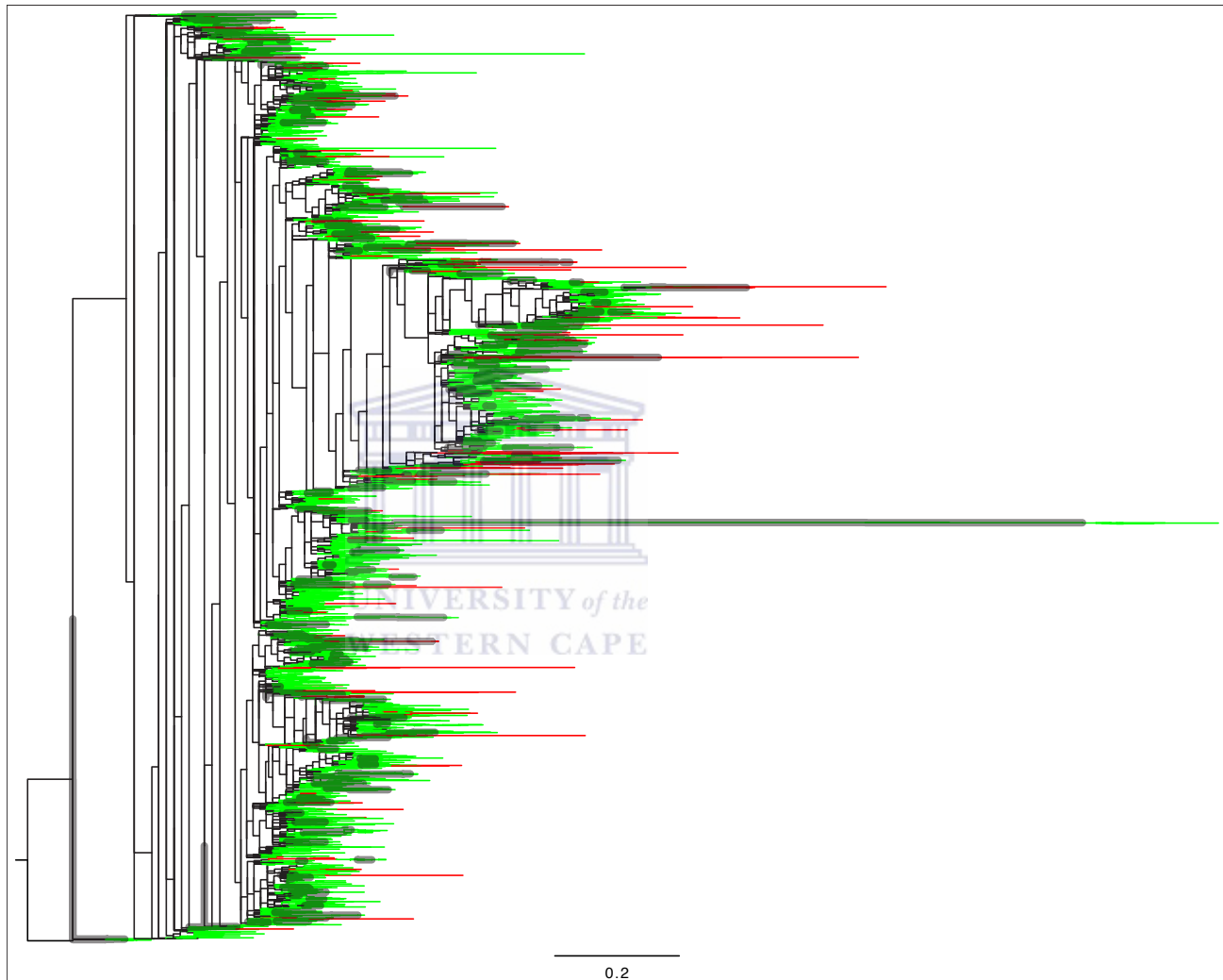


Figure 3.7: Distribution of CCR5 (green) and CXCR4-using (red) sequences with bootstrap values greater than or equal to 70 highlighted. The tree was rooted with the HXB2 reference sequence (shown in black). Horizontal branch lengths are drawn to scale with the bar at the bottom indicating nucleotide substitutions per site.

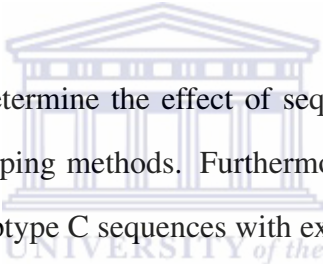
3. RESULTS

The phylogeny was fairly well supported, with bootstrap estimates of 70% and above occurring throughout the tree (Figure 3.7). The greatest support for the phylogeny at bootstrap confidence values of 70% and above was found at the terminal nodes, with few basal branches providing support at this level. However, the backbone of the lineage was not well supported, with little support found deeper in the phylogeny. This is further supported by the many short branch lengths seen deeper in the tree. Of note is a long branch midway in the phylogeny representing a group of highly divergent sequences.



Chapter 4

Discussion



This is the first study to determine the effect of sequences from dual tropic viruses on the sensitivity of genotyping methods. Furthermore, it is the first comprehensive study using all available subtype C sequences with experimentally verified coreceptor tropism to evaluate the performance of various genotyping tools at accurately predicting CXCR4-usage in HIV-1 subtype C. It is also the first study to undertake a large-scale analysis of the emergence and prevalence of CXCR4 coreceptor usage in HIV-1 group M subtype C, using a geographically diverse dataset.

4.1 Ability to Account for Ambiguous Nucleotide Positions

The presence of ambiguous nucleotide calls, particularly within the codons encoding for amino acid positions 11, 24 and 25, can substantially reduce the ability of approaches to correctly predict coreceptor usage (Sing et al., 2007b). The clinical ap-

4. DISCUSSION

plication of genotypic tropism testing must often account for viral sequences with such ambiguous nucleotide calls. Previous studies have shown that the accuracy of genotypic tropism prediction methods is lower on clinically derived data than on clonal data (Sing et al., 2007b; Bozek et al., 2013). In a recent study, Bozek and colleagues examined the effect of amino-acid ambiguities on the prediction accuracy of several clinical models for tropism prediction. The results of three datasets were compared: one without any ambiguous sequences, another with ambiguous sequences and the third dataset having all ambiguous position replaced by gaps. It was found that combined information from both types of positions is important for tropism prediction (Bozek et al., 2013). Geno2pheno is the only one of the tools tested that is capable of accounting for ambiguous base positions in its genotypic predictions, while none of the other methods were designed to handle ambiguous base positions. Thus, in our study it was found that the ability to account for ambiguous nucleotide positions in geno2pheno gives it a distinct advantage over all of the other approaches tested.

4.2 Effect of Dual Tropic Viruses on Prediction of CXCR4-usage

Dual-tropic viruses are a unique class of viruses in that they can enter host cells using either CCR5 or CXCR4 chemokine receptors. However, some dual-tropic viruses can exhibit preferential use of one of these (Berger et al., 1998; Huang et al., 2007, 2009). From a clinical perspective, it is imperative that genotyping approaches correctly identify the CXCR4-using capabilities of dual-tropic viruses. Genotyping algorithms have been shown to vary widely in their predictive ability of CXCR4-usage in subtype B

4. DISCUSSION

dual-tropic viruses (Mefford et al., 2008). In general, approaches were observed to underestimate the frequency of CXCR4-usage in dual tropic viruses (Mefford et al., 2008). Thus, the effect of dual-tropic viruses on the accuracy of each of the tested genotyping approaches was assessed.

In this study it appears that, in subtype C at least, the ability of approaches to predict CXCR4-usage in dual tropic viruses directly correlates with their ability to predict CXCR4-usage in CXCR4-exclusive viruses. Such an observation does not appear to hold true in subtype B, however, where some methods with high sensitivity for prediction of CXCR4-using viruses in subtype B (Garrido et al., 2008) show low accuracy for the prediction of CXCR4-usage in subtype B dual-tropic viral sequences (Mefford et al., 2008). This is most likely due to a limited number of CXCR4-using sequences in the training datasets. Geno2pheno, however, does show high accuracy for the prediction of CXCR4-usage in subtype B dual-tropic viruses (Mefford et al., 2008). This is probably due to the inclusion of a high number of CXCR4-using sequences in the training dataset.

4.3 Genotypic Algorithm Evaluation

In selecting how conservative the detection of CXCR4-usage should be, German Treatment Guidelines were chosen as a significance level for the geno2pheno tool in this study, as the method most suitable for triplicate sampling (Deutsche-AIDS-Gesellschaft, 2012). The German Treatment Guidelines describe three case scenarios for the prescription of CCR5 antagonists based on various false positive rates (FPR), where the FPR is the probability of classifying an R5 virus as X4. The higher the FPR, the more

4. DISCUSSION

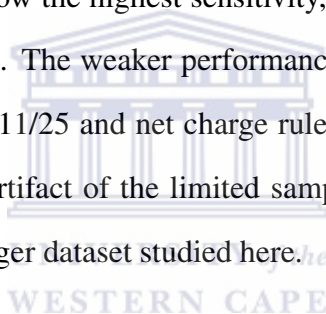
specific the algorithm is at detecting CCR5 usage.

Taking into account the latest findings in published literature as well as from scientific conferences, the German Guidelines prescribe the following treatment procedure: for patients with several treatment options, the use of CCR5 antagonist is only recommended if CXCR4-using viruses are detected at a FPR above 15%. For patients with limited treatment options, the administration of a CCR5 antagonist may also be considered at a FPR between 5% and 15%. If the FPR is however less than 5%, the risk of a false negative prediction is very high – in about one third of the predictions using a FPR of 5%, CXCR4-using viruses are not detected. In these cases, CCR5 antagonist therapy is generally not recommended. Thus, cut-offs of 5%, 10% and 20% were used in this study to determine coreceptor usage with geno2pheno. For HIV-1 subtype C it was found that a FPR of 5% was most accurate.

While predicting CXCR4-usage with high sensitivity is important, the ability to correctly identify R5 variants as CCR5-using is equally as important in reducing the amount of false positives that would result in incorrect clinical interpretations. Thus, the specificity (proportion of CCR5-tropic viruses correctly predicted as R5) of each approach was also calculated. All the approaches tested had high sensitivities, and performed well in their ability to predict CCR5 usage. These high specificity values are consistent with previous observations in both HIV-1 subtype B and non-B subtypes that all approaches, in general, are better at correctly predicting CCR5-usage than CXCR4-usage (Poveda et al., 2009; Jensen et al., 2006; Raymond et al., 2010; Garrido et al., 2008; Seclen et al., 2010).

4. DISCUSSION

Raymond and colleagues had previously evaluated nine of the 13 genotyping approaches studied here using a smaller, geographically limited subtype C dataset comprising 55 CCR5 and 15 CXCR4 viral sequences sampled from Malawi and France (Raymond et al., 2010). They reported that the optimal approach for subtype C genotyping was a combination of the 11/ 25 and net charge rules (Raymond approach) with sensitivity and specificity for CXCR4-usage prediction in subtype C of 93.3% and 96.4% respectively. However, when compared to the other approaches tested using a larger and more geographically diverse dataset, Raymonds method is not the optimal approach. While it does show the highest sensitivity, it also has the lowest specificity of all the approaches tested. The weaker performance on our comprehensive subtype C dataset of the combined 11/25 and net charge rule proposed by Raymond and colleagues is most likely an artifact of the limited sample size/diversity in their dataset that is not present in the larger dataset studied here.



Web C-PSSM was the only one of the genotypic algorithms tested that was designed based on a subtype C sequence training dataset. In describing C-PSSM, Jensen and colleagues used a dataset consisting of 228 CCR5 sequences and 51 CXCR4 sequences (from 200 and 20 subjects respectively) (Jensen et al., 2006) and reported a sensitivity of 75%, substantially less than the 90% sensitivity reported here, with comparable specificities of 94% and 92%.

Similarly, Garrido and colleagues evaluated the performance of eight of the subtype B designed approaches studied here on their ability to predict HIV-1 tropism in non-B subtypes (Garrido et al., 2008). When compared to our results, they found that geno2pheno performed considerably poorer on non-B subtypes, with a sensitivity of

4. DISCUSSION

61% and a specificity of 73%. However, their study was composed (in part) of a relatively small number of subtype C sequences, forming only 8.7% of the total non-B subtype dataset.

In our study, two variants of the wetcat package, C4.5 and C4.5 with p8-p12, had the lowest sensitivities of all approaches evaluated, and higher specificity values. These findings are consistent with previous observations on both subtype B and non-B subtype viral sequences (Raymond et al., 2010; Garrido et al., 2008; Seclen et al., 2010).

While some methods are extremely sensitive at correctly predicting CXCR4-use, the optimum approach for clinical implementation also needs to be highly specific in correctly identifying viruses that do not use CXCR4, as neither is a true measure on its own of how good an algorithm is at predicting tropism. Thus a measure taking both values into account is required, and an accuracy score was calculated for each of the approaches tested that takes into account an approach's sensitivity and specificity. Based on this method, an approach with a very high sensitivity and low specificity scored poorly in accuracy while a fairly high sensitivity and specificity scored better in accuracy. Thus, by taking both sensitivity and specificity into account, accuracy is the best way to summarise the complete performance of a method.

4.4 Multiple Sequence Alignment Using RAMICS

The second dataset collated for this study consisted of 12,121 unique sequences, representing all available subtype C V3 loop sequences. Because this many sequences could not be accurately aligned manually, an automated approach was used to generate

4. DISCUSSION

multiple sequence alignments. RAMICS was selected in favour of other widely-used multiple sequence alignment methods, including Muscle, for its ability to generate accurate, biologically relevant sequence alignments.

RAMICS is able to take into account both the nucleotide and the amino acid for every position in the reference sequence, and compares the likelihood of insertions, deletions and mutations at each position of the reference sequence. In so doing, it was able to correctly account for the two amino acid insertion in the HXB2 reference sequence in its global alignments. This feature of RAMICS was recently highlighted in a study by Wright and colleagues, who compared the multiple sequence alignment capabilities of RAMICS with Muscle (Wright *et al.*, currently under review in *Nucleic Acids Research*). In their study, RAMICS was able to generate significantly more accurate alignments of HIV-1 subtype C V3 loop sequences when compared to Muscle (Figure 4.1).

4.5 Prevalence and Patterns of CXCR4-usage in HIV-1 Subtype C

HIV-1 Subtype C accounts for over 50% of global infections and over 95% of infections in southern African countries (Hemelaar et al., 2011). Despite this, little is known about the characteristics of this subtype, particularly those relating to its coreceptor usage patterns and prevalence. In this study, we assembled the largest dataset of subtype C sequences to help elucidate the prevalence and patterns of CXCR4-usage in HIV-1 Subtype C.

4. DISCUSSION

In the LANL database records, early reports of subtype C sequences were low with initial reports of fewer than 10 sequences per year. From the late 1980's onwards, this figure began steadily increasing, with thousands of sampled subtype C sequences reported in the following years. This surge in reported subtype C sequences could possibly be attributed to the high infectivity of HIV-1 subtype C, leading to a greater number of observed infections. It may also be indicative of an increased interest in subtype C research, leading to a greater number of sequences being sampled, especially in regions in Africa where HIV research centers are well established.

HIV-1 subtype C was first identified in North East Africa in the early 1980's (Salmi-
nen et al., 1996; McCormack et al., 2002) and in this study we found that sequences
were first recorded in HIV database records from as early as 1984. HIV-1 subtype
C sequences are spread globally, with database records appearing for 77 countries.
While the majority of countries have relatively few representative subtype C sequences
recorded, some countries have a disproportionately higher number. On closer inspec-
tion, the countries with the highest number of subtype C sequences are all developing
countries, with most of these countries being in eastern and southern Africa regions.
This pattern is a reflection of the global subtype C pandemic, which sees the highest
number of HIV infections occurring in eastern and southern Africa as well as in India
(Neogi et al., 2010). In this study too, we see that India is one of the highest contribut-
ing countries to the global records of subtype C sequences.

Conflicting reports on coreceptor usage in HIV-1 subtype C have been published.
Earlier studies reported a 'remarkably' low frequency of CXCR4-using sequences,

4. DISCUSSION

with CCR5-using sequences dominating throughout infection, suggesting that a switch to CXCR4-usage was rare or never occurred in subtype C (Abebe et al., 1999). However, an increasing number of more recent studies have shown that a switch from CCR5 to CXCR4-usage can occur, and does so in a manner similar to subtype B (Cilliers et al., 2003; Pollakis et al., 2004).

Using a substantially larger dataset with 12,121 subtype C sequences spanning 26 years, we show that a switch to CXCR4-usage is seen in subtype C for well over 20 years. This finding indicates that although seldom reported, a switch to CXCR4-usage has consistently taken place in subtype C over time, suggesting that CXCR4-usage has not evolved recently in this subtype. The lack of reported subtype C CXCR4-usage in earlier studies may be as a result of sampling artifact or inadequate techniques to determine coreceptor tropism rather than to a biological difference in this subtype.

Furthermore, it has been suggested that between 30-50% of subtype C infected individuals exhibit a change to CXCR4-usage during disease progression (Connell et al., 2008; Kassaye et al., 2009; Michler et al., 2008; Cilliers et al., 2003; Papathanasopoulos et al., 2002; Johnston et al., 2003). Esbjörnsson and colleagues reported a frequency of 15% CXCR4-usage in subtype C sequences (Esbjörnsson et al., 2010), while a study by Connell and colleagues reported that of the 20 South African AIDS patients they examined (19 of which had subtype C infections), 30% of primary isolates were CXCR4-using. These higher figures, as compared to previous studies, have been suggested to indicate an increase in frequency of CXCR4-usage in subtype C over time (Connell et al., 2008).

4. DISCUSSION

In our study, CXCR4-using sequences, at less than 5%, form a relatively small proportion of the total number of sequences in the dataset of globally-derived sequences. Although this figure is much lower than previous reports, the earlier reports were typically based on a small sample size, with sequences derived from a single study group or country. The figure we report is likely a more accurate estimate of the frequency of CXCR4-usage in subtype C as it is calculated as a proportion of a larger, more inclusive dataset, representing sequences from 77 countries over a period of 26 years.

The high frequency of CXCR4-usage reported in some studies has also been attributed to ART (anti-retroviral therapy) exposure. Pramanik Sollerkvist and colleagues found that the frequency of CXCR4 use in subtype C patients failing ART's was higher compared to treatment naive patients (Pramanik Sollerkvist et al., 2013). Duri and colleagues report that CCR5 usage dominated in the 28 treatment naive mother-infant groups in their study, and that a switch to CXCR4-usage rarely happened in this patient group (Duri et al., 2011). It has been suggested that ARTs create a suitable environment for a switch to CXCR4-usage to take place, complicating the administration of CCR5 inhibitors to treatment-experienced patients (Pramanik Sollerkvist et al., 2013).

4.6 Study Limitations

Despite the significance of these findings, it is important to note the limitations of this study. One major caveat is that sequences in this study do not represent the entire spectrum of globally circulating subtype C sequences, and are not a true reflection of

4. DISCUSSION

the subtype C epidemic. Rather, they indicate the frequency of sequences observed in database records and are not an accurate measure of the actual prevalence of subtype C in a specific country. Because the sequence data wasn't sampled/sequenced for this study, we can only infer from observations in the data, however we cannot definitely say the observations are true. Thus, we can't draw specific conclusions about prevalence of coreceptor usage in these countries.

Another known caveat is the inability to account for differences between phenotypic assays used to determine tropism, particularly the differences between older assays with a generally lower sensitivity for minor variants and the newer more sensitive assays. Furthermore, data on disease stage or therapy stage was not available for every sequence, while it is known that samples obtained during early or late infections would have a significant impact on the prevalence of a specific tropism.

The use of bulk or Sanger sequencing for tropism prediction may be considered as a further confounder in this study. Here, a consensus is taken of the viral population in an infected individual, and the sequence tested genotypically may not be derived from only one virus. Ideally, genotypic testing on SGA sequences, the sequences of clones or NGS sequence data would be a more accurate approach and would provide a better prediction of whether X4 variants are present or absent in an individual's viral population.

A final point worth mentioning, is the handling of multiple sequences for the same patient. In this regard, only one randomly selected sequence was retained for further analysis, accounting for approximately 5200 discarded sequences - each potentially

4. DISCUSSION

representing an outlier sequence. However, each sequence being randomly selected essentially negates this concern, as the possibility that the chosen sequence - for each patient with greater than 2 representative sequences - is an outlier would be minimised.



Chapter 5

Conclusions

This study adds to the limited characterisation of CXCR4-usage in HIV-1 subtype C. Using a comprehensive, geographically diverse dataset, we find that geno2pheno (FPR₅) is the most accurate approach available for the prediction of coreceptor tropism in HIV-1 subtype C viral sequences, with an accuracy of 94% (89% sensitivity and 99% specificity). Coupled with its high accuracy, the ability of geno2pheno to account for ambiguous nucleotide calls in V3 sequences gives it a distinct advantage over all other approaches for coreceptor genotyping of sequence data generated from population-based sequencing. Web C-PSSM, the only tool tested that was designed on subtype C sequence data, had a slightly lower prediction ability compared to subtype B-trained geno2pheno.

Based on these findings, we conclude that the geno2pheno coreceptor tool may be used as a reliable genotypic predictor in clinical settings to establish the viability of CCR5-antagonist therapies using drugs such as Maraviroc. At approximately USD100-200, genotypic sequencing provides a rapid and cost effective alternative to phenotypic testing, particularly in resource limited areas. The significance of genotypic testing is further highlighted when compared to the cost of phenotypic testing,

5. CONCLUSIONS

with the Trofile assay being undertaken at a cost of approximately USD1500 - an amount which few individuals can afford in the most severely affected regions of the world.

Furthermore, we report that in HIV-1 group M subtype C, sequences from dual-tropic viruses have minimal effect on the performance of genotypic tools and the optimal approaches for prediction of CXCR4-usage in sequence from viruses that use CXCR4 exclusively also perform best at predicting CXCR4-use in dual-tropic viral variants. Based on this, it appears that viral genotyping of envelope sequences from subtype C infected individuals is feasible with the correct approach and can be undertaken with a high degree of confidence that CXCR4-usage will be accurately identified in both CXCR4-exclusive and dual tropic variants.

In determining HIV-1 subtype C prevalence and patterns of CXCR4-usage, we find that a switch to CXCR4-usage is seen in subtype C for well over 20 years. A switch from CCR5 to CXCR4-usage has consistently taken place in subtype C over time, suggesting that CXCR4-usage has not evolved recently.

We find that in our dataset, which constitutes the largest collection of subtype C sequences, the overall frequency of CXCR4-using sequences seen is not only lower than the prevalence previously reported for subtype B sequences, but also lower than found in other studies on subtype C. We report a frequency of 5% for the switch to CXCR4-usage in subtype C. These observations are important in understanding the rapid spread of HIV-1 subtype C in the developing world and may have broad implications for the design of intervention and treatment strategies.

5. CONCLUSIONS



Chapter 6

Appendix A



6. APPENDIX A

Table 6.1: Tables detailing the uncorrected numbers of true positives (CXCR4-usage correctly predicted in CXCR4-using sequences), true negatives (CCR5-usage correctly predicted in CCR5-using sequences), false positives (CXCR4-usage incorrectly predicted in CCR5-using sequences) and false negatives (CCR5-usage incorrectly predicted in CXCR4-using sequences) predicted by each of the approaches. Results are shown for (A) CXCR4-using sequences, (B) CXCR4-exclusive sequences and (C) dual-tropic sequences.

A

Method	TP	TN	FP	FN
Web PSSM _{-sinsi}	44	348	1	14
Web PSSM _{-X4R5}	43	338	10	14
Web C-PSSM	52	317	29	6
Geno2Pheno_FPR ₅	50	344	5	6
Geno2Pheno_FPR ₁₀	50	328	21	6
Geno2Pheno_FPR ₂₀	51	299	50	5
WetCat_C4.5	23	346	3	34
WetCat_C4.5 _{pos.8&12}	23	348	1	35
WetCat _{-PART}	32	348	1	26
WetCat _{-SVM}	36	344	3	21
11/24/25 Charge Rule	38	340	9	20
11/25 Charge Rule	34	347	2	24
Raymond Approach	55	262	82	2

B

Method	TP	TN	FP	FN
Web PSSM _{-sinsi}	19	348	1	6
Web PSSM _{-X4R5}	19	338	10	6
Web C-PSSM	22	317	29	3
Geno2Pheno_FPR ₅	22	344	5	3
Geno2Pheno_FPR ₁₀	22	328	21	3
Geno2Pheno_FPR ₂₀	22	299	50	3
WetCat_C4.5	10	346	3	15
WetCat_C4.5 _{pos.8&12}	10	348	1	15
WetCat _{-PART}	13	348	1	12
WetCat _{-SVM}	16	344	3	9
11/24/25 Charge Rule	16	340	9	9
11/25 Charge Rule	13	347	2	12
Raymond Approach	25	262	82	0

6. APPENDIX A

C

Method	TP	TN	FP	FN
Web PSSM _{-sinsi}	25	348	1	8
Web PSSM _{-X4R5}	24	338	10	8
Web C-PSSM	30	317	29	3
Geno2Pheno_FPR ₅	28	344	5	3
Geno2Pheno_FPR ₁₀	28	328	21	3
Geno2Pheno_FPR ₂₀	29	299	50	2
WetCat_C4.5	13	346	3	19
WetCat_C4.5 _{pos.8&12}	13	348	1	20
WetCat_PART	19	348	1	14
WetCat_SVM	20	344	3	12
11/24/25 Charge Rule	22	340	9	11
11/25 Charge Rule	21	347	2	12
Raymond Approach	30	262	82	2

Chapter 7

Appendix B



7. APPENDIX B

Table 7.1: Number of HIV-1 subtype C sequences recorded per country.

Country	Number of sequences recorded
Georgia	1
Hungary	1
Indonesia	1
Korea (South)	1
Lebanon	1
Latvia	1
Mexico	1
Romania	1
Sudan	1
Thailand	1
Uruguay	1
Vietnam	1
Yemen	1
Argentina	2
Austria	2
Greece	2
Malaysia	2
Nigeria	2
New Zealand	2
Philippines	2
Belarus	3
Estonia	3
Italy	3
Venezuela	3
Gabon	4
Norway	4
Taiwan	4
Rwanda	5
Singapore	5
Gambia	6
Iran	6
Cameroon	7
Cuba	8
Germany	8
Djibouti	8
Finland	8
Guinea-Bissau	8
Myanmar	8
Russian Federation	8

Table continued...

7. APPENDIX B

Table 7.1 continued

Country	Number of sequences recorded
Seychelles	8
Spain	9
Bangladesh	11
Belgium	11
Somalia	11
Angola	12
Czech Republic	12
Israel	13
Senegal	13
Fiji	15
Japan	15
Australia	16
Cyprus	17
Great Britain	24
Papua New Guinea	32
Portugal	33
Netherlands	36
Nepal	36
China	38
Denmark	43
France	53
Sweden	89
Congo	90
Uganda	105
United States	105
Kenya	107
Mozambique	110
Burundi	119
Brazil	122
Switzerland	148
Botswana	225
Ethiopia	467
Tanzania	479
Zimbabwe	612
India	1201
South Africa	2265
Malawi	2462
Zambia	2804
Unknown	18
Total:	12121

Chapter 8

Appendix C

From Publication:



Crous, S., Shrestha, RK. and Travers, SA. (2012). Appraising the performance of genotyping tools in the prediction of coreceptor tropism in HIV-1 subtype C viruses. *BMC Infectious Diseases*, 12:203. doi:10.1186/1471-2334-12-203

Available online: <http://www.biomedcentral.com/1471-2334/12/203>

8. APPENDIX C

Conference oral and poster presentations:

1) Crous, S., Shrestha, RK. and Travers, SA. *Characterising coreceptor usage in HIV-1 group M subtype C*. 19th International HIV Dynamics and Evolution Conference, 2012, Asheville, North Carolina, USA. (Poster presented by SA. Travers)

2) Crous, S., Shrestha, RK. and Travers, SA. *Characterising coreceptor usage in HIV-1 group M subtype C*. Joint South African Genetics Society and South African Society for Bioinformatics and Computational Biology: The Data-mining Revolution, September 2012, Stellenbosch, South Africa. (Oral presentation, presented by S. Crous)

3) Crous, S., Shrestha, RK. and Travers, SA. *Characterising coreceptor usage in HIV-1 group M subtype C*. UWC Faculty of Science Postgraduate Research Open Day 2012, Bellville, Cape Town. (Oral presentation, presented by S. Crous)

RESEARCH ARTICLE

Open Access

Appraising the performance of genotyping tools in the prediction of coreceptor tropism in HIV-1 subtype C viruses

Saleema Crous, Ram Krishna Shrestha and Simon A Travers*

Abstract

Background: In human immunodeficiency virus type 1 (HIV-1) infection, transmitted viruses generally use the CCR5 chemokine receptor as a coreceptor for host cell entry. In more than 50% of subtype B infections, a switch in coreceptor tropism from CCR5- to CXCR4-use occurs during disease progression. Phenotypic or genotypic approaches can be used to test for the presence of CXCR4-using viral variants in an individual's viral population that would result in resistance to treatment with CCR5-antagonists. While genotyping approaches for coreceptor-tropism prediction in subtype B are well established and verified, they are less so for subtype C.

Methods: Here, using a dataset comprising V3-loop sequences from 349 CCR5-using and 56 CXCR4-using HIV-1 subtype C viruses we perform a comparative analysis of the predictive ability of 11 genotypic algorithms in their prediction of coreceptor tropism in subtype C. We calculate the sensitivity and specificity of each of the approaches as well as determining their overall accuracy. By separating the CXCR4-using viruses into CXCR4-exclusive (25 sequences) and dual-tropic (31 sequences) we evaluate the effect of the possible conflicting signal from dual-tropic viruses on the ability of a of the approaches to correctly predict coreceptor phenotype.

Results: We determined that geno2pheno with a false positive rate of 5% is the best approach for predicting CXCR4-usage in subtype C sequences with an accuracy of 94% (89% sensitivity and 99% specificity). Contrary to what has been reported for subtype B, the optimal approaches for prediction of CXCR4-usage in sequence from viruses that use CXCR4 exclusively, also perform best at predicting CXCR4-use in dual-tropic viral variants.

Conclusions: The accuracy of genotyping approaches at correctly predicting the coreceptor usage of V3 sequences from subtype C viruses is very high. We suggest that genotyping approaches can be used to test for coreceptor tropism in HIV-1 group M subtype C with a high degree of confidence that they will identify CXCR4-usage in both CXCR4-exclusive and dual tropic variants.

Keywords: Human immunodeficiency virus, Coreceptor, Chemokine receptors, CXCR4, CCR5, Genotype, Phenotype, Subtype C

Background

To enable cell entry by HIV, the gp120 glycoprotein, present in a trimeric arrangement on the surface of a HIV virion, must first bind to a CD4 receptor on the target cell [1-3]. This binding induces a conformational change in the gp120/gp41 trimer complex [4,5] thereby enabling binding of a chemokine receptor, either CCR5 or CXCR4 [6]. CCR5-tropic viruses are associated with

primary transmission and can persist throughout infection [6]. In as many as 50% of HIV-1 subtype B infections, a switch to CXCR4-usage has been observed and this switch is generally regarded as an indicator of disease progression [7-10]. Early studies of HIV-1 subtype C suggested that a switch to CXCR4-usage was less common in subtype C compared to subtype B [11,12], however more recent studies have suggested that between 30-50% of subtype C infected individuals exhibit a change to CXCR4-usage during disease progression [13-18].

* Correspondence: simon@sanbi.ac.za
South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Belville 7535, South Africa

8. APPENDIX C

Dual-tropic viruses (R5X4) capable of using either CCR5 or CXCR4 for host cell entry have been described [19] as have dual-tropic viruses that, while capable of using either receptor for cell entry, exhibit preferential use of either CCR5 (dual-R) or CXCR4 (dual-X) [20,21]. Detecting the presence of dual-tropic viruses in an individual's viral population is difficult however, as a mixed population of R5 and X4 viruses will be identified as dual in a population-based phenotyping assay.

Determining the coreceptor usage profile of an individual's viral population has been used as an indicator of disease progression and in more recent years as an approach for detecting resistance to CCR5 antagonists such as maraviroc [22-24]. Phenotypic assays, such as Monogram Bioscience's Trofile™ assay [25], are the most effective means of elucidating the coreceptor tropism of a viral population. These approaches, however, are expensive, laborious and unavailable for routine use in all laboratories [26,27]. Thus, genotyping approaches have been suggested to be a viable alternative for routine coreceptor tropism testing [28]. While many amino acid positions throughout gp120 have been suggested to influence coreceptor affinity and tropism [29-35], the V3 loop appears to be the strongest determinant of coreceptor tropism with amino acid mutations affecting V3 net charge, charge at positions 11, 24 and 25 and glycan binding patterns all implicated in causing a switch from CCR5- to CXCR4-usage [36-41].

Early genotypic algorithms predicted the coreceptor tropism of HIV-1V3 sequences using the properties of the amino acids at positions 11 and 25 while later algorithms account for various properties of the entire V3 loop [39,40,42-45]. With the exception of C-PSSM [43] and the Raymond combined 11/25 and net charge rules [46], all of these approaches have been optimised for coreceptor tropism prediction in subtype B and show varying levels of sensitivity at predicting CXCR4-usage in subtype B [47].

Despite HIV-1 subtype C accounting for almost 60% of worldwide HIV infections [48], the genetic determinants of the switch in coreceptor use are less-well understood than in subtype B. Conflicting reports have been published with some suggesting that these determinants are the same for subtype C as subtype B [46], while others have presented evidence to the contrary [43]. Jensen and colleagues developed the only subtype C specific genotyping tool with a reported sensitivity of 75% [43] while others evaluated the ability of this and other algorithms trained on subtype B data at correctly predicting CXCR4-use in subtype C sequence data [46]. They found that the most appropriate approach for predicting CXCR4-usage in subtype C were C-PSSM and their combined 11/25 and net charge rule [46]. When specificity was considered, however, Raymond and

colleagues approach was significantly better than C-PSSM (96.4% versus 81.8%). The dataset used in this study, however, did not represent the entire spectrum of HIV-1 subtype C diversity in that it had a limited number of phenotyped sequences (55 R5 and 15 X4 sequences) collected from only two countries (Malawi and France).

In this study we have collated a large dataset consisting of all obtainable subtype C sequences with experimentally verified coreceptor tropism and used this to evaluate the performance of various genotyping tools at accurately predicting CXCR4-usage in HIV-1 subtype C. Further, we determine the effect of sequences from dual-tropic viruses on the sensitivity of genotyping methods.

Results and discussion

In total 731 HIV-1 group M subtype C V3 sequences with experimentally verified coreceptor tropism were retrieved. Only one representative sequence for each individual was retained reducing the total number of sequences to 405. The final analysis dataset (available on request) contained sequences from 349 CCR5-using and 56 CXCR4-using viruses. Sequences from CXCR4-using viruses were further separated into R5X4 (dual-tropic) and CXCR4-exclusive viruses with 31 and 25 sequences, respectively, comprising these datasets.

The coreceptor usage of every sequence in each of the datasets was predicted using all of the genotyping approaches. 23 of the sequences tested contained at least one ambiguous nucleotide position. Geno2pheno is the only one of the tools tested that is capable of accounting for ambiguous positions in its genotypic predictions [44]. To assess all of the other approaches, we translated the nucleotide sequences into all the possible combinations of amino acid sequences and if one or more of these translated sequences was predicted as CXCR4-using, the genotyping call for the original sequence was taken as X4. For each of the 23 sequences, all possible translations of the sequence had the same coreceptor tropism prediction for each method. Thus, in this data, ambiguous positions did not affect the genotypic predictions. However, in many cases the presence of ambiguous nucleotide calls, particularly within the codons encoding for amino acid positions 11, 24 and 25, would substantially reduce the ability of approaches to correctly predict coreceptor usage [44]. Thus, the ability to account for ambiguous nucleotide positions in geno2pheno gives it a distinct advantage over all of the other approaches tested here.

The sensitivity of each of the tested approaches at predicting X4 viruses in the CXCR4-using dataset (dual-tropic and CXCR4-exclusive combined) varied widely from 40-97% (Table 1 and Figure 1). The method by Raymond and colleagues performed best with 97%

8. APPENDIX C

Table 1 Performance of genotyping approaches at predicting CXCR4-usage in viral sequences from individuals infected with HIV-1 group M subtype C

Method	CXCR4-using sensitivity (%)	Specificity
PSSM _{sim}	76	100
PSSM _{X4R5}	75	97
C-PSSM	90	92
Geno2Pheno _{FPR5}	89	99
Geno2Pheno _{FPR10}	89	94
Geno2Pheno _{FPR20}	91	86
WetCat _{C4.5}	40	99
WetCat _{C4.5 pos. 8&12}	40	100
WetCat _{PART}	55	100
WetCat _{SVM}	63	99
11/24/25	68	97
11/25	60	99
Raymond	97	76

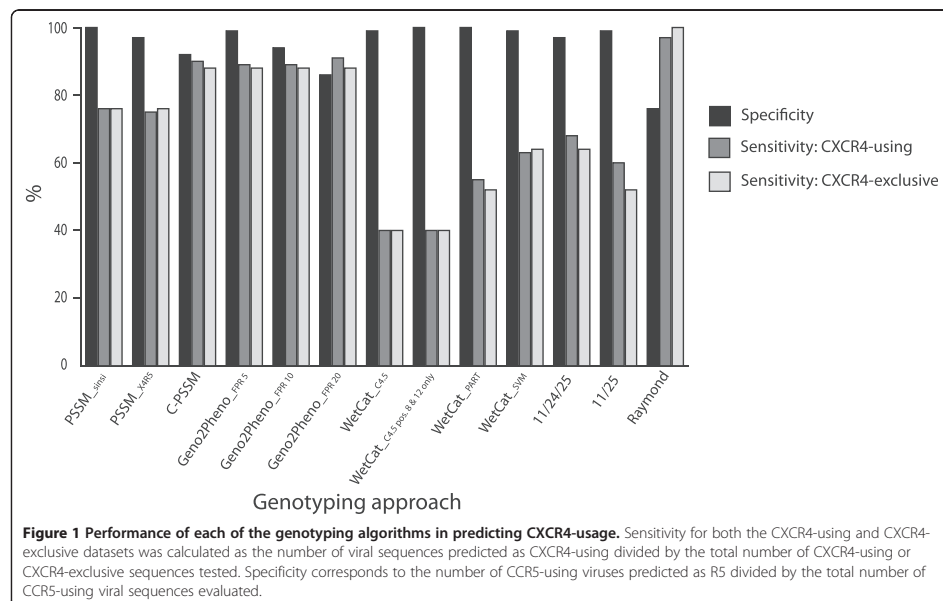
Sensitivity corresponds to the ability of the approach to predict CXCR4-use, while specificity corresponds to the ability to correctly predict CCR5-use.

sensitivity while Geno2pheno (FPR20) and C-PSSM exhibited high sensitivities greater than 90%. Two variants of the wetcat package, C4.5 and C4.5 with p8-p12, performed most poorly with sensitivities of 40%,

consistent with previous observations on both subtype B and non-B subtypes [46,47,49].

While predicting CXCR4-usage with high accuracy is important, the ability to correctly identify R5 variants as CCR5-using is equally as important in reducing the amount of false positives that would result in incorrect clinical interpretations. Thus, we also calculated the specificity (proportion of CCR5-tropic viruses correctly predicted as R5) of each approach. All approaches performed well with three having 100% specificity, eight having specificity greater than 90% and geno2pheno (FPR20) and Raymond exhibiting lower specificity of 86% and 76% respectively (Table 1 and Figure 1). These high specificity values are consistent with previous observations in both HIV-1 subtype B and non-B subtypes that all approaches, in general, are better at correctly predicting CCR5-usage than CXCR4-usage [22,43,46,47,49].

Raymond and colleagues had previously evaluated nine of the 13 approaches studied here using a smaller, geographically limited subtype C dataset comprising 55 R5 and 15 X4 viral sequences sampled from Malawi and France [46]. They reported that the optimal approach for subtype C genotyping was a combination of the 11/25 and net charge rules with sensitivity and specificity for CXCR4-usage prediction in subtype C of 93.3% and 96.4% respectively. Using the larger and more



8. APPENDIX C

geographically diverse dataset studied here, we estimate sensitivity of 97% and a specificity of 76% for this approach. Compared to the other approaches tested, however, Raymond's method is not the optimal approach. While it does show the highest sensitivity, it also has the lowest specificity of all the approaches tested (Table 1). For the other approaches we find that sensitivity increases by as much as 22% for five of the approaches relative to the Raymond study, while the sensitivity of PSSM_{sinsi}, PSSM_{X4R5} and C-PSSM drops by 4%, 5% and 3% respectively. We suggest that the weaker performance on our comprehensive subtype C dataset of the combined 11/25 and net charge rule proposed by Raymond and colleagues is most likely an artifact of the limited sample size/diversity in their dataset that is not present in the larger dataset studied here. In describing C-PSSM, Jensen and colleagues used a dataset consisting 228 R5 sequences and 51 X4 sequences (from 200 and 20 subjects respectively) [43] and reported a sensitivity of 75%, substantially less than the 90% sensitivity reported here, with comparable specificities of 94% and 92%.

While some methods are extremely sensitive at correctly predicting CXCR4-use, the optimum approach for clinical implementation also needs to be highly specific in correctly identifying viruses that do not use CXCR4. Thus, we have calculated an accuracy score for each of the approaches tested that takes into account an approach's sensitivity and specificity (Table 2). For the CXCR4-using dataset, we find that three of the 13 approaches tested have an accuracy of 90% or greater at

predicting coreceptor usage in HIV-1 group M subtype C viral sequences with geno2pheno (FPR5) being the most accurate of all approaches tested with an accuracy of 94% (89% sensitivity and 99% specificity, Table 2). Two variants of the wetcat package, C4.5 and C4.5 with p8-p12, both perform poorest with accuracy scores of 70% (Table 2).

Dual-tropic viruses are a unique class of viruses in that they can enter host cells using either CCR5 or CXCR4 chemokine receptors, however, some dual-tropic viruses can exhibit preferential use of one of these [19-21]. From a clinical perspective, it is imperative that genotyping approaches correctly identify the CXCR4-using capabilities of dual-tropic viruses. Genotyping algorithms have been shown to vary widely in their predictive ability of CXCR4-usage in subtype B dual-tropic viruses [50]. In general, approaches were observed to underestimate the frequency of CXCR4-usage in dual tropic viruses [50]. Thus, we sought to investigate the effect of dual-tropic viruses on the accuracy of each of the genotyping approaches tested. The CXCR4-using viruses were separated into CXCR4-exclusive and dual-tropic viral sequences and the accuracy of each of the approaches at correctly predicting coreceptor tropism was calculated (Table 2). When dual-tropic sequences are excluded, the accuracy of three of the approaches increases minimally, with four methods showing no change in accuracy and six showing a slight decrease of 1% in accuracy (Table 2). Similarly, when the dual-tropic viruses were studied separately there was minimal effect on the accuracy of each of the approaches (Table 2). There was significant variability in the ability of the approaches to accurately predict CXCR4-usage in dual-tropic viruses, ranging from 40% (wetcat C4.5 with p8-p12) to 94% (Geno2pheno FPR20) of sequences from dual-tropic viruses predicted as CXCR4-using (Figure 2). It appears that, in subtype C at least, the ability of approaches to predict CXCR4-usage in dual tropic viruses directly correlates with their ability to predict CXCR4-usage in CXCR4-exclusive viruses. Such an observation does not appear to hold true in subtype B, however, where some methods with high sensitivity for prediction of CXCR4 viruses in subtype B [47], show low accuracy for the prediction of CXCR4-usage in subtype B dual-tropic viral sequences [50]. Geno2pheno, however, does show high accuracy (90%) for the prediction of CXCR4-usage in subtype B dual-tropic viruses [50].

Conclusion

Using a comprehensive, geographically diverse dataset, we find that geno2pheno (FPR5) is the most accurate approach for the prediction of coreceptor tropism in HIV-1 subtype C viral sequences. Coupled with its high accuracy, the ability of geno2pheno to account for

Table 2 Accuracy of genotyping approaches at correctly predicting coreceptor tropism

Method	CXCR4-using accuracy	CXCR4-exclusive accuracy	R5X4 accuracy
PSSM _{sinsi}	88	88	88
PSSM _{X4R5}	86	87	86
C-PSSM	91	90	91
Geno2Pheno _{FPR5}	94	93	94
Geno2Pheno _{FPR10}	92	91	92
Geno2Pheno _{FPR20}	88	87	90
WetCat _{C4.5}	70	70	70
WetCat _{C4.5 pos. 8&12}	70	70	70
WetCat _{PART}	77	76	79
WetCat _{SVM}	81	82	81
11/24/25	81	81	82
11/25	79	76	82
Raymond	86	88	85

Accuracy scores are presented for a combined dataset containing CXCR4-using viruses (both CXCR4-exclusive and dual-tropic viruses) as well as separately for the CXCR4-exclusive and dual-tropic viral sequences.

8. APPENDIX C

approach was not designed to account for ambiguous nucleotide positions, all possible combinations of amino acid sequences were output and a worst-case scenario approach was employed whereby if one of these translated sequences was predicted as CXCR4-using, the genotyping call for the original sequence was taken as X4.

Viral sequences were separated into three distinct categories (R5, X4 and R5X4) based upon their experimentally verified viral phenotype. Dual-tropic and CXCR4-tropic viruses were studied both separately and together (as CXCR4-using) in order to determine the affect of the conflicting signal of dual-tropic viruses on sensitivity estimates. The sensitivity of each approach for CXCR4 prediction was calculated as the number of predicted X4 viruses in the CXCR4-using dataset divided by the total number of sequences in the CXCR4-using dataset. The specificity of each approach for CXCR4 prediction was calculated as the number of predicted R5 viruses in the CCR5-using dataset divided by the total number of sequences in the CCR5-using dataset. The same method was used to calculate the sensitivity and specificity of each genotyping method on the CXCR4-exclusive and dual-tropic datasets.

Further, an overall accuracy score for each of the approaches used was calculated using:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

where, for the CXCR4-using dataset, TP corresponds to the number of CXCR4-using sequences predicted as CXCR4-using, TN the number of R5 sequences predicted as CCR5-using, FP the number of R5 sequences predicted as CXCR4-using and FN the number of CXCR4-using sequences predicted as CCR5-using. For the CXCR4-exclusive dataset the TP and FN values were calculated only for sequences phenotypically determined to exclusively use CXCR4. For each calculation we normalized the TP and FN values relative to the TN and FP values to account for the disproportionate number of sequences representing the positive (CXCR4-using or CXCR4-exclusive) and negative (CCR5) datasets (see Additional file 1: Table S1 for the uncorrected values).

Additional file

Additional file 1: Table S1. Tables detailing the uncorrected numbers of true positives (CXCR4-usage correctly predicted in CXCR4-using sequences), true negatives (CCR5-usage correctly predicted in CCR5-using sequences), false positives (CXCR4-usage incorrectly predicted in CCR5-using sequences) and false negatives (CCR5-usage incorrectly predicted in CXCR4-using sequences) predicted by each of the approaches. Results are shown for (A) CXCR4-using sequences, (B) CXCR4-exclusive sequences and (C) dual-tropic sequences.

Abbreviations

HIV-1: Human immunodeficiency virus type 1; FPR: False positive rate; CXCR4: C-X-C chemokine receptor type 4; CCR5: C-C chemokine receptor type 5.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SC collated the dataset, tested the various genotyping approaches and wrote the first version of the manuscript. RKS wrote software for the translation of ambiguous nucleotides and for the approach by Raymond and colleagues as well as collating results from a number of the approaches used. SAAT conceived the study, participated in its design and wrote the final manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Conor Meehan for providing the scripts for implementing the 11/25 and 11/24/25 rules. This work was supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of South Africa (grant # 64751) as well as a National Research Foundation of South Africa Blue Skies grant (grant # 75899) and a student bursary to RKS from Atlantic Philanthropies (grant # 62302).

Received: 5 March 2012 Accepted: 27 August 2012

Published: 2 September 2012

References

1. Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA: **The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus.** *Nature* 1984, **312**(5996):763–767.
2. Maddon PJ, Dalgleish AG, McDougal JS, Clapham PR, Weiss RA, Axel R: **The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain.** *Cell* 1986, **47**(3):333–348.
3. McDougal JS, Maddon PJ, Dalgleish AG, Clapham PR, Littman DR, Godfrey M, Maddon DE, Chess L, Weiss RA, Axel R: **The T4 glycoprotein is a cell-surface receptor for the AIDS virus.** *Cold Spring Harb Symp Quant Biol* 1986, **51**(Pt 2):703–711.
4. Sattentau QJ, Moore JP: **Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding.** *J Exp Med* 1991, **174**(2):407–415.
5. Liu J, Bartesaghi A, Borgnia MJ, Sapiro G, Subramaniam S: **Molecular architecture of native HIV-1 gp120 trimers.** *Nature* 2008, **455**(7209):109–113.
6. Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, Cayanan C, Maddon PJ, Koup RA, Moore JP, et al: **HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CR5.** *Nature* 1996, **381**(6584):667–673.
7. Koot M, Keet IP, Vos AH, de Goede RE, Roos MT, Coutinho RA, Miedema F, Schellekens PT, Tersmette M: **Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS.** *Ann Intern Med* 1993, **118**(9):681–688.
8. Hazenberg MD, Otto SA, Hamann D, Roos MT, Schuitemaker H, de Boer RJ, Miedema F: **Depletion of naive CD4 T cells by CXCR4-using HIV-1 variants occurs mainly through increased T-cell death and activation.** *AIDS* 2003, **17**(10):1419–1424.
9. Levine B, Sadora DL: **HIV and CXCR4 in a kiss of autophagic death.** *J Clin Invest* 2006, **116**(8):2078–2080.
10. Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR: **Change in coreceptor use correlates with disease progression in HIV-1-infected individuals.** *J Exp Med* 1997, **185**(4):621–628.
11. Abebe A, Demissie D, Goudsmit J, Brouwer M, Kuiken CL, Pollakis G, Schuitemaker H, Fontanet AL, Rinke de Wit TF: **HIV-1 subtype C syncytium- and non-syncytium-inducing phenotypes and coreceptor usage among Ethiopian patients with AIDS.** *AIDS* 1999, **13**(11):1305–1311.
12. Pollakis G, Abebe A, Kliphuis A, Chalaby M, Bakker M, Mengistu Y, Brouwer M, Goudsmit J, Schuitemaker H, Paxton WA: **Phenotypic and genotypic comparisons of CCR5- and CXCR4-tropic human immunodeficiency virus type 1 biological clones isolated from subtype C-infected individuals.** *J Virol* 2004, **78**(6):2841–2852.

8. APPENDIX C

Crous *et al.* *BMC Infectious Diseases* 2012, **12**:203
<http://www.biomedcentral.com/1471-2334/12/203>

Page 7 of 8

13. Connell BJ, Michler K, Capovilla A, Venter WD, Stevens WS, Papatathanasopoulos MA: **Emergence of X4 usage among HIV-1 subtype C: evidence for an evolving epidemic in South Africa.** *AIDS* 2008, **22**(7):896–899.
14. Kassaye S, Johnston E, McColgan B, Kantor R, Zijenah L, Katzenstein D: **Envelope coreceptor tropism, drug resistance, and viral evolution among subtype C HIV-1-infected individuals receiving nonsuppressive antiretroviral therapy.** *J Acquir Immune Defic Syndr* 2009, **50**(1):9–18.
15. Michler K, Connell BJ, Venter WD, Stevens WS, Capovilla A, Papatathanasopoulos MA: **Genotypic characterization and comparison of full-length envelope glycoproteins from South African HIV type 1 subtype C primary isolates that utilize CCR5 and/or CXCR4.** *AIDS Res Hum Retroviruses* 2008, **24**(5):743–751.
16. Cilliers T, Nhlapo J, Coetzer M, Orlovic D, Ketas T, Olson WC, Moore JP, Trkola A, Morris L: **The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C.** *J Virol* 2003, **77**(7):4449–4456.
17. Papatathanasopoulos MA, Cilliers T, Morris L, Mokili JL, Dowling W, Birk DL, McCutchan FE: **Full-length genome analysis of HIV-1 subtype C utilizing CXCR4 and intersubtype recombinants isolated in South Africa.** *AIDS Res Hum Retroviruses* 2002, **18**(12):879–886.
18. Johnston ER, Zijenah LS, Mutetwa S, Kantor R, Kittinunvorakoon C, Katzenstein DA: **High frequency of syncytium-inducing and CXCR4-tropic viruses among human immunodeficiency virus type 1 subtype C-infected patients receiving antiretroviral treatment.** *J Virol* 2003, **77**(13):7682–7688.
19. Berger EA, Doms RW, Fenyo EM, Korber BT, Littman DR, Moore JP, Sattentau QJ, Schuitemaker H, Sodroski J, Weiss RA: **A new classification for HIV-1.** *Nature* 1998, **391**(6664):240.
20. Huang W, Eshleman SH, Toma J, Franses S, Stawiski E, Paxinos EE, Whitcomb JM, Young AM, Donnell D, Mmiro F, *et al*: **Coreceptor tropism in human immunodeficiency virus type 1 subtype D: high prevalence of CXCR4 tropism and heterogeneous composition of viral populations.** *J Virol* 2007, **81**(15):7885–7893.
21. Huang W, Eshleman SH, Toma J, Stawiski E, Whitcomb JM, Jackson JB, Guay L, Musoke P, Parkin N, Petropoulos CJ: **Vertical transmission of X4-tropic and dual-tropic HIV-1 in five Ugandan mother-infant pairs.** *AIDS* 2009, **23**(14):1903–1908.
22. Poveda E, Seclen E, Gonzalez Mdel M, Garcia F, Chueca N, Aguilera A, Rodriguez JJ, Gonzalez-Lahoz J, Soriano V: **Design and validation of new genotypic tools for easy and reliable estimation of HIV tropism before using CCR5 antagonists.** *J Antimicrob Chemother* 2009, **63**(5):1006–1010.
23. Poveda E, Briz V, Quinones-Mateu M, Soriano V: **HIV tropism: diagnostic tools and implications for disease progression and treatment with entry inhibitors.** *AIDS* 2006, **20**(10):1359–1367.
24. Collins S, i-Base H: **Treatment failure and tropism changes in maraviroc trial related to previously undetected CXCR4, rather than a mutational shift from CCR5.** In *XVI International HIV Drug Resistance Workshop*. Barbados: HIV Treatment Bulletin; 2007.
25. Whitcomb JM, Huang W, Franses S, Limoli K, Toma J, Wrin T, Chappay C, Kiss LD, Paxinos EE, Petropoulos CJ: **Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism.** *Antimicrob Agents Chemother* 2007, **51**(2):566–575.
26. Prosperi MC, Bracciale L, Fabbiani M, Di Giambenedetto S, Razzolini F, Meini G, Colafigli M, Marzocchetti A, Cauda R, Zazzi M, *et al*: **Comparative determination of HIV-1 co-receptor tropism by enhanced sensitivity profile, gp120 V3-loop RNA and DNA genotyping.** *Retrovirology* 2010, **7**:56.
27. Sierra S, Kaiser R, Thielens A, Lengauer T: **Genotypic coreceptor analysis.** *Eur J Med Res* 2007, **12**(9):453–462.
28. McGovern RA, Thielens A, Mo T, Dong W, Woods CK, Chapman D, Lewis M, James I, Heera J, Valdez H, *et al*: **Population-based V3 genotypic tropism assay: a retrospective analysis using screening samples from the A4001029 and MOTIVATE studies.** *AIDS* 2010, **24**(16):2517–2525.
29. Rizzuto C, Sodroski J: **Fine definition of a conserved CCR5-binding region on the human immunodeficiency virus type 1 glycoprotein 120.** *AIDS Res Hum Retroviruses* 2000, **16**(8):741–749.
30. Rizzuto CD, Wyatt R, Hernandez-Ramos N, Sun Y, Kwong PD, Hendrickson WA, Sodroski J: **A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding.** *Science* 1998, **280**(5371):1949–1953.
31. Boyd MT, Simpson GR, Cann AJ, Johnson MA, Weiss RA: **A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism.** *J Virol* 1993, **67**(6):3649–3652.
32. Bergeron L, Sullivan N, Sodroski J: **Target cell-specific determinants of membrane fusion within the human immunodeficiency virus type 1 gp120 third variable region and gp41 amino terminus.** *J Virol* 1992, **66**(4):2389–2397.
33. Ross TM, Cullen BR: **The ability of HIV type 1 to use CCR-3 as a coreceptor is controlled by envelope V1/V2 sequences acting in conjunction with a CCR-5 tropic V3 loop.** *Proc Natl Acad Sci U S A* 1998, **95**(13):7682–7686.
34. Hoffman NG, Seillier-Moisewitsch F, Ahn J, Walker JM, Swanstrom R: **Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop.** *J Virol* 2002, **76**(8):3852–3864.
35. Nabatov AA, Pollakis G, Linnemann T, Kliphuis A, Chalaby MI, Paxton WA: **Intrapatient alterations in the human immunodeficiency virus type 1 gp120 V1V2 and V3 regions differentially modulate coreceptor usage, virus inhibition by CC/CXC chemokines, soluble CD4, and the b12 and 2.G12 monoclonal antibodies.** *J Virol* 2004, **78**(11):524–530.
36. Clevevstig P, Pramank L, Leitner T, Ehrnst A: **CCR5 use by human immunodeficiency virus type 1 is associated closely with the gp120 V3 loop N-linked glycosylation site.** *J Gen Virol* 2006, **87**(Pt 3):607–612.
37. Pollakis G, Kang S, Kliphuis A, Chalaby MI, Goudsmit J, Paxton WA: **N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization.** *J Biol Chem* 2001, **276**(16):13433–13441.
38. Polzer S, Dittmar MT, Schmitz H, Schreiber M: **The N-linked glycan g15 within the V3 loop of the HIV-1 external glycoprotein gp120 affects coreceptor usage, cellular tropism, and neutralization.** *Virology* 2002, **304**(1):70–80.
39. Fouchier RA, Groenink M, Kootstra NA, Tersmette M, Huisman HG, Miedema F, Schuitemaker H: **Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule.** *J Virol* 1992, **66**(5):3183–3187.
40. Cardozo T, Kimura T, Philpott S, Weiser B, Burger H, Zolla-Pazner S: **Structural basis for coreceptor selectivity by the HIV type 1 V3 loop.** *AIDS Res Hum Retroviruses* 2007, **23**(3):415–426.
41. Resch W, Hoffman N, Swanstrom R: **Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks.** *Virology* 2001, **288**(1):51–62.
42. Jensen MA, Li FS, Van't Wout AB, Nickle DC, Shriner D, He HX, McLaughlin S, Shankarappa R, Margolick JB, Mullins JL: **Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences.** *J Virol* 2003, **77**(24):13376–13388.
43. Jensen MA, Coetzer M, Van't Wout AB, Morris L, Mullins JK: **A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences.** *J Virol* 2006, **80**(10):4698–4704.
44. Sing T, Low AJ, Beerenwinkel N, Sander O, Cheung PK, Domingues FS, Buch J, Daumer M, Kaiser R, Lengauer T, *et al*: **Predicting HIV coreceptor usage on the basis of genetic and clinical covariates.** *Antivir Ther* 2007, **12**(7):1097–1106.
45. Pillai S, Good B, Richman D, Corbett J: **A new perspective on V3 phenotype prediction.** *AIDS Res Hum Retroviruses* 2003, **19**(2):145–149.
46. Raymond S, Delobel P, Mavignier M, Ferradini L, Cazabat M, Souyris C, Sandres-Saune K, Pasquier C, Marchou B, Massip P, *et al*: **Prediction of HIV type 1 subtype C tropism by genotypic algorithms built from subtype B viruses.** *J Acquir Immune Defic Syndr* 2010, **53**(2):167–175.
47. Garrido C, Roulet V, Chueca N, Poveda E, Aguilera A, Skrabal K, Zahonero N, Carlos S, Garcia F, Faudon JL, *et al*: **Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes.** *J Clin Microbiol* 2008, **46**(3):887–891.
48. Requejo H: **Worldwide molecular epidemiology of HIV.** *Rev Saude Publica* 2006, **40**(2):331–345.
49. Seclen E, Garrido C, Gonzalez Mdel M, Gonzalez-Lahoz J, de Mendoza C, Soriano V, Poveda E: **High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists.** *J Antimicrob Chemother* 2010, **65**(7):1486–1492.

8. APPENDIX C

Crous *et al.* *BMC Infectious Diseases* 2012, **12**:203
<http://www.biomedcentral.com/1471-2334/12/203>

Page 8 of 8

50. Mefford ME, Gory PR, Kunstman K, Wolinsky SM, Gabuzda D: **Bioinformatic prediction programs underestimate the frequency of CXCR4 usage by RSX4 HIV type 1 in brain and other tissues.** *AIDS Res Hum Retroviruses* 2008, **24**(9):1215–1220.
51. Maddison WP, Maddison DR: **MacClade**. 45th edition Sunderland: Sinauer; 1992.

doi:10.1186/1471-2334-12-203

Cite this article as: Crous *et al.*: Appraising the performance of genotyping tools in the prediction of coreceptor tropism in HIV-1 subtype C viruses. *BMC Infectious Diseases* 2012 **12**:203.



Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



8. APPENDIX C

References

- Abebe, A., Demissie, D., Goudsmit, J., Brouwer, M., Kuiken, C. L., Pollakis, G., Schuitemaker, H., Fontanet, A. L., and Rinke de Wit, T. F. (1999). HIV-1 subtype C syncytium- and non-syncytium-inducing phenotypes and coreceptor usage among Ethiopian patients with AIDS. *AIDS*, 13(11):1305–11. [14](#), [19](#), [65](#)
- Alkhatib, G. (2009). The biology of CCR5 and CXCR4. *Current Opinion HIV AIDS*, 4(2):96–103. [14](#), [15](#)
- Baba, M., Nishimura, O., Kanzaki, N., Okamoto, M., Sawada, H., Iizawa, Y., Shiraishi, M., Aramaki, Y., Okonogi, K., Y., O., Meguro, K., and Fujino, M. (1999). A small-molecule, nonpeptide CCR5 antagonist with highly potent and selective anti-HIV-1 activity. *Proc Natl Acad Sci U S A.*, 96(10):5698–703. [17](#)
- Barré-Sinoussi, F., Chermann, J., Rey, F., Nugeyre, M., Chamaret, S., Gruest, J., Dautoguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 20;220(4599):868–71. [3](#)
- Berger, E., Doms, R., Fenyo, E., Korber, B., Littman, D., Moore, J., Sattentau, Q., Schuitemaker, H., Sodroski, J., and Weiss, R. (1998). A new classification for HIV-1. *Nature*, 391(6664):240. [15](#), [16](#), [57](#)
- Berger, E., Murphy, P., and Farber, J. (1999). Chemokine receptors AS HIV-1 CORECEPTORS: Roles in Viral Entry, Tropism, and Disease. *Annu. Rev. Immunol.*, 17:657–700. [16](#)
- Bergeron, L., Sullivan, N., and Sodroski, J. (1992). Target cell-specific determinants

8. APPENDIX C

- of membrane fusion within the human immunodeficiency virus type 1 gp120 third variable region and gp41 amino terminus. *J. Virol.*, 66(4):2389–2397. [8](#), [15](#), [18](#)
- Boyd, M., Simpson, G., Cann, A., Johnson, M., and Weiss, R. (1993). A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. *J Virol.*, 67(6):3649–3652. [18](#)
- Bozek, K., Lengauer, T., Sierra, S., Kaiser, R., and Domingues, F. (2013). Analysis of Physicochemical and Structural Properties Determining HIV-1 Coreceptor Usage. *PLoS Comput Biol.*, 9(3). [57](#)
- Briggs, J., Riches, J., Glass, B., Bartonova, V., Zanetti, G., and Kräusslich, H.-G. (2009). Structure and assembly of immature HIV. *Proc Natl Acad Sci U S A.*, 106(27):11090–11095. [10](#)
- Buzon, V., Natrajan, G., Schibli, D., Campelo, F., Kozloz, M., and Weisenhorn, W. (2010). Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. *PLoS Pathog*, 6(5):e1000880. doi:10.1371/journal.ppat.1000880. [14](#), [15](#)
- Cardozo, T., Kimura, T., Philpott, S., Weiser, B., Burger, H., and Zolla-Pazner, S. (2007). Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. *AIDS Res Hum Retroviruses*, 23(3):415–26. [18](#), [28](#), [29](#)
- CDC (1981). Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men - New York City and California. *MMWR*, 4(30):306–8. [3](#)
- CDC (1982a). Current Trends Update on Acquired Immune Deficiency Syndrome (AIDS) - United States. *MMWR*, 31(37):507–508, 513–514. [3](#)

8. APPENDIX C

CDC (1982b). Persistent, generalized lymphadenopathy among homosexual males. *MMWR*, 31:249–52. [3](#)

CDC (2006). Twenty-Five Years of HIV/AIDS - United States, 1981-2006. *MMWR*, 55(21):585 –589. [3](#)

Cecilia, D., Kulkarni, S., Tripathy, S., Gangakhedkar, R., Paranjape, R., and Gadkari, D. (2000.). Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India. *Virology*, 271:253–258. [19](#)

Chan, D., Fass, D., Berger, J., and Kim, P. (1997). Core Structure of gp41 from the HIV Envelope Glycoprotein. *Cell*, 89:263–273. [12](#), [13](#), [14](#), [15](#)

Chan, D. and Kim, P. (1998). HIV Entry and Its Inhibition. *Cell*, 93:681–684. [15](#)

Chiu, T. and Davies, D. (2004). Structure and function of HIV-1 integrase. *Curr Top Med Chem.*, 4(9):965–77. [8](#)

Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W., Moore, J., Trkola, A., , and Morris, L. (2003). The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. *Journal of Virology*, 77(7):4449–4456. [15](#), [17](#), [19](#), [65](#)

Clevestig, P., Pramanik, L., Leitner, T., and Ehrnst, A. (2006). CCR5 use by human immunodeficiency virus type 1 is associated closely with the gp120 V3 loop N-linked glycosylation site. *J Gen Virol.*, 87(Pt 3):607–12. [18](#)

Coiras, M., López-Huertas, M., Pérez-Olmeda, M., and Alcamí, J. (2009). Understanding HIV-1 latency provides clues for the eradication of long-term reservoirs. *Nature Reviews Microbiology*, 7:798–812. [8](#)

8. APPENDIX C

- Collins, S. and iBase, H. (2007). Treatment failure and tropism changes in maraviroc trial related to previously undetected CXCR4, rather than a mutational shift from CCR5. In *HIV Treatment Bulletin*. XVI Intl Drug Resistance Workshop 16 Barbados 2007. [17](#)
- Connell, B., Michler, K., Capovilla, A., Venter, W., Stevens, W., and Papathanasopoulos, M. (2008). Emergence of X4 usage among HIV-1 subtype C: evidence for an evolving epidemic in South Africa. *AIDS*, 22(7):896–9. [15](#), [19](#), [35](#), [65](#)
- Connor, R., Sheridan, K., Ceradini, D., Choe, S., and Landau, N. (1997). Change in coreceptor use correlates with disease progression in HIV-1-infected individuals. *J Exp Med.*, 185(4):621–8. [17](#)
- Cormier, E. and 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. *J Virol*, 76(17):8953–7. [18](#)
- Dalglish, A., Beverley, P., Clapham, P., Crawford, D., Greaves, M., and Weiss, R. (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*, 312(5996):763–7. [13](#)
- Deutsche-AIDS-Gesellschaft (2012). Empfehlungen zur Bestimmung des HIV-1-Korezeptor-Gebrauchs: Anhang zu Therapieleitlinien. [58](#)
- Doehle, B., Chang, K., Rustagi, A., McNevin, J., McElrath, M., and Gale, M., J. Vpu mediates depletion of interferon regulatory factor 3 during HIV infection by a lysosome-dependent mechanism. *J Virol*, 86(16):8367–74. [10](#)

8. APPENDIX C

- Doms, R. and Trono, D. (2000). The plasma membrane as a combat zone in the HIV battlefield. *Genes Dev.*, 14(21):2677–88. [16](#)
- Doranz, B., Lu, Z., Rucker, J., Zhang, T., Sharron, M., Cen, Y., Wang, Z., Guo, H., Du, J., Accavitti, M., Doms, R., and Peiper, S. (1997). Two distinct CCR5 domains can mediate coreceptor usage by human immunodeficiency virus type 1. *J. Virol.*, 71(9):6305–6314. [17](#)
- Dragic, T., Litwin, V., Allaway, G., Martin, S., Huang, Y., Nagashima, K., Cayanan, C., Maddon, P., Koup, R., Moore, J., and Paxton, W. (1996). HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature*, 381:667 – 673. [13](#), [14](#)
- Duri, K., Soko, W., Gumbo, F., Kristiansen, K., Mapingure, M., Stray-Pedersen, B., and Muller, F. (2011). Genotypic analysis of human immunodeficiency virus type 1 env V3 loop sequences: bioinformatics prediction of coreceptor usage among 28 infected mother-infant pairs in a drug-naive population. *AIDS Res Hum Retroviruses*, 27(4):411–9. [66](#)
- Esbjörnsson, J., Månsson, F., Martínez-Arias, W., Vincic, E., Biague, A., da Silva, Z., Fenyö, E., Norrgren, H., and 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(23). [19](#), [65](#)
- Fouchier, R., Groenink, M., Kootstra, N., Tersmette, M., Huisman, H., Miedema, F., and Schuitemaker, H. (1992). Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J. Virol.*, 66(5):3183–3187. [17](#), [18](#), [28](#), [29](#)

8. APPENDIX C

- Freed, E. (2001). HIV-1 replication. *Somat Cell Mol Genet.*, 26(1-6):13–33. [7](#), [8](#), [10](#)
- Gallo, R., Sarin, P., Gelmann, E., Robert-Guroff, M., Richardson, E., Kalyanaraman, V., Mann, D., Sidhu, G., Stahl, R., Zolla-Pazner, S., Leibowitch, J., and Popovic, M. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):865–867. [3](#)
- Gao, F., Bailes, E., Robertson, D., Chen, Y., Rodenburg, C., Michael, S., Cummins, L., Arthur, L., Peeters, M., Shaw, G., Shar, P., and Hahn, B. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature*, 397(6718):436–41. [4](#)
- Garrido, C., Roulet, V., Chueca, N., Poveda, E., Aguilera, A., Skrabal, K., Zahonero, N., Carlos, S., Garcia, F., Faudon, J., Soriano, V., and de Mendoza, C. (2008). Evaluation of eight different bioinformatics tools to predict viral tropism in different Human Immunodeficiency Virus type 1 subtypes. *J Clin Microbiol*, 46(3):887–91. [19](#), [58](#), [59](#), [60](#), [61](#)
- Goudsmit, J. (1997). *Viral Sex; The Nature of AIDS.*, volume Pg. 51-58. Oxford University Press. New York, New York. [7](#)
- Hahn, B., Shaw, G., De Cock, K., and Sharp, P. (2000). AIDS as a Zoonosis: Scientific and Public Health Implications. *Science*, 287(5453):607–614. [4](#), [5](#)
- Han, X., An, M., Zhang, W., Zhao, B., Chu, Z., Takebe, Y., and Shang, H. (2013). Genome sequences of a novel HIV-1 circulating recombinant form CRF59_01B identified among men who have sex with men in northeastern China. *Genome Announc.*, 1(3). [7](#)
- Hazenbergh, M., Otto, S., Hamann, D., Roos, M., Schuitemaker, H., de Boer, R.,

8. APPENDIX C

- and Miedema, F. (2003). Depletion of naive CD4 T cells by CXCR4-using HIV-1 variants occurs mainly through increased T-cell death and activation. *AIDS*, 17(10):1419–24. [14](#)
- Hemelaar, J., Gouws, E., Ghys, P., and Osmanov, S. (2011). Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS*, 25:679–689. [62](#)
- Ho, D., Neumann, A., Perelson, A., Chen, W., Leonard, J., and Markowitz, M. (1995). Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature*, 373(6510):123–6. [5](#)
- Hoffman, N., Seillier-Moiseiwitsch, F., Ahn, J., Walker, J., and Swanstrom, R. (2002). Variability in the Human Immunodeficiency Virus Type 1 gp120 Env Protein Linked to Phenotype-Associated Changes in the V3 Loop. *J Virol.*, 76(8):3852–3864. [18](#)
- Huang, W., Eshleman, S., Toma, J., Fransen, S., Stawiski, E., Paxinos, E., Whitcomb, J., Young, A., Donnell, D., Mmiro, F., Musoke, P., Guay, L., Jackson, J., Parkin, N., and Petropoulos, C. (2007). Coreceptor tropism in Human Immunodeficiency Virus type 1 subtype D: high prevalence of CXCR4 tropism and heterogeneous composition of viral populations. *J Virol.*, 81(15):7885–93. [15](#), [16](#), [57](#)
- Huang, W., Eshleman, S., Toma, J., Stawiski, E., Whitcomb, J., Jackson, J., Guay, L., Musoke, P., Parkin, N., and Petropoulos, C. (2009). Vertical transmission of X4-tropic and dual-tropic HIV-1 in five Ugandan mother-infant pairs. *AIDS.*, 23(14):1903–8. [15](#), [16](#), [57](#)
- Huang, Y., Paxton, W., Wolinsky, S., Neumann, A., Zhang, L., He, T., Kang, S., Ceraadini, D., Jin, Z., Yazdanbakhsh, K., Kunstman, K., Erickson, D., Dragon, E., Lan-

8. APPENDIX C

- dau, N., Phair, J., Ho, D., and Koup, A. (1996). The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med*, 2(11):1240–1243. [17](#)
- Hummel, S., Schmidt, D., Kremeyer, B., Herrmann, B., and Oppermann, M. (2005). Detection of the CCR5-Delta32 HIV resistance gene in Bronze Age skeletons. *Genes Immun.*, 6(4):371–4. [17](#)
- Jensen, M., Coetzer, M., van 't Wout, A., Morris, L., and Mullins, J. (2006). A reliable phenotype predictor for Human Immunodeficiency Virus type 1 subtype C based on envelope V3 sequences. *J Virol*, 80(10):4698–704. [18](#), [19](#), [20](#), [25](#), [27](#), [59](#), [60](#)
- Jensen, M., Li, F., van 't Wout, A., Nickle, D., Shriner, D., He, H., McLaughlin, S., Shankarappa, R., Margolick, J., and Mullins, J. (2003). Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of Human Immunodeficiency Virus type 1 env V3 loop sequences. *J Virol*, 77(24):13376–88. [18](#), [27](#)
- Johnston, E., Zijenah, L., Mutetwa, S., Kantor, R., Kittinunvorakoon, C., and Katzenstein, D. (2003). High frequency of syncytium-inducing and CXCR4-tropic viruses among Human Immunodeficiency Virus type 1 subtype C-infected patients receiving antiretroviral treatment. *J Virol.*, 77(13):7682–7688. [15](#), [65](#)
- Kassaye, S., Johnston, E., McColgan, B., Kantor, R., Zijenah, L., and Katzenstein, D. (2009). Envelope coreceptor tropism, drug resistance, and viral evolution among subtype C HIV-1-infected individuals receiving nonsuppressive antiretroviral therapy. *J Acquir Immune Defic Syndr*, 50(1):9–18. [15](#), [65](#)
- Keele, B., Van Heuerswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M., Bibollet-Ruche, F., Chen, Y., Wain, L., Liegeois, F., Loul, S., Ngole, E., Bienvenue, Y.,

8. APPENDIX C

- Delaporte, E., Brookfield, J., Sharp, P., Shaw, G., Peeters, M., and Hahn, B. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313(5786):523–6. [4](#)
- Kiselyeva, Y., Nedellec, R., Ramos, A., Pastore, C., Margolis, L., and Mosier, D. (2007). Evolution of CXCR4-using Human Immunodeficiency Virus type 1 SF162 is associated with two unique envelope mutations. *J. Virol.*, 81(7):3657–3661. [16](#)
- Koot, M., Keet, I., Vos, A., de Goede, R., Roos, M., Coutinho, R., Miedema, F., Schellekens, P., and Tersmette, M. (1993). Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann Intern Med*, 118(9):681–8. [14](#)
- Korber, B., Foley, B., Kuiken, C., Pillai, S., and J.G, S. (1998). Numbering Positions in HIV Relative to HXB2CG. *AIDS*, page 102. [26](#)
- Kwong, P., Wyatt, R., Robinson, J., Sweet, R., Sodroski, J., and Hendrickson, W. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393:648–659. [13](#)
- Lederman, M., Penn-Nicholson, A., Cho, M., and Mosier, D. (2006). Biology of CCR5 and its role in HIV infection and treatment. *JAMA*, 296(7):815–826. [17](#)
- Lemey, P., Pybus, O., Wang, B., Saksena, N., Salemi, M., and Vandamme, A. (2003). Tracing the origin and history of the HIV-2 epidemic. *PNAS*, 100(11):6588–6592. [4](#)
- Levine, B. and Sodora, D. (2006). HIV and CXCR4 in a kiss of autophagic death. *J Clin Invest*, 116(8):2078–80. [14](#)

8. APPENDIX C

- Liu, J., Bartesaghi, A., Borgnia, M., Sapiro, G., and Subramaniam, S. (2008). Molecular architecture of native HIV-1 gp 120 trimers. *Nature*, 455(7209):109–113. [13](#), [14](#)
- Maddison, W. and Maddison, D. (1992). *MacClade: Analysis of phylogeny and character evolution*, volume Version 3. Sinauer Associates, Sunderland, Massachusetts. [26](#)
- Maddon, P., Dalglish, A., McDougal, J., Clapham, P., Weiss, R., and Axel, R. (1986). The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. *Cell*, 47(3):333–48. [13](#)
- Marlink, R., Marlink, R., Ricard, D., M’Boup, S., Kanki, P., Romet-Lemonne, J., N’Doye, I., Diop, K., Simpson, M., and Greco, F. e. a. (1994). Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science*, 265(5178):1587–1590. [4](#)
- McCormack, G., Glynn, J., Crampin, A., Sibande, F., Mulawa, D., Bliss, L., Broadbent, P., Abarca, K., Pönnighaus, J., Fine, P., and Clewley, J. (2002). Early evolution of the human immunodeficiency virus type 1 subtype C epidemic in rural Malawi. *J. Virol.*, 76(24):12890–9. [64](#)
- McDougal, J., Maddon, P., Dalglish, A., Clapham, P., Littman, D., Godfrey, M., Maddon, D., Chess, L., Weiss, R., and Axel, R. (1986). The T4 glycoprotein is a cell-surface receptor for the AIDS virus. *Cold Spring Harb Symp Quant Biol.*, 51(Pt 2):703–11. [7](#), [8](#), [12](#), [13](#)
- McGovern, R., Thielen, A., Mo, T., Dong, W., Woods, C., Chapman, D., Lewis, M., James, I., Heera, J., Valdez, H., and Harrigan, P. (2010). Population-based V3

8. APPENDIX C

- genotypic tropism assay: a retrospective analysis using screening samples from the A4001029 and MOTIVATE studies. *AIDS*, 24(16):2517–2525. [18](#)
- Mefford, M., Gorry, P., Kunstman, K., Wolinsky, S., and Gabuzda, D. (2008). Bioinformatic prediction programs underestimate the frequency of CXCR4 usage by R5X4 HIV type 1 in brain and other tissues. *AIDS Res Hum Retroviruses*, 24(9):1215–20. [58](#)
- Michael, N. (1999). Host genetic influences on HIV-1 pathogenesis. *Curr Opin Immunol*, 11:466–474. [5](#)
- Michler, K., Connell, B., Venter, W., Stevens, W., Capovilla, A., and Papatheopoulos, M. (2008). Genotypic characterization and comparison of full-length envelope glycoproteins from South African HIV type 1 subtype C primary isolates that utilize CCR5 and/or CXCR4. *AIDS Res Hum Retroviruses*, 24(5):743–5. [15](#), [65](#)
- Miller, M., Farnet, C., and Bushman, F. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.*, 71(7):5382–5390. [8](#)
- Mohammadi, P., Desfarges, S., Bartha, I., Joos, B., Zangger, N., Muñoz, M., Günthard, H., Beerenwinkel, N., Telenti, A., and Ciuffi, A. (2013). 24 Hours in the life of HIV-1 in a T cell line. *PLoS Pathog*, 9(1):e1003161. [7](#)
- Moore, J., Parren, P., and Burton, D. (2001). Genetic subtypes, humoral immunity, and Human Immunodeficiency Virus type 1 vaccine development. *J. Virol.*, 75(13):5721–5729. [7](#)
- Nabatov, A., Pollakis, G., Linnemann, T., Kliphuis, A., Chalaby, M., and Paxton, W.

8. APPENDIX C

- (2004). Inpatient alterations in the human immunodeficiency virus type 1 gp120 V1V2 and V3 regions differentially modulate coreceptor usage, virus inhibition by CC/CXC chemokines, soluble CD4, and the b12 and 2G12 monoclonal antibodies. *J Virol.*, 78(1):524–30. [18](#)
- Neogi, U., Prarthana, S., D'Souza, G., DeCosta, A., Kuttiatt, V., Ranga, U., and Shet, A. (2010). Co-receptor tropism prediction among 1045 Indian HIV-1 subtype C sequences: Therapeutic implications for India. *AIDS Research and Therapy*, 7(24). [64](#)
- Pagán, I. and Holguín, Á. (2013). Reconstructing the timing and dispersion routes of HIV-1 subtype B epidemics in the Caribbean and Central America: a phylogenetic story. *PLoS One*, 8(7):e69218. doi:10.1371/journal.pone.0069218. [7](#)
- Papathanasopoulos, M., Cilliers, T., Morris, L., Mokili, J., Dowling, W., Birx, D., and McCutchan, F. (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):879–886. [15](#), [65](#)
- Pillai, S., Good, B., Richman, D., and Corbeil, J. (2003). A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses*, 19(2):145–9. [18](#), [28](#)
- Plantier, J., Leoz, M., Dickerson, J., De Oliveira, F., Cordonnier, F., Lemée, V., Darnaud, F., Robertson, D., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nat Med*, 15(8):871–2. [5](#)
- Pollakis, G., Abebe, A., Kliphuis, A., Chalaby, M., Bakker, M., Mengistu, Y., Brouwer, M., Goudsmit, J., Schuitemaker, H., and Paxton, W. (2004). Phenotypic and genotypic comparisons of CCR5- and CXCR4-tropic human immunodeficiency virus

8. APPENDIX C

- type 1 biological clones isolated from subtype C-infected individuals. *J Virol*, 78(6):2841–52. [14](#), [19](#), [65](#)
- Pollakis, G., Kang, S., Kliphuis, A., Chalaby, M., Goudsmit, J., and Paxton, W. (2001). N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. *J Biol Chem.*, 276(16):13433–41. [13](#), [18](#)
- Polzer, S., Dittmar, M., Schmitz, H., and Schreiber, M. (2002). The N-linked Glycan g15 within the V3 Loop of the HIV-1 External Glycoprotein gp120 Affects Coreceptor Usage, Cellular Tropism, and Neutralization. *Virology*, 304(1):70–80. [13](#), [18](#)
- Posada, D. and Crandall, K. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14:817–818. [53](#)
- Poveda, E., Alcamí, J., Paredes, R., Córdoba, J., Gutiérrez, F., Llibre, J., Delgado, R., Pulido, F., Iribarren, J., García Delatoro, M., Hernández Quero, J., Moreno, S., and García, F. (2010). Genotypic determination of HIV tropism - clinical and methodological recommendations to guide the therapeutic use of CCR5 antagonists. *AIDS Rev*, 12(3):135–48. [17](#), [19](#)
- Poveda, E., Briz, V., Quinones-Mateu, M., and Soriano, V. (2006). HIV tropism: diagnostic tools and implications for disease progression and treatment with entry inhibitors. *AIDS*, 20(10):1359–67. [17](#)
- Poveda, E., Seclen, E., Gonzalez Mdel, M., Garcia, F., Chueca, N., Aguilera, A., Rodriguez, J., Gonzalez-Lahoz, J., and Soriano, V. (2009). Design and validation

8. APPENDIX C

- of new genotypic tools for easy and reliable estimation of HIV tropism before using CCR5 antagonists. *J Antimicrob Chemother*, 63(5):1006–10. [17](#), [59](#)
- Pramanik Sollerkvist, L., Gaseitsiwe, S., Mine, M., Sebetso, G., Mphoyakgosi, T., Diphoko, T., Essex, M., and Ehrnst, A. (2013). Increased CXCR4 use of HIV-1 subtype C identified by population sequencing in patients failing antiretroviral treatment compared with treatment-naïve patients in Botswana. *AIDS Res Hum Retroviruses*. [66](#)
- Preston, B. and Dougherty, J. (1996). Mechanisms of retroviral mutation. *Trends in Microbiology*, 4(16 - 21). [5](#)
- Preston, B., Poiesz, B., and Loeb, L. (1988). Fidelity of HIV-1 reverse transcriptase. *Science*, 242(4882):1168 – 1171. [5](#)
- Prosperi, M., Bracciale, L., Fabbiani, M., Di Giambenedetto, S., Razzolini, F., Meini, G., Colafigli, M., Marzocchetti, A., Cauda, R., Zazzi, M., and De Luca, A. (2010). Comparative determination of HIV-1 co-receptor tropism by Enhanced Sensitivity Trofile, gp120 V3-loop RNA and DNA genotyping. *Retrovirology*, 7:56. [18](#)
- Quiñones-Kochs, M., Buonocore, L., and Rose, J. (2002). Role of n-linked glycans in a human immunodeficiency virus envelope glycoprotein: Effects on protein function and the neutralizing antibody response. *J. Virol.*, 76(9):4199–4211. [13](#)
- Raymond, S., Delobel, P., Mavigner, M., Ferradini, L., Cazabat, M., Souyris, C., Sandres-Saune, K., Pasquier, C., Marchou, B., Massip, P., and Izopet, J. (2010). Prediction of HIV type 1 subtype C tropism by genotypic algorithms built from subtype B viruses. *J Acquir Immune Defic Syndr*, 53(2):167–75. [19](#), [20](#), [26](#), [28](#), [30](#), [59](#), [60](#), [61](#)

8. APPENDIX C

- Requejo, H. I. Z. (2006). Worldwide molecular epidemiology of HIV. *Rev. Saúde Pública*, 40(2):331–45. [19](#)
- Resch, W., Hoffman, N., and Swanstrom, R. (2001). Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, 288(1):51–62. [18](#)
- Rizzuto, C. and Sodroski, J. (2000). Fine definition of a conserved CCR5-binding region on the human immunodeficiency virus type 1 glycoprotein 120. *AIDS Res Hum Retroviruses.*, 16(8):741–9. [18](#)
- Rizzuto, C., Wyatt, R., Hernández-Ramos, N., Sun, Y., Kwong, P., Hendrickson, W., and Sodroski, J. (1998). A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science*, 280(5371):1949–53. [18](#)
- Roberts, J., Bebenek, K., and Kunkel, T. (1988). The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171–1173. [5](#)
- Robertson, D., Anderson, J., Bradac, J., Carr, J., Foley, B., Funkhouser, R., Gao, F., Hahn, B., Kalish, M., Kuiken, C., Learn, G., Leitner, T., McCutchan, F., Osmanov, S., Peeters, M., Pieniazek, D., Salminen, M., Sharp, P., Wolinsky, S., and Korber, B. (2000). HIV-1 nomenclature proposal. *Science*, 288(5463):55–6. [5](#)
- Robertson, D., Hahn, B., and Sharp, P. (1995). Recombination in AIDS viruses. *J Mol Evol.*, 40(3):249–59. [4](#)
- Ross, T. and Cullen, B. (1998). The ability of HIV type 1 to use CCR-3 as a coreceptor is controlled by envelope V1/V2 sequences acting in conjunction with a CCR-5 tropic V3 loop. *Proc Natl Acad Sci U S A.*, 95(13):7682–6. [8](#), [12](#), [18](#)

8. APPENDIX C

- Rotta, I. and Almeida, S. (2011). Genotypical diversity of HIV clades and central nervous system impairment. *Arq Neuropsiquiatr*, 69(6):964–972. [7](#)
- Salminen, M., Johansson, B., Sonnerborg, A., Ayehunnie, S., Gotte, D., Leinikki, P., Burke, D., and McCutchan, F. (1996). Full-length sequence of an Ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. *AIDS Res Hum Retroviruses*, 12:1329–39. [64](#)
- Sander, O., Sing, T., Sommer, I., Low, A., Cheung, P., Harrigan, P., Lengauer, T., and Domingues, F. (2007). Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput Biol*, 3(3):e58. [29](#)
- Santiago, M., Rodenburg, C., Kamenya, S., Bibollet-Ruche, F., Gao, F., Bailes, E., Meleth, S., Soong, S., Kilby, J., Moldoveanu, Z., Fahey, B., Muller, M., Ayoub, A., Nerrienet, E., McClure, H., Heeney, J., Pusey, A., Collins, D., Boesch, C., Wrangham, R., Goodall, J., Sharp, P., Shaw, G., and Hahn, B. (2002). SIVcpz in wild chimpanzees. *Science*, 295(5554):465. [4](#)
- Sattentau, Q. and Moore, J. (1991). Conformational changes induced in the human immunodeficiency virus envelope glycoprotein by soluble CD4 binding. *J Exp Med.*, 174(2):407–15. [8](#), [13](#), [14](#), [15](#)
- Seclen, E., Garrido, C., Gonzalez Mdel, M., Gonzalez-Lahoz, J., de Mendoza, C., Soriano, V., and Poveda, E. (2010). High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists. *J Antimicrob Chemother*, 65(7):1486–92. [59](#), [61](#)
- Sharp, P. and Hahn, B. (2011). Origins of HIV and the AIDS Pandemic. *Cold Spring Harb Perspect Med*, 1(1). [4](#)

8. APPENDIX C

- Sierra, S., Kaiser, R., Thielen, A., and Lengauer, T. (2007). Genotypic coreceptor analysis. *Eur J Med Res*, 12(9):453–62. [18](#)
- Sing, T., Low, A., Beerenwinkel, N., Sander, O., Cheung, P., Domingues, F., Büch, J., Däumer, M., Kaiser, R., Lengauer, T., and Harrigan, P. (2007a). Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther.*, 12(7):1097–106. [18](#)
- Sing, T., Low, A. J., Beerenwinkel, N., Sander, O., Cheung, P. K., Domingues, F. S., Buch, J., Daumer, M., Kaiser, R., Lengauer, T., and Harrigan, P. R. (2007b). Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther.*, 12(7):1097–106. [26](#), [27](#), [38](#), [56](#), [57](#)
- Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690. [36](#)
- Starcich, B., Hahn, B., Shaw, G., McNeely, P., Modrow, S., Wolf, H., Parks, E., Parks, W., Josephs, S., Gallo, R., and et al (1986). Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell*, 45(5):637–48. [13](#)
- Subbarao, S. and Schochetman, G. (1996). Genetic variability of HIV-1. *AIDS*, 10(Suppl A:S13-23). [5](#)
- Tebit, D. and Arts, E. (2011). Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect Dis*, 11:45 – 56. [4](#)
- Temin, H. (1993). Retrovirus variation and reverse transcription: abnormal strand

8. APPENDIX C

transfers result in retrovirus genetic variation. *Proc. Natl. Acad. Sci. U. S. A.*, 90:6900–6903. [5](#), [8](#)

UNAIDS (2013). Unaided 2012 global report. [1](#), [3](#), [7](#)

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., and Brennan, C. (2011). Confirmation of Putative HIV-1 Group P in Cameroon. *J Virol.*, 85(3):1403–140. [5](#)

Vandekerckhove, L., Verhofstede, C., Demecheleer, E., De Wit, S., Florence, E., Fransen, K., Moutschen, M., Mostmans, W., Kabeya, K., Mackie, N., Plum, J., Vaira, D., Van Baelen, K., Vandenbroucke, I., Van Eygen, V., Van Marck, H., Vogelaers, D., Geretti, A., and Stuyver, L. (2011a). Comparison of phenotypic and genotypic tropism determination in triple-class-experienced HIV patients eligible for maraviroc treatment. *J Antimicrob. Chemother.*, 66(2):265–272. [18](#)

Vandekerckhove, L., Wensing, A., Kaiser, R., Brun-Vezinet, F., Clotet, B., De Luca, A., Dressler, S., Garcia, F., Geretti, A., Klimkait, T., Korn, K., Masquelier, B., Perno, C., Schapiro, J., Soriano, V., Sönnnerborg, A., Vandamme, À., Verhofstede, C., Walter, H., Zazzi, M., and Boucher, C. (2010). Consensus statement of the european guidelines on clinical management of hiv-1 tropism testing. *J Internat. AIDS Soc.*, 13(Sup 4):O7. [18](#)

Vandekerckhove, L., Wensing, A., Kaiser, R., Brun-Vezinet, F., Clotet, B., De Luca, A., Dressler, S., Garcia, F., Geretti, A., Klimkait, T., Korn, K., Masquelier, B., Perno, C., Schapiro, J., Soriano, V., Sonnerborg, A., Vandamme, A., Verhofstede, C., Walter, H., Zazzi, M., and Boucher, C. (2011b). European guidelines on the

8. APPENDIX C

clinical management of HIV-1 tropism testing. *Lancet Infect Dis*, 11(5):394–407.

[18](#)

Verhofstede, C., Nijhuis, M., and Vandekerckhove, L. (2012). Correlation of coreceptor usage and disease progression. *Curr Opin HIV AIDS*, 7(5):432–9. [17](#)

Vidal, N., Peeters, M., Mulanga-Kabeya, C., Nzilambi, N., Robertson, D., Ilunga, W., Sema, H., Tshimanga, K., Bongo, B., and Delaporte, E. (2000). Unprecedented degree of Human Immunodeficiency Virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.*, 74(22):10498–10507. [5](#)

Ward, M., Lycett, S., Kalish, M., Rambaut, A., and Leigh Brown, A. (2013). Estimating the Rate of Intersubtype Recombination in Early HIV-1 Group M Strains. *J. Virol*, 87(4):1967–1973. [5](#)

Weber, J. (2001). The pathogenesis of HIV-1 infection. *Br Med Bull*, 58:61–72. Weber, J Review England British medical bulletin Br Med Bull. 2001;58:61-72. [8](#), [10](#), [12](#)

Wei, X., Ghosh, S., Taylor, M., Johnson, V., Emini, E., Deutsch, P., Lifson, J., Bonhoeffer, S., Nowak, M., and Hahn, B. e. a. (1995). Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, 373(6510):117–22. [5](#), [10](#), [12](#)

Weissenhorn, W., Dessen, A., Harrison, S., Skehel, J., and Wiley, D. (1997). Atomic structure of the ectodomain from HIV-1 gp41. *Nature*, 387:426 – 430. [15](#)

Wertheim, J. and Worobey, M. (2009). Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol.*, 5(5). [4](#)

8. APPENDIX C

Westby, M. and van der Ryst, E. (2005). CCR5 antagonists: host-targeted antivirals for the treatment of HIV infection. *Antiviral Chemistry Chemotherapy*, 16(339–354).

[17](#)

Whitcomb, J., Huang, W., Fransen, S., Limoli, K., Toma, J., Wrin, T., Chappey, C., Kiss, L., Paxinos, E., and Petropoulos, C. (2007). Development and characterization of a novel single-cycle recombinant-virus assay to determine Human Immunodeficiency Virus type 1 coreceptor tropism. *Antimicrob Agents Chemother*, 51(2):566–

575. [17](#)

Wu, Y. and March, J. (2003). Gene transcription in HIV infection. *Microbes Infect.*, 5(11):1023–7. [8](#), [10](#)

Wyatt, R., Kwong, P., Desjardins, E., Sweet, R., Robinson, J., Hendrickson, W., and Sodroski, J. (1998). The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature*, 393:705–711. [13](#)

Zhang, L., Hi, T., Huang, Y., Chen, Z., Guo, Y., Wu, S., Kuntsman, K., Brown, R., Phair, J., Neumann, A., Ho, D., and Wolinsky, S. (1998). Chemokine coreceptor usage by diverse primary isolates of Human Immunodeficiency Virus type 1. *J. Virol.*, 72(11):9307–9312. [17](#)