PHYLOGENETIC ANALYSIS AND MOLECULAR EPIDEMIOLOGICAL CHARACTERIZATION OF HIV INFECTIONS DIAGNOSED IN SOUTH AUSTRALIA

Karen Gail Hawke, B Behav Sc, BSc (Hons) Medicine

Flinders University School of Medicine

Faculty of Health Science

Adelaide, South Australia

A thesis submitted to Flinders University of South Australia for the degree of

Doctor of Philosophy

24 November 2015

TABLE OF CONTENTS

LIST OF FIGURES

TABLES

TABLE OF ABBREVIATIONS

°C – degrees Celsius

µL – microliters

3TC – Lamivudine

A – Adenine

ABC – Abacavir

AIDS – Acquired Immunodeficiency Syndrome

ART – Antiretroviral Treatment

ARV – Antiretroviral

ATV – Atazanavir

BEAST – Bayesian Evolutionary Analysis Sampling Trees

BLAST – Basic Local Alignment Search Tool

bp – base pairs

C – Cytosine

CCR5 – Chemokine (C-C motif) receptor 5

CD4 – cluster of differentiation 4

CRF – Circulating recombinant form

d4T – Stavudine

ddI – Didanosine

DNA – Deoxyribonucleic acid

DRV – Darunavir

EFV – Efavirenz

env – envelope gene

ETR – Etravirine

F – Forward Primer

FPV – Fosamprenavir

FTC – Emtricitabine

g – Gravitational constant

G – Guanine

gp – glycoprotein

GTR – General Time Reversible Model

HIV – Human Immunodeficiency Virus

HIV-1 – Human Immunodeficiency Virus type 1

HIV-2 – Human Immunodeficiency Virus type 2

IDV – Indinavir

Indels – Insertions and Deletions

jpHMM – jumping profile Hidden Markov Model

LANL BLAST – Los Alamos National Laboratory

LMIC – Low to Middle Income Country

LPV – Lopinavir

M – Molecular marker

MEGA – Molecular Evolutionary Genetics Analysis

$MgCl_2$ – Magnesium chloride

mM – millimoles

mRNA – messenger RNA

MSM – Men who have Sex with Men

MTCT – Mother to Child Transmission

NCBI – National Center for BioInformatics

NFV – Nelfinavir

NJ – Neighbor Joining

nm – nanometers

NNI – Nearest Neighbor Interchange

NNRTI – Non-Nucleoside Reverse Transcriptase Inhibitor

NRTI – Nucleoside Reverse Transcriptase Inhibitor

NVP – Nevirapine

PAUP – Phylogenetic analysis using Parsimony

PCR – Polymerase Chain Reaction

PHYLIP – PHYLogeny Inference Package

PI – Protease inhibitor

pol – polymerase gene

PR – Protease

R – Reverse Primer

RDP – Recombination Detection Program

rev – regulator gene

RIP – Recombination Identification Program

RNA – Ribonucleic acid

RPV – Rilpivirine

RT – Reverse Transcriptase

SCUEAL – Subtype Classification Using Evolutionary ALgorithms

SIV – Simian Immunodeficiency Virus

SHIV – Simian-Human Immunodeficiency Virus

SQV – Saquinavir

T – Thymine

T_A – Annealing temperature

tat – transactivation gene

TDF – Tenofovir

T_M – Melting temperature

TPV – Tipranavir

UK – United Kingdom

UNAIDS – Joint United Nations Programme on HIV and AIDS

UPGMA – Unweighted Pair Group Method with Arithmetic Means

URF – Unique Recombinant Form

US – United States

USA – United States of America

UV – Ultraviolet light

vif – Viral infectivity factor

vpr – viral protein R

vpu – viral protein U

WHO – World Health Organization

ZDV – Zidovudine

SUMMARY ABSTRACT

The global diversity of circulating HIV-1 subtypes and circulating recombinant forms (CRFs) is growing, including an increasing number of new complex variants. This has serious implications for effective vaccine and treatment strategies worldwide, and may affect sensitivity of diagnostic and viral load assays. In Australia, subtype B has historically been dominant among men who have sex with other men (MSM) but in recent years new non-B infections have appeared, predominantly via heterosexual contact, which follows a global trend. To date, there have been two studies (not including the one resulting from this PhD) published on this change in subtype distribution in Australia, noting a significant increase in the proportion of CRFs and non-B subtypes circulating across Victoria (2012) and Western Australia (2015).

To understand epidemics globally it is also important to monitor the transmission of drug resistant strains, known as transmitted drug resistance (TDR), which is rising within the adult population globally. In 2012, UNAIDS assessed global TDR prevalence at 3.1% and likely to increase. TDR can reduce efficacy of antiretroviral therapy and while subtypes display similar drug sensitivity, some may have greater propensity to develop certain mutations. In 2009, an Australian study reported a TDR prevalence of 16% in Victoria.

Since 2000, subtype and drug resistance data have been collected as part of routine genotypic resistance testing in South Australia. This PhD examined all newly diagnosed, genotyped HIV cases in South Australia between 2000 and 2013. Initial genotypic drug resistance testing was based on a 1098bp region of the *Pol* gene, encompassing the protease (PR) and reverse transcriptase (RT) genes. The Stanford calibrated population resistance (CPR) online subtyping tool was used to interpret subtype and drug resistance mutations (Chapter Four and Five), and the 2009 WHO list of surveillance mutations was used to determine TDR. A 530bp region spanning gp41 of the *Env* gene was then sequenced

(Chapter Six), and analysis of the sequenced *pol* and *env* data was conducted using Maximum Likelihood. Online subtyping tools (Stanford CPR, jumping profile Hidden Markov Model (jpHMM), REGA, Subtype Classification Using Evolutionary ALgorithms (SCUEAL), National Center for Biotechnology Information (NCBI), and COntext-based Modeling for Expeditious Typing (COMET)) were also used to determine subtype (Chapter Seven).

Molecular epidemiological analysis of the HIV *pol* gene identified an evolving epidemic in South Australia, from predominantly subtype B infection among the MSM community or Australian-born heterosexual population, to an increasing proportion of non-B infections imported from overseas by Australian travelers, or overseas-born people. These non-B infections include an increasing number of complex recombinants that are predominantly acquired in high prevalence countries where multiple strains co-circulate.

While the prevalence of TDR decreased over time, the overall rate in the HIV infected population was high, largely due to the forward transmission of K103N amongst MSM infected with subtype B infection. There was also evidence of a moderate but declining rate of Non-Nucleoside Reverse Transcriptase Inhibitor (NNRTI) and Nucleoside Reverse Transcriptase Inhibitor (NRTI) resistance in the non-B cohort who were diagnosed within the last 10 years. This resistance corresponds with the time period NNRTI and NRTI treatments were introduced, then improved, overseas.

Phylogeny and rapid online subtyping tool analysis of the *pol* and *env* genes found a significantly greater diversity of circulating HIV variants than previously identified using the Stanford CPR tool to assign the *pol* region alone, including the identification of possible unique recombinant forms. Using two gene regions also identified a larger number of possible transmission events, demonstrating the complexity of this continually evolving virus. Transmission events were predominantly between male/female heterosexual pairings

carrying non-B infections acquired overseas, with some evidence of subtype B transmission events occurring between male only pairs and male/female pairs, and a number of larger male subtype B clusters predominantly acquired within Australia.

These studies were the first in Australia to link genetic and routine surveillance data to provide an in-depth characterisation of the changing HIV epidemic in Australia over the last 14 years. The findings greatly improve our knowledge of HIV subtype distribution and transmission dynamics in South Australia. They also demonstrate that analysis of multiple gene regions using modern phylogenetic methods provides additional valuable information about the HIV epidemic structure within a region of interest. In addition, we now have an enhanced understanding of treatment failure and treatment adherence rates of current first-line regimens and how these differ between geographic origin and subtype.

It is crucial to conduct accurate and detailed surveillance in order to identify factors and molecular mechanisms that affect transmission, replication, and resistance, improve research efforts into strains widely circulating in high prevalence areas, and target subtype-specific prevention and treatment options for at-risk populations.

Despite decreasing rates, the number of TDR strains circulating in the MSM population in Australia remain a concern, which highlights the need for continued surveillance, education and early diagnosis. There is also a steady influx of people migrating from high prevalence countries and a growing proportion of these are now being diagnosed with HIV, including some with drug resistant strains.

As ART availability continues to increase globally, it is also important to ensure early diagnosis, pre-treatment drug resistance testing, effective treatment regimens and ongoing surveillance are available worldwide. It is imperative that efforts are targeted toward increasing and expanding HIV testing, not only to identify infection early but also to test for transmitted resistance before treatment commences. Resistance patterns in non-B persons

infected overseas may influence treatment choice and viral suppression beyond our current understanding of the historical subtype B infection circulating in Australia, and these cases should be monitored closely.

In addition to the change in geographic segregation of HIV subtypes and CRFs, the traditional risk-group segregation is also becoming less distinct, with growing evidence of crossover between the MSM, Intravenous Drug Use (IDU) and heterosexual populations. This combined with the increasing genetic diversity including complex and unique variants, highlights the need for careful monitoring of new infections, and the immediate need for non-B subtype research in Australia.

Despite evidence that shows social structures have far greater impact on the HIV epidemic than behavioral or treatment strategies, integration and acceptance of social strategies into prevention and control efforts has been very slow, and the prevention focus remains on modifying individual behavior and increasing treatment. Future research should combine genetic and epidemiological data, and use it to create social strategies that identify and counteract issues that impact on individual and societal behavior, such as socioeconomic disadvantage and geographical region, political and cultural systems and gender inequity, which all have considerable effect on the incidence of HIV.

In 2014 ethics approval was given for a national project named the *Australian Molecular Epidemiology Network* to commence. This project is the result of collaboration between South Australia, Western Australia, Victoria, New South Wales and Queensland, and is the first national study in Australia to collect data from all newly diagnosed HIV cases between 2005 and 2018 and phylogenetically analyze sequences to assess subtype distribution, transmission patterns both within each state and across state borders, and to identify the prevalence of unique variant forms of HIV that may be circulating.

DECLARATION OF ORIGINALITY

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

ACKNOWLEDGMENTS

I would like to extend my deepest gratitude to the following people and institutions, without which the completion of this thesis would not have been possible:

Rodney Ratcliff – Words cannot express the thanks I have for your time, patience, integrity and positive outlook. A thousand times thank you, you are a gem of a human being and have taught me what it is to respect and believe in myself.

Paul Ward – Your unwavering belief in me, your incredible resilience and wisdom, have all carried me not just through this PhD, but continue to guide me through life. I'm so glad I met you.

Russell Waddell, David Gordon and John Kaldor – between you I'm fortunate to have been under your guidance from everything to statistical genius (Russell) attention to detail and rigorous standards (David) and brilliant critical thinking (John). A special thank you to John and Russell for encouraging me to stay in Adelaide and look at the wealth of data waiting to be explored.

My partner Daniel – for the support, the patience, the belief in me. For trying to understand the beast that is a PhD, and for encouraging me to continue when things got hard.

My 8 year old son Truman – for asking me if I'll be at your PhD graduation one day - one sentence that made all of this worthwhile and let me know I did the right thing.

My mentors, Kay and Bob – you started this journey with me, and haven't left my side since. You inspire me constantly, thank you for finding me. So much love and respect to you both.

My family and friends – for always sharing in my successes, and supporting me through my challenges. I'm blessed to be surrounded by such wonderful people. ☺

Finally, this thesis is dedicated to **Suzy (12 December 2013)** and **Chris (3 September 2014),** my shining lights.

Suzy, you've been there for so many milestones in my lifetime, I so wish you were here to hold my hand as I reach this one.

Chris, the completion of this PhD is bittersweet for me. For so long my study and your illness existed side by side. I always thought we would conquer both… maybe in a way we did. It was an absolute privilege to know you all those precious years my friend.

I love you both, may you rest in peace.

PUBLICATIONS

Hawke, KG; Waddell, RG; Gordon, DL; Ratcliff, RM; Ward, PR; Kaldor, JM. 2013. HIV non-B subtype distribution: emerging trends and risk factors for imported and local infections newly diagnosed in South Australia. AIDS Res. Hum. Retroviruses. **29**:311–317

CHAPTER 1: INTRODUCTION

The human immunodeficiency virus (HIV) is part of the family of retroviruses that integrate themselves into the genomes of host cells. Almost one-tenth of the human genome is made up of retrovirus DNA, predominantly remnants of previous retroviruses that merged with human DNA over the course of human evolution.[1]

HIV exists in two types, HIV-1 and HIV-2.[2] HIV-1 is the most common and widely distributed,[3] accounting for over 99% of HIV infections worldwide.[4] HIV-1 is further classified into distinct lineages, Main (M), New (N), Outlier (O) and Putative (P), reflecting four separate introductions of simian immunodeficiency viruses into humans,[3,5,6] and there are eight distinct lineages of HIV-2.[7] Molecular and phylogenetic analyses of HIV have led to discoveries about how HIV evolves over time and by geographic region. HIV-1 has spread rapidly worldwide, while HIV-2 has remained mostly confined to Africa. Over 99% of the global HIV epidemic is attributable to the M lineage[4] and there are at least nine phylogenetically distinct subtypes, A–D, F–H, J and K.[8,9] HIV-1 naturally mutates at an extremely fast rate, about one million times faster than eukaryote DNA.[2] Selection pressure because of antiretroviral treatment (ART) also affects the rate of mutation and leads to treatment failure and TDR. When combined with the length of time a strain has been established in a geographic area or risk group, this has led to the evolution of many circulating strains, including 75 circulating recombinant forms (CRFs) and a large number of transmitted drug resistant mutations.[2,10–13]

Historically, HIV-1 subtypes and drug resistant strains have been associated with specific parts of the world and with particular modes of transmission.[10,14–20] However, the distribution of subtypes and CRFs is becoming more heterogeneous globally because of population mobility and sexual/intravenous drug contact between different population groups.[21,22] TDR has also increased in low- and middle-income countries (LMICs) because

of the increased use of antiretroviral therapy (ART) and associated issues that come with resource-poor countries. In Australia, immigration has led to importation of an increasing number of genetically variable HIV strains with possible subsequent effects on treatment and HIV diagnosis. Further epidemiological analysis is needed to assess patterns, causes and effects of HIV infection, while phylogenetic analysis (hereafter referred to as phy) can elucidate how biological sequences are genetically related and provide estimates of hypothetical common ancestors.

Phylogeny and current rapid online subtyping programs are valuable but under-used tools in Australia. Using these resources for accurate identification of viral variants and assessment of transmission patterns, then linking these data with epidemiological information in South Australia, will aid understanding about the course of the HIV global epidemic and will greatly improve knowledge of non-B subtype infections circulating in Australia.[7] Accurate reporting of TDR in newly diagnosed individuals will ensure greater understanding of current and historical treatment failure overseas and locally, and identify forward transmission of resistant virus to assist with creating targeted strategies for prevention or intervention.

In the following chapter, the HIV genome, history of the HIV global epidemic, genetic diversity, drug resistance, phylogeny and the use of online subtyping tools will be reviewed.

CHAPTER 2: LITERATURE REVIEW.

Part 1

Part 2

Part 3

Part 1

2.1 Retroviruses and viral diversity

The human immunodeficiency virus (HIV) is part of the family of retroviruses. Unlike other lentiviruses that are DNA viruses (DNA genome), retroviruses are RNA viruses transcribed into DNA by the enzyme reverse transcriptase, then into mRNA and then viral proteins. Retroviruses share three common genes: *Gag* creates the inner shell that houses the viral genes, *Env* creates the surface environment so the virus can attach to host cells, and *Pol* makes three enzymes that assist with replication and insertion of the viral genes into the host DNA and maturation of viral proteins.[1] A feature of retroviruses is that during their replication, the virus integrates itself into the genomes of certain host cells. Retrovirus DNA comprises 8% of the human genome, and is predominantly remnants of previous retroviruses that merged with human DNA over the course of human evolution.[1]

Nucleotides are the building blocks of DNA. Sequencing is the identification of the order of nucleotides for a particular gene. There are four types of nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C).[2] The genome of all living organisms is made up of a combination of these four building blocks. Homology refers to the similarity between sequences, while divergence represents the difference. For example, if 98% of a given nucleotide sequence between two isolates is the same, they are 98% similar or 2% divergent. The degree to which isolates are divergent will determine whether they are classed as subtypes of species or distinct species.

Like other lentiviruses, HIV infection does not occur as a single wild-type virus, but rather as a viral population of homogeneous strains known as quasispecies.[23,24] This is because of the error-prone nature of the reverse transcriptase enzyme during viral replication combined with the very high number of virus particles created.[23,25] RNA viruses have an extraordinary capacity to change the cells they infect and replicate in (cell tropism) and to

evade host immune responses and external antivirals.[26] Many nucleotide substitutions occur during viral replication of the initial infecting strain and in response to host immune and treatment pressures, which leads to quasispecies.[23,24] The intrahost variation of circulating quasispecies can exceed 5%, and the frequency of different quasispecies fluctuates. Proviral DNA also stores an archive of the original HIV strains within a host and, in the absence of selection pressure, these strains can reappear and become the dominant circulating form. Recently, studies have focused on examining the proportion of infectious strains passing from the infected host to an infected recipient. Data shows that a large proportion (as high as 80% in some cases) of newly diagnosed individuals are infected by a single virus strain,[27–31] which is called a genetic bottleneck.[32] However this restricted transmission of viral strains is dependent on the route of transmission, approximately 20% of productive infections are caused by a multiple virus strains in heterosexual transmission, but a greater viral diversity (40%) passes from host to recipient in men who have sex with men (MSM) and Intravenous drug users (IDU; 60%). One of the reasons for this bottleneck in heterosexual transmission is thought to be due to the more protective vaginal mucosal barrier.[27–31,33]

In addition to the viral diversity created by nucleotide substitutions, HIV strains are prone to recombine with one another, and large nucleotide insertions and deletions (indels) occur frequently which also lead to further divergent strains.[23] Recombination occurs by strand switching, which occurs when RNA transcripts from two separate proviruses form part of a HIV virion. When this virion enters a host cell, the RT jumps between the RNA transcripts, and the result is a recombinant retroviral DNA sequence.[34] Any subsequent virions produced will be this recombinant type.[34] Recombination can occur through concomitant infection (co-infection), where a person is infected with multiple strains of HIV at the time of initial infection or shortly thereafter, especially in populations where multiple subtypes circulate.[2,35] It can also occur from sequential infection (super-infection), where a person is

initially infected with one HIV strain then infected later with another HIV strain during a different exposure. This is especially likely for groups engaging in high-risk behaviors such as having multiple sexual partners or intravenous drug use.[36] Recombination can also occur between other recombinant viruses. The high rate of genetic diversity mixed with risky behavior and a large number of circulating infections will in turn determine the ongoing rate of co- and super-infections. It presents a challenge with respect to prevention and treatment strategies, and for development of a vaccine and cure.

2.2 Simian Immunodeficiency Virus (SIV)

Initially, HIV diversity was associated with multiple cross-species transmission events of Simian Immunodeficiency Virus (SIV) from non-human primates (NHPs) to humans. Soon after the discovery of HIV-1 in 1983, scientists began suspecting that human AIDS was of simian origin. The first isolate of simian immunodeficiency virus (SIVmac) was extracted from an immune deficient rhesus macaque (*Macaca mulatta*) that displayed clinical symptoms similar to AIDS.[37] In 1985, antibodies to SIV were discovered in individuals from Senegal, West Africa, suggesting that there may be another type of human retrovirus. In 1986, HIV-2 was isolated and characterized from West African patients living in France.[38]

It has now been clearly demonstrated that transmission of SIV from NHPs to humans has resulted in human HIV infection. SIV has infected at least 45 species of NHPs over thousands of years, and over time these primates have become natural hosts to species-specific infection.[7,39] SIV infections in these natural hosts can show high levels of viral replication but remain typically non-pathogenic. While SIV is generally host species-specific, several strains have infected humans, most probably from NHP blood entering human hosts through cuts and wounds during preparation of NHP 'bush meat'. HIV-1 in humans is most closely related to SIVcpz from chimpanzees (*Pan troglodytes troglodytes*)

and SIVgor isolated from gorillas. HIV-2 is most closely related to SIV_{smm} from sooty mangabeys (*Cercocebus atys*).[7] Cross-species transmission of SIV from chimpanzees (SIV_{cpz}), sooty mangabeys (SIV_{smm}) and gorillas (SIV_{gor}) to humans led to the two HIV strains, HIV-1, and HIV-2.[39]

2.3 The HIV genome

The HIV genome is approximately 9719 base pairs in length and comprised of five main genes, five accessory genes, 15 proteins and seven structural elements.[40] Figure 1 shows each gene in detail. The *Gag* gene encodes the capsid proteins and associates with the plasma membrane to assist virus assembly. *Tat* and *Rev* are both regulatory genes for HIV-1 gene expression, while *Vif*, *Vpr*, *Vpu* and *Nef* are multifunctional accessory genes that assist in the viral replication by inhibiting host antiviral factors.[40,41] The *pol* and *env* genes are two often used in phy and these are described in detail in the next section.

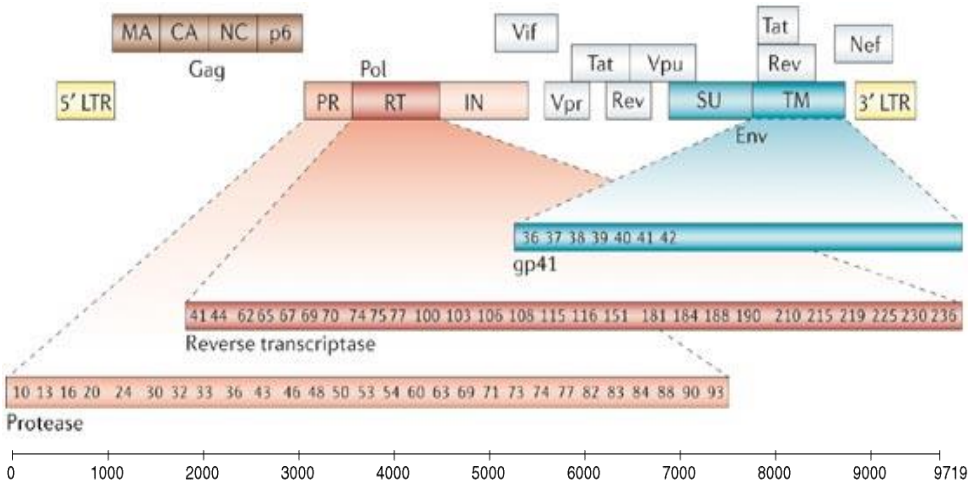


Figure 1. The HIV genome. (Lengauer & Sing 2006)[42] The zoomed areas shown below the genome denote locations of known drug resistance mutations.

2.3.1 The life cycle of HIV-1

HIV replicates inside host cells, specifically inside human immune system cells. It uses the host cell's own machinery to replicate and integrate itself into the host DNA.[43] Figure 2 shows the life cycle of HIV.[19]

An HIV particle attaches to a host cell by its envelope proteins, gp120 (surface protein) and gp41 (anchor protein for gp120). *Env* encodes the enzyme gp160, and Furin (a host cell protease) cleaves gp160 to form gp120 and gp41. When gp120 binds to a host CD4 receptor it causes gp41 to become exposed from the viral envelope and assists in cell fusion.[45]

There are two protein receptors on the host cell membrane, the initial binding receptor, CD4, and a co-receptor that can differ by cell type. The co-receptor most commonly used is CCR5, which is found on dendritic cells, macrophages, regulatory T cells, and, importantly, memory and effector helper T cells. The other co-receptor, CXCR4, is used later in infection. The virion's envelope fuses with the host cell membrane and the contents of the virion enter the cell: two copies of the single stranded RNA HIV genome and three proteins encoded by *pol*,[40] reverse transcriptase (RT) that transcribes virus RNA into virus DNA, integrase (IN) that splices virus DNA into the host cell DNA genome, and protease (PR) that plays a role in preparing new viral proteins to be released from the cell.[43]

The virus RT uses the viral RNA template to synthesize viral DNA, then the virus IN splices this HIV DNA into the host genome. This new HIV DNA is transcribed to HIV mRNA by the host cell's own RNA polymerase, and then the HIV mRNA is translated into HIV precursor proteins by the host cell's ribosomes. The virus PR released by the virion into the cell then cleaves the precursor proteins into mature viral proteins and these assemble into virions within the cytoplasm. The virions bud from the host cell membrane out to the extracellular space, with the virus envelope being formed by the host cell

membrane. These new virions will mature and infect new cells or be transmitted to a new host.[43]

It is important to note that HIV uses the host cell's own machinery to replicate. This makes treatment and eradication of the virus very difficult, because any drug that interferes with the HIV cycle will also affect the host cell's enzymatic functions and thus have significant side effects.[43]

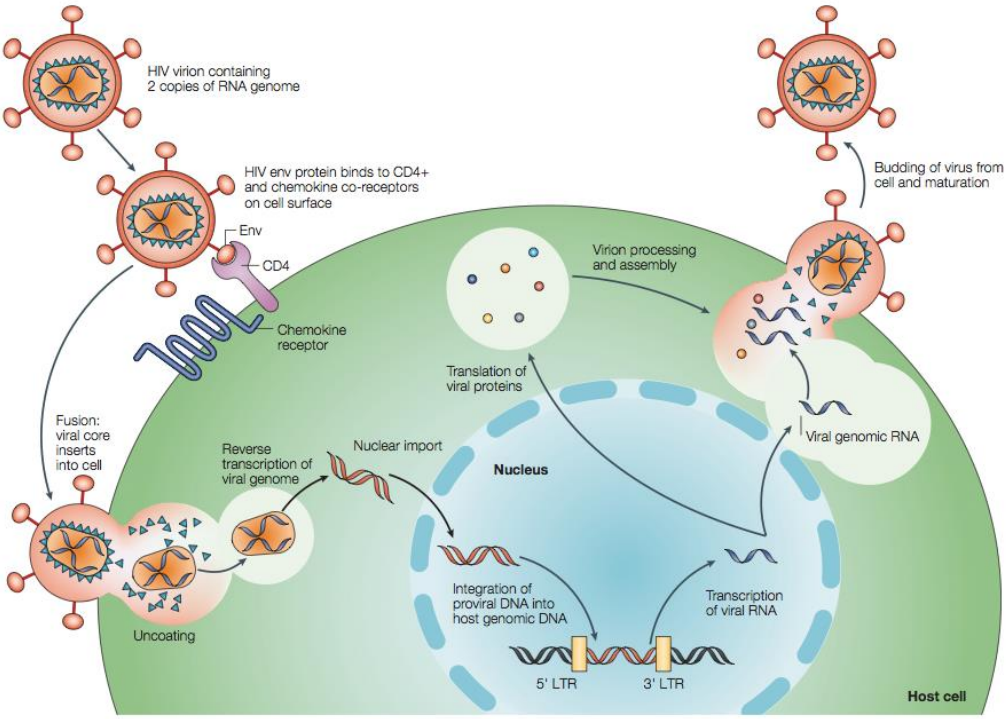


Figure 2. The life cycle of a HIV virion (Rambaut et al 2004).[44]

2.3.2 Phases of Infection

There are three main phases of infection. Phase 1 is the acute stage, where HIV virions enter the host and begin to replicate rapidly in the host cells. More and more virions accumulate in the blood, and the number of CD4+ T cells falls as the HIV infection destroys them as part of the replication process. At this stage, the T cells that are most

affected are the memory helper cells in lymphoid tissue, especially in the gut. The gut is very vulnerable to pathogens, so the loss of T cells causes major damage to the immune response. The acute phase ends when the viral load in the blood begins to stabilize and decline. This may be because there is depletion of CD4-containing cells to invade, and/or an increase in mobilization of killer T cells which target cells containing the virus.[43] During this time, the host experiences a mild to severe influenza-like illness, frequently not attributed to HIV infection, and this can determine how quickly after infection diagnosis occurs.

In Phase 2, the virus enters the chronic or asymptomatic stage. During this time, the CD4+ T cell count begins to recover as the immune system continues its efforts to fight the virus. During the chronic phase, the immune system remains highly activated, which is both helpful and harmful. Killer T cells continue to destroy infected host cells but a steady supply of CD4+ helper T cells is also being produced, which gives virions a home in which to replicate. Because virions are still present, the helper cells are also being constantly stimulated to divide into memory and short-lived effector cells, which depletes lineages quickly. This puts the thymus under pressure to create naïve helper T cells, but age and HIV both affect thymic function and impair bone marrow and lymph nodes. The immune system is fighting a losing battle to protect host cells from infection, and slowly progress erodes. The chronic phase ends when the viral load starts climbing again and the CD4+ T cell count declines to under 200 cells per cubic millimeter.[43] In the beginning of this phase patients start to feel better, and even if untreated may continue to live symptom free for many years.[46]

In the final phase, opportunistic infections occur. Helper T cells are too depleted and the immune system can no longer fight the HIV infection. Opportunistic fungal and

bacterial infections are detected, the patient is diagnosed as having AIDS and, without treatment, death may occur in two to three years.[43]

2.4 HIV types, groups and subtypes

HIV isolates that differ by more than 50% are classed as different types, either HIV-1 or HIV-2.[2] HIV-1 is the most common and widely distributed,[3] accounting for the majority of HIV infections worldwide.[4] HIV-1 is further classified into distinct lineages, Main (M), New (N), Outlier (O) and Putative (P), reflecting four separate introductions of SIV into humans,[3,5,6] and there are eight distinct lineages of HIV-2.[7]

Most recent common ancestor (MRCA) timing estimates the HIV-1 group M strain moved from NHP to human infection in around 1930. In 1998, Zhu *et al*. reported the oldest HIV-1 group M sequence known at the time, the ZR59. The isolate was taken in 1959 from an adult male living in the Democratic Republic of Congo (DRC). Phy showed the sequence branching off from the Subtype D lineage, which in turn branched off from Subtype B. It was half the distance from the root compared with strains from 1998, suggesting it was significantly older.[47] In 2008, Worobey *et al*. reported a sequence from the same part of the DRC, collected in 1960. The isolate 'DRC60' was found to cluster with Subtype A, and was again found much closer to the root than current strains.[48] These two sequences made a convincing case that the HIV-1 group M epidemic began before 1960. When MRCA timing was recalculated using these strains, a new origin year of 1921 (CI=1908–1933) was estimated, using a constant population size.[49] The clustering around different subtypes and the isolates originating from both male and female patients suggest that there was already wide genetic diversity in DRC by 1921, and that transmission was primarily heterosexual.[49]

The first strain that identified group O was discovered in 1990 and it is mostly restricted to Cameroon, Gabon and other countries bordering the DRC.[50] Group N was discovered in Cameroon in 1995,[50] and in 2009 a new HIV-1 variant was reported there that did not cluster

phylogenetically with the M, N or O groups. It was designated the prototype of a new lineage – group P (putative), and is believed to originate from gorillas rather than chimpanzees, although the woman in whom the sequence was found reported no exposure to gorillas or bush meat, suggesting the strain was already circulating in the human population in Cameroon.[5,47]

Over 99% of the global HIV epidemic is attributable to the M lineage[4] and there are at least nine phylogenetically distinct subtypes, A–D, F–H, J and K.[8,9] Subtypes A and F also have sub-subtypes, A1–A4 and F1/F2. Subtypes B and D show many similarities and could be considered sub-subtypes of a single subtype, however historically it has been difficult to change subtype designations.[7] Two other subtypes (E and I) were initially identified based on the *env* gene, which is highly variable. However a mosaic structure along the entire genome was identified during full genome sequencing, and subtype E was reclassified to CRF01_AE, and subtype I to CRF04_cpx.[51]

HIV-1 mutates at an extremely fast rate – about one million times faster than eukaryote DNA.[2] The longer a strain has been established in an area, the greater the time available for mutation to permit evolution into different strains, including recombinant forms incorporating different subtypes.[2] Both intra- and inter-subtype variation occurs in HIV-1. Within a subtype (intrasubtype), the threshold of genetic variation is 8–17%, while to distinguish between subtype (intersubtype) variation is 17–50%, depending on the subtype and the region of the genome being examined.[52] The *pol* gene is less divergent than *env* because of the critical enzymes it encodes to ensure viral replication, namely RT, PR and IN as mentioned previously, and *env* evolves faster than *pol* due to continuing antibody escape and greater immunological pressure.[53] Hence the *env* gene is the most diverse genetically,[7] while the *gag* gene is even less divergent than *pol* because of the functional constraint imposed by the encoding for core proteins.[7,20] Differences in evolutionary rates have been

found between regions, with some under significant functional constraint, such as component proteins that perform regulatory, functional, and structural roles. These constraints impact on the likelihood of nucleotide change and therefore viral diversity along specific parts of the HIV genome.[54] This can also differ by subtype/CRF, and it is thought that this heterogeneity of selective pressure impacts on viral fitness between the group M subtypes.[55]

Despite constraints on diversity, intersubtype recombinant (ISR) viruses have increased rapidly in the last five years and there are now 75 major circulating recombinant forms (CRFs) ≤http://www.hiv.LANL BLAST.gov /content/sequence/HIV/CRFs/CRFs.html≥, last accessed 12 Sept 2015) and multiple unique recombinant forms (URFs).[4,56] For an ISR to be considered a CRF rather than a URF, at least three epidemiologically unlinked cases must be identified that are identical genetically. Intrasubtype recombination (InSR) is more difficult to detect and to date, studies have largely been restricted to subtype B and C viruses.[57–59] A relatively recent study in Tanzania found a high prevalence of InSR along the *env* region for subtype A, B, C and D, and authors postulated that InSR may not always be the result of co- or super-infections but may also be due to an intra-host pool of viral quasispecies that develops over time.[60]

2.5 Subtype variation by geography and route of transmission.

Historically, HIV-1 subtypes have been associated with specific parts of the world and with particular modes of transmission.[14] However, the distribution of subtypes and CRFs is becoming more heterogeneous globally, related to population mobility, diversity of sexual contacts through travel and migration, the increase in IDU and the impact of antiretroviral therapies.[21,22] This highlights a type of founder effect, where the subtype or CRF remains the most common strain transmitted within an area or risk group but transmission branches out to other geographic regions or populations.[7,14,52] The role played by biological and viral

factors in the ease of transmissibility is still unclear,[52] but many reports suggest that CRFs may have improved biological capacity compared to the parent strains. For example, a very recent Cuban study assessed rapid disease progression against epidemiological, clinical, viral and immunological parameters and found 19_cpx, (recombinant of subtypes A, D and G) was very evolutionary fit and patients with this CRF had higher viral loads at diagnosis compared to patients infected with parent subtypes.[61] It was found to cause rapid progression to AIDS and authors suggest the driver may be the subtype D PR region; subtype D is known to be associated with faster disease progression.[61] Accurate identification of subtypes, CRFs and URFs globally and the linking of these with epidemiological data will assist greatly in understanding the course of the HIV global epidemic.[7]

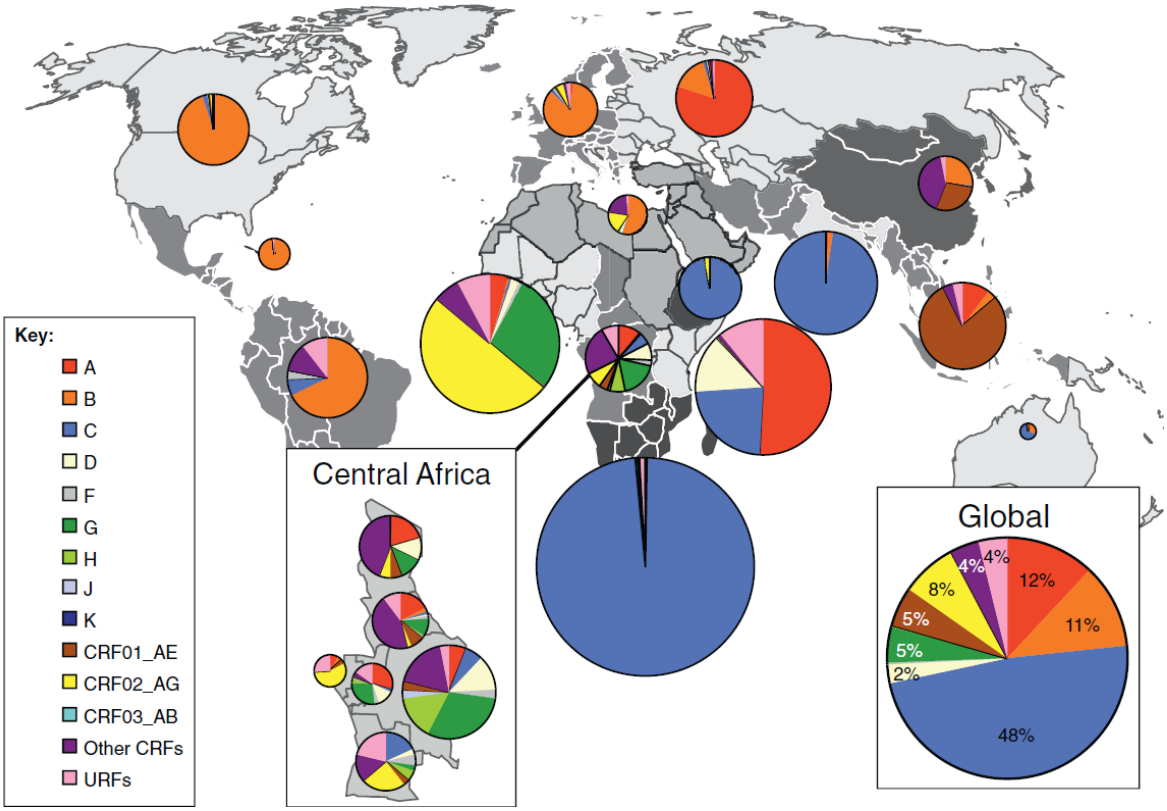


Figure 3. Global distribution of HIV-1 subtypes and recombinants. (Hemelaar 2012)[33]

Figure 3 shows the prevalence of subtypes and CRFs by geographic region between 2004 and 2007.[33] Subtype C accounts for approximately 48% of all HIV-1 infections worldwide.[7] It is the predominant subtype in eastern Africa and India,[14] and in southern Africa it accounts for almost 100% of infections.[7] It is most often transmitted by heterosexual contact.[52]

Subtype A has a global prevalence of 12%. In east Africa, eastern Europe and central Asia it is mainly transmitted through heterosexual contact but in eastern Europe and central Asia there are very high transmission rates through injecting drug use and iatrogenic events, because of central Asia's location on a major drug trafficking route.[18,62]

Subtype B accounts for 11% of global infections. It is thought to have originated in central Africa, spread through a single transmission from Congo to Haiti and then onward to the United States in the mid-20th century.[63,64] Subtype B infections are very rare in sub-Saharan Africa. It is a subtype largely confined to developed regions including North America, Western Europe and Australia, and is predominantly transmitted by MSM and injecting drug use.[14,52] In developed countries subtype B has been the principal infective agent but there is a growing proportion of non-B subtypes transmitted through heterosexual contact.[14]

The CRFs 02_AG and 01_AE have the fourth and fifth highest prevalence globally, with 8% and 5% of all infections respectively. 02_AG is mainly located in West Africa and is transmitted by heterosexual contact, while 01_AE circulates mostly in Southeast Asia, transmitted through heterosexual contact and injecting drug use.[7] In South Asia, 01_AE was first transmitted mostly by heterosexual contact and subtype B was found mainly among people who injected drugs, but 01_AE is now the major strain transmitted by both routes, with a number of 01_AE/B recombinants circulating.[7] Phy of 02_AG is uncovering increasing viral diversity among 02_AG strains. Subtype discordance has been found between two or more subgenomic regions, either because of co-infection or super-infection

which can lead to URFs or completely new recombinant strains.[65] Over time, transmission of both 02_AG and 01_AE has increased in other regions of the world because of travel and migration.[66]

Subtypes D and G account for 7% of global infections while CRFs comprise 4% and it is thought that URFs comprise another 4%, though some reports state URFs can account for more than 30% of new infections in regions where multiple subtypes and CRFs co-circulate.[52,67] Subtypes F, H, J and K combined represent ≤1% globally.[52] Subtype D is mainly found in East Africa while subtype G is found in West Africa.[52] Of the F, H, J and K subtypes, only F has been found to be widely distributed globally and evenly spread across regions, while H, J and K are predominantly localized to central, southern and western Africa respectively.[46]

One-fifth of HIV infections worldwide are therefore thought to be a recombinant of some form.[52] Africa contains the widest variety of URFs followed by Latin America,[52] while different CRFs are associated with certain regions or transmission routes. For example, CRF07_BC and 08_BC are associated with injecting drug use in China,[7] while in West Africa CRF02_AG accounts for 50–80% of HIV infections but prevalence of CRF06_cpx, (an A, G, K and J recombinant) can reach 20–50% in some regions.[7]

The epidemic in the DRC is thought to be the epicenter, as shown by the high rate and variety of co-circulating strains. It is believed that the origin of group M strain variation began here and spread outward to the rest of Africa and eventually the world through increased rates of global travel and permanent migration.[7] Hence HIV subtype distribution is dynamic and the rate of viral diversity and complexity across the globe will continue to increase as travel becomes easier and more affordable, and more migrants and refugees leave countries of high HIV prevalence. An example is France, where the widely circulating subtype B virus among native MSM and heterosexual populations has now recombined with

the 02_AG virus from West African migrants to create a 02_AG/B recombinant.[68]

2.6 Changing prevalence and effects of subtype

As mentioned previously, subtype B has predominated in developed countries[8,9,69] where transmission is primarily through male-to-male sex.[9,70,71] However, the prevalence of non-B infections is increasing. Recent studies in France have found non-B prevalence rates of 42–48% in newly diagnosed HIV infections.[72] In Italy, non-B prevalence rates rose from 25% in 2000 to over 60% in 2008, with African ethnicity and heterosexual acquisition as independent predictors.[9] In a Maryland cohort the non-B prevalence rate was 13%, with the majority of non-B infections from Washington, DC.[73] In one broad population-based study in the United States a national non-B prevalence rate of 5.1% was reported in newly diagnosed infections.[74] In a recently published Australian study, Chibo and Birch (2012) found a non-B prevalence rate of 22% in a Victorian cohort.[69] There is evidence that subtypes and CRFs differ clinically and phenotypically, by way of replication fitness, rate of disease progression, co-receptor utilization, transmission route, transmissibility, accuracy of current diagnostic assays, response to therapy and development of drug resistance mutations.[3,21,22,65,75–77] Many studies have shown that patients infected with subtype D and some CRFs can present with more rapid disease progression, most likely because of its propensity to be dual tropic,[4,14,77–79] while a recent study found very fast disease progression for the newly identified 02_AG/A3 recombinant compared with subtype A virus.[80]

Historically, diagnostic and drug resistance tools were designed based on subtype B isolates, which may affect identification of non-B virus and associated resistance mutations.[17] All subtypes and CRFs appear to be equally sensitive to initial treatment,[14,81] however, inherent mutations present in particular subtypes or CRFs and cross-species transmission routes may lead to subtype-specific differences in drug resistance mechanisms.[15,81,82] In addition, subtypes may differ in terms of transmission efficiency of

resistant strains and genetic barrier differences, which can alter responses to therapies targeting PR and RT. It is therefore critical to understand how to identify, characterize and clinically manage people infected with non-B HIV.

2.7 HIV surveillance

Acquired immune deficiency syndrome (AIDS) was first recognized around 1980, and the etiologic agent of AIDS, human immunodeficiency virus (HIV-1) was identified in 1983. This global epidemic is one of the worst seen globally in modern times.[7] To date, over 65 million people have been diagnosed with HIV, with an estimated 36.9 million people living with HIV infection or AIDS in 2015 compared with 8 million in 1990.[43,83–85] Of those living with HIV infection or AIDS, approximately 40% are young adults, and 6% (or about 2 million) are children. The majority of people affected with HIV live in sub-Saharan Africa, where the adult prevalence rate can range from 0% to 39%.[83] The current prevalence rate in Australia is 0.1%.[86]

There is no vaccine or cure for HIV or AIDS, but ART can effectively reduce and control replication of the virus. The number of new global HIV infections annually has declined from an estimated 5.4 million in 1999 to an estimated 2.5 million in 2011[52,87] and an estimated 2 million in 2015.[85] As access to treatment increases, there is a corresponding decrease in HIV/AIDS related deaths. Deaths were estimated at 2.8 million in 1999, compared with 1.7 million in 2011[87,88] and 1.2 million in 2015.[85] The number of people receiving ART has increased from 13.6 million in 2014 to 15 million in 2015,[85] and 41% of adults in need of ART are now receiving it, compared with only 23% in 2010.[85]

About 50% of people infected with HIV are female, while children from birth to 14 years account for around 9% of all HIV infections (estimate range: 3.1–3.8 million) with approximately 330,000 new infections occurring each year, and young people aged 15–24 years account for 45% of new HIV infections.[89] In total, people aged under 25 years

constitute over half of the population living with HIV. Mother-to-child-transmission (MTCT) however, has decreased over the last decade because of the rigorous expansion of prevention programs[88,89] and the improvement in ART regimens during pregnancy and breastfeeding from monotherapy, which can often lead to TDR, to a more efficacious combination therapy.[90] In developing countries, where MTCT is highest, antiretroviral treatment for pregnant women increased from 9% in 2004 to 33% in 2007.[89] However, the number of children living with HIV has increased from 1.6 million in 2001 to 3.3 million (estimate range, 3.1–3.8 million) in 2011. This can be explained in part by a decrease in HIV-related mortality because of the increased access to ART,[88,89] and the continued relatively high rate of MTCT in the most economically disadvantaged regions where adequate combination therapies during pregnancy, time of birth and breastfeeding are not widely available.[90] However, it is believed that the high number of children living with HIV is largely because of iatrogenic transmission, especially in central Asia, Russia and sub-Saharan Africa.[62,91]

Almost all (96%) global HIV infections originate in LMICs,[89] with the global epidemic branching into three major patterns: a predominantly heterosexual epidemic in sub-Saharan Africa, the Caribbean and surrounding countries,[89] a disproportionately high HIV prevalence in sex workers and people who inject drugs, especially in Asia and low income countries like Papua New Guinea, and a high HIV prevalence in men who have sex with men (MSM) worldwide, including in developed nations such as Australia and the US.[89,92] Historically in Australia, the HIV epidemic has predominantly been within the MSM population, although this pattern is changing as more heterosexual infections are reported.[66,69]

Prevention efforts remain a subject of conjecture globally. Prevention and control of HIV rely on three types of strategies: behavioral (the individual), treatment (the virus) and social (the environment).[93] To date, global prevention efforts have focused on *individual risk*

behavior, for example education, increasing condom awareness, treatment adherence and needle exchange programs, and *treatment strategies,* including best time to begin treatment according to immune system response, up-scaling of ART availability and affordability, improving drug regimens, identification of new target sites along the genome, and the introduction of pre- and post-exposure prophylaxis.[86,93] Although current treatment regimens are very effective at suppressing viral load and reducing transmission rates, there are still major challenges to overcome with regard to treatment availability, adherence and drug resistance, which are predominantly related to social structures such as a region's financial status and cultural norms. There is also growing viral diversity globally,[23] which makes finding a cure or vaccine very difficult.[83,93]

Social strategies that tackle issues such as socioeconomic disadvantage and geographical region, political and cultural systems and gender inequity can have considerable impact on the incidence of HIV. Sub-Saharan Africa is a prime example. While the proportion of people engaging in high-risk behaviors in sub-Saharan Africa is no different to that of Australia, the incidence of HIV is much higher, in large part because of political and cultural oppression, economic deprivation, and poor access to medical care. However, despite evidence that shows social structures have far greater impact on the HIV epidemic than behavioral or treatment strategies, integration and acceptance of social strategies into prevention and control efforts has been very slow, and the prevention focus remains on modifying individual behavior and increasing treatment.[93]

Prophylaxis is increasingly being used as a preventative method for people who are deemed to be at high risk of HIV infection, including people who practice unsafe sex with multiple partners, uninfected partners of people living with HIV, people who inject drugs, and sex workers.[94] Pre-exposure prophylaxis (PrEP) can prevent HIV infection in the event of an exposure risk. Oral Truvada®, taken once daily, contains the antiretroviral medications

tenofovir and emtricitabine. Clinical trials of the efficacy of PrEP have found reduced risk rates of HIV infection of 44–92% among MSM, 62% among heterosexual men and women, 75–90% among HIV discordant couples, and 49–74% among people who inject drugs.[95–98] While PrEP is taken on an ongoing basis, post-exposure prophylaxis (PEP) is a once-only 28-day course of treatment used within 72 hours after suspected exposure to reduce the likelihood of infection. PEP can be a two- or three-drug regimen. Because of the ethical sensitivity of PEP, efficacy research has largely been confined to human case studies and non-human laboratory studies, both of which have found significant decreased risk of acquiring HIV after taking PEP.[99–101]

In terms of HIV treatment, developed countries have the highest rate of ART coverage, principally because of government-subsidized health care such as the Australian Medicare system. In Australia, 88% of HIV diagnosed people are receiving treatment,[102] compared with just under half of HIV positive people (47%) in Latin America, 44% in the Caribbean, 41% in sub-Saharan Africa, 36% in Asia, 18% in Eastern Europe and Central Asia, and only 14% in the Middle East and North Africa.[85]

Of the 2 million new HIV infections that occur globally every year, Australia accounts for approximately 1000, or 0.05%.[102] When broken down by state, Queensland had the highest rate of HIV infection in 2014 (5.3 per 100,000 population) followed by Victoria (2.5) and New South Wales (4.7) while South Australia had the lowest (2.5).

South Australia has a relatively small population (1.68 million) compared with Victoria (5.79 million) and New South Wales (7.5 million). It is a major urban environment located on the plain bounded by the Saint Vincent's Gulf and Mount Lofty Ranges, with the majority of people residing in or around the city of Adelaide. The remainder of the state ranges from agricultural land with rural towns, and a significant amount of the top half of the state is principally arid with very little population. Historically, the spread of HIV infection in South

Australia has been predominantly confined to a small subset of the MSM community residing in urban areas surrounding Adelaide, with a small proportion of heterosexual acquired infections acquired locally and overseas.

The overall proportion of Australians living with HIV is approximately 0.1%, although this prevalence increases alarmingly to 17% among MSM.[86] Prevention and treatment programs were very effective in reducing the number of new infections in Australia in the 1980s and 1990s, but national surveillance data have shown an increase in new diagnoses annually since 1999, stabilizing between 2012 and 2014.[21,102–104] A number of factors have contributed to this, including an increase in testing and increased occurrence of risk behaviors in MSM and heterosexual communities. The increase has occurred despite approximately 88% of diagnosed people receiving ART, with high treatment adherence as shown by the majority of treated people having undetectable viral loads, and a doubling of testing rates across the country.[102]

Approximately 35,000 people have been diagnosed with HIV in Australia since 1982, with a further 12% possibly infected and undiagnosed.[102] Late diagnoses comprised 28% of all newly diagnosed infections in 2014, defined as being infected for at least four years before being tested.[102] Late diagnosis was highest among people born in Southeast Asia and sub-Saharan Africa. Seventy percent of HIV infections were transmitted through male-to-male sex although the proportion of infections acquired heterosexually is increasing, predominantly in people from countries with high HIV prevalence.[66] The proportion of undiagnosed HIV infection is of real concern for at-risk populations such as MSM and people who inject drugs; in 2008 Wilson estimated that one in three newly diagnosed males had been infected by the 13% of undiagnosed men in the population and that condom use had decreased.[105] In 2014, almost 40% of MSM in Australia reported unprotected anal intercourse, a rise of approximately 6% over the last decade.[86] Of those interviewed, 15% of

Australians who inject drugs reported needle sharing.[86] Likewise, the high proportion of late diagnosed infections among people infected overseas has serious implications for contact tracing within Australia and overseas, treatment options and ongoing health care.

Over the last 25 years, a number of HIV surveillance studies in Australia have reported the issues mentioned above, including the annual HIV surveillance report by the Kirby Institute (University of NSW), which reports national and state/territory rates of new diagnoses, incidence of HIV, exposure routes, risk behaviors and testing prevalence.[21,86,103] Surveillance systems for monitoring HIV are very important for identifying trends and targets for prevention programs, education and further research.[106] Global surveillance of the HIV epidemic has been undertaken through a joint collaboration of the World Health Organization (WHO) and the Joint United Nations Programme on HIV/AIDS (UNAIDS) since 1996.[107] In the early days of HIV surveillance, a first generation model was used that relied solely on patient data once the infection had progressed to AIDS. In 2000, the second generation surveillance (SGS) model was introduced.[107] This model incorporated collection and management of a wide range of data from HIV and AIDS notifications, STI surveillance, and behavioral and clinical information. The data is analyzed both globally and by key regions and populations most at risk, which allows creation of HIV estimates and projections via mathematical modelling that can be used to implement public health programs.[107]

However, to date the molecular epidemiology of HIV has not been included in the SGS, despite the fact it is reported frequently in many parts of the world as a tool for surveillance of genetic diversity and to monitor transmission and geographic pathways of genetic variants.[6,73,108,109] Molecular epidemiology in HIV research involves the use of molecular biology techniques to assess differences in nucleic acids among multiple HIV sequences, and to investigate how these differ among human populations.[110] For example, Hong Kong was recently the first country in Asia to map transmission of new recombinant HIV-1

subtypes.[108] Large genetic diversity among HIV-1 strains was found, including two recombinant strains that are newly circulating. A Latvian study recently found different levels of TDR across HIV subtypes,[111] while in 2012 a Victorian study found an increasing proportion of non-B subtypes that were related to an increased proportion of female infections.[69] Most recently, two studies in Malaysia discovered new recombinant forms of HIV (CRD58_01B and CRF74_01B) circulating amongst people injecting drugs.[13,112]

Variations in HIV are always occurring. The HIV-1 group M virus which first jumped the species barrier from simian to human has now mutated into a plethora of subtypes that circulate today. Antiretroviral therapies and other prevention strategies also influence the rate of mutation and evolution of the virus over time. The variability can be examined closely by identifying transmission patterns of different strains through genetic analysis, and tracking global and local changes in regard to ease of transmission, risk factors, prevalence by geographic region and virulence.[22,110] This type of analysis has potential large-scale implications for education, prevention strategies and therapy administration and development. It is therefore crucial that viral diversity is incorporated into the SGS model, or that a third-generation model is created to ensure the tracking and monitoring of viral diversity across time, geography, transmission route and other relevant demographic and epidemiological information.

In 2000, South Australia became the first Australian state to integrate genotypic drug resistance testing as part of routine HIV reporting and surveillance system. The data from this reported in Chapter Four provide the first analysis of Australian trends and molecular epidemiology of HIV *pol* subtype distribution over the past 14 years.

Part 2

2.8 TDR

The introduction of ART has dramatically changed HIV epidemics in many regions of the world,[113] and its availability is having positive impacts in even the most socially disadvantaged regions, although there is much still to be done.[114] Current triple combination therapy is highly effective at suppressing viral load if taken correctly. The dosage is better tolerated and the number and severity of side effects has decreased. This has led to dramatically increased life expectancy rates among people living with HIV.[115]

There has been much debate in recent times about the test and treat approach, whereby patients begin treatment as soon as they are diagnosed rather than waiting until their CD4+ T cell count falls below 500 cells/mm^3, which is the current WHO treatment guideline.[116] Findings from the Strategic Timing of Antiretroviral Treatment (START) study have established that it is more beneficial for all people infected with HIV to begin treatment immediately rather than to wait. START is a randomized controlled trial that began in 2009 and spans 35 countries. It was designed specifically to identify the optimal time to begin treatment. However Ammaranond *et al*. (2003) predicted that if treatment rates increased to 85%–90% of the infected population, then rates of resistant virus among newly transmitted virus could eventually be as high as 70%.[12]

Despite advances in ART availability and successful viral suppression, almost one-quarter of people experience treatment failure during first-line regimens, which often results in secondary or acquired drug resistance whereby mutated strains become resistant to certain drugs or drug classes[117] These mutations can arise because of non-adherence to treatment, subtype propensity for resistance or selection pressure of certain drugs. The global increase in ART access and extended life expectancy may lead to an increase in *primary* or *transmitted* drug resistance (TDR). TDR mutations (TDRMs) are resistance mutations that

have arisen in a host because of treatment failure and which are then transmitted to another individual during HIV infection. Forward transmission of TDR from a treatment naïve person to an uninfected individual can also occur.

It was first believed that resistant virus would pass to an untreated individual and rapidly revert to wild type in the absence of drug pressure. However it is now widely accepted that TDR strains can remain the dominant population for an extended period of time.[118] Figure 4 shows the difference between acquired and transmitted drug resistance, and acquired drug resistance will be discussed in more detail in section 2.9.

2.9 The history of HIV treatment

To combat HIV infection, viral target sites for inhibition were identified. One of the earliest drugs created, azidothymidine (AZT) targeted the virus's own RT to block reverse transcription without, hopefully, harmful side effects to the host.[43] Earlier it was mentioned that each virion transfers two HIV RNA strands into a host cell, along with the viral proteins RT, IN and PR. RT uses the HIV RNA to create a complementary strand of DNA, using the host cell's own nucleotides. AZT is very similar to the nucleotide thymidine, and fools RT into incorporating it into the HIV DNA.[43] Thymidine contains a hydroxyl group (-OH), which is needed for attachment of the next nucleotide of a DNA strand. AZT has an azide group ($-N_3$) in place of the hydroxyl group, effectively terminating the DNA strand and thus preventing new viral proteins and hence virions being assembled.[43]

Initially there were very good results with AZT. A reduction in the loss of CD4+ T cells was achieved and viral load was suppressed. However, AZT also interfered with the function of host enzyme polymerase, interrupting DNA synthesis.[43] Within a few years, treatment started failing, and investigations found that because of a lack of error correction by RT, the HIV genome had mutated and was substituting amino acids at the RT active site that blocked the effect of AZT. Viral replication was able to continue even in the presence of AZT. HIV

has the highest mutation rate of any virus or organism in the world. Thousands of generations of HIV replications take place inside an individual, so a single strain of HIV can produce hundreds of different RT variants over the course of an infection.[43]

In the presence of AZT, only resistant strains could replicate, and these became the dominant patient strains. In addition, because natural selection favors non-mutant virions, it was discovered that the predominant circulating strain would rapidly revert to wild-type strain if the selection pressure afforded by drug presence was removed by cessation of AZT treatment, Figure 4.[43,118] This discovery created the realization that effective antiretroviral drugs were needed.

2.10 HIV drug classes

The first drug class created for HIV treatment were nucleoside analogue reverse transcriptase inhibitors (NRTIs). When NRTIs are phosphorylated inside the host cell they inhibit RT, which terminates the transcription and ultimately viral replication.[11] One of the first NRTI drugs was zidovudine (ZDV), however it was soon shown to have adverse side effects and had no long-term benefit.[10,11] The design and introduction of new HIV treatment progressed slowly; the NRTIs didanosine (ddI), lamivudine (3TC), stavudine (d4T) and zalcitabine (ddC) were eventually developed and used as single-class dual therapies after clinical trials such as Delta and the AIDS Clinical Trials Group deemed dual therapy had superior efficacy.[10] Any one-drug method used alone soon resulted in the same outcome as AZT.[43] Therefore, to increase the number of mutations required for treatment escape, more than one drug needed to be used simultaneously.

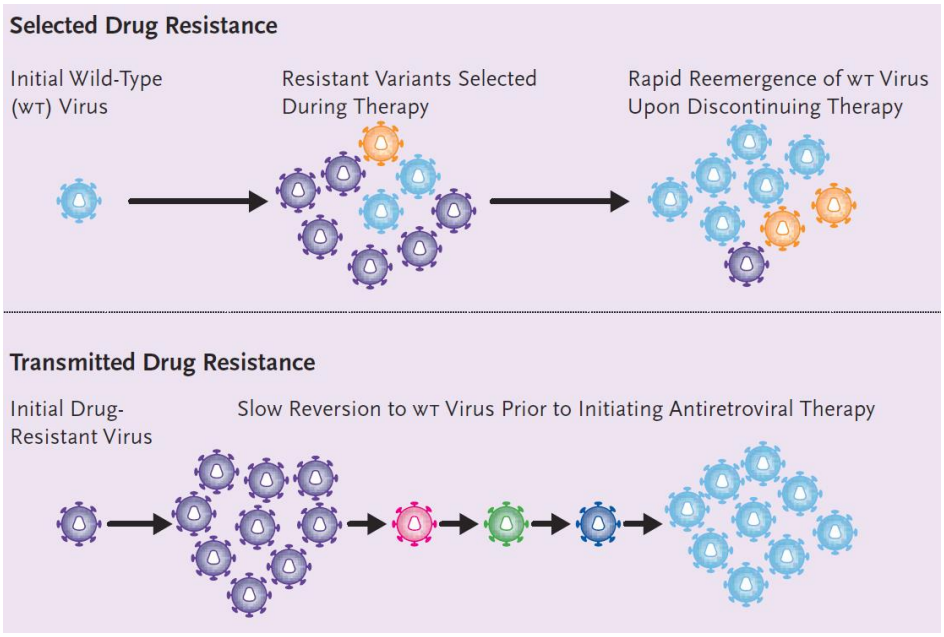


Figure 4. Acquired versus transmitted drug resistance. (Kuritzkes, 2004)[118]

During the mid-1990s there was greater understanding of HIV replication with the discovery that around 10 billion virions that affected CD4+ T cells and other immune targets were produced daily in one individual.[10] This new understanding led to routine viral load testing to assess treatment response. Two new drug classes also emerged at this time, protease inhibitors (PIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). PIs work by inhibiting the activity of PR to cleave proteins for the final assembly of new virions, and NNRTIs bind directly to RT preventing transcription.

By 1996, a new treatment regimen called Highly Active Antiretroviral Therapy (HAART) which combined drugs from two or all three drug classes became the new standard first-line treatment in developed countries.[10,43] HAART does not cure HIV, because reservoirs of HIV genomes are still present in chromosomes of resting white blood cells and possibly other tissues. However initial use of HAART afforded very rapid and sustained results with a marked decrease in AIDS-related illness and hospitalizations. In fact, in just a few years the number of deaths fell by almost 70% in the US, and within 5 years the number

of AIDS diagnoses in Australia had fallen from 953 in 1994 to 216 in 1999.[10,43,88] However, unpleasant treatment side effects were common with PIs, including persistent nausea and diarrhea, and the drugs had a tendency to interact with other medications. NNRTIs caused rashes and hepatotoxicity.[11] In addition to adverse effects, multiple combination therapies required a large number of tablets to be taken daily, and dose intervals had to be exact to prevent further side effects and possible selection pressure. This made treatment adherence more difficult than with mono or dual therapies, and approximately 12 months after the initiation of HAART, patients were experiencing treatment failure, observed by detectable viral load and resistance mutations to the new drug classes.[10] Drug resistance was highest in those whose disease had progressed to AIDS, those who had previously been on mono or dual therapy and those who had not adhered to treatment.

As a result, pharmaceutical companies have sought to improve therapies by combining multiple drugs into a single pill for a once-daily dosage regimen. This has led to a marked improvement in adherence to treatment and stability in TDR prevalence.[10] Protease inhibitors are infrequently used in first-line regimens because of the side effects mentioned, and consequently TDR is mostly seen with NNRTI and NRTI.[10]

EFV and NVP are two of the most commonly used NNRTI drugs globally, both in high-income countries and LMICs.[15] However, both drugs have a low genetic threshold to resistance. A single mutation can lead to treatment failure and there is a high level of cross-resistance to other NNRTI drugs, which is problematic for second-line regimens. Etravirine (ETV) and Rilpivirine (RPV) are commonly used in second-line regimens that effectively suppress NVP and/or EFV resistant viruses.[119,120]

A triple combination therapy of two NRTIs and one NNRTI is the current preferred first-line regimen; usually with NVP or EFV as the NNRTI, and 3TC, AZT, or d4T as the NRTI.[15] A PI is often used for second-line regimens in conjunction with two NRTIs; NFV or

lopinavir/ritonavir (LPV/r) with ABC and TDF or ddI.[121]

Two other drug classes have also been developed, fusion inhibitors and integrase inhibitors (INSTIs). Fusion inhibitors prevent HIV from entering host cells by interfering with *env*-gp120 or *env*-gp41, or interfering with the receptors on the host cell, CXCR4 and CCR5. Enfuvirtide is the only one commercially available but it has not been widely used, including in Australia. This is because of the twice-daily injection regimen, which leads to an injection-site reaction in over 90% of patients.[122] No resistance to enfuvirtide has been reported in South Australia.

INSTIs prevent IN from splicing HIV DNA into host DNA, which prevents transcription of new viral RNA. Raltegravir was the first approved INSTI in 2007, and it has been found to be a potent inhibitor, active against multiple strains resistant to other drug classes, strains using both the CXCR4 and CCR5 co-receptors, wild-type strains and even HIV-2. Drug resistance mutations to INSTIs have only been reported recently compared with RTIs and PIs, because of their much later introduction.[123,124]

As discussed earlier, PEP and PrEP are increasingly being used as treatment-as-prevention strategies.[95,96,98,125] These prophylaxes are highly effective when taken correctly, but drug resistance can occur if treatment adherence is not rigorous.

Drug therapies have also dramatically improved for the prevention of vertical transmission from mother to child (PMTCT). In LMICs, especially sub-Saharan Africa, the preferred treatment was the use of a single-dose NVP regimen for the mother at the onset of labor and for the child within 72 hours of birth, but this was found to lead to rapid selection and transmission of virus that was resistant to NVP. This had significant implications for the use of first-line regimens containing NVP and EFV. The single-dose regimen has now been replaced by more efficacious triple combination regimens.

Guidelines have also changed for treating children under three years of age. First-line

ART should be started soon after birth using a LPV or RTV based first-line regimen rather than NVP, because these have been found to be more efficacious.[126]

2.11 The rate of TDR by geography

TDR is a growing issue globally through sexual contact, MTCT, iatrogenic transmission and injecting drug use,[18,62,117,127] with prevalence rates ranging from 0% in countries where ART is relatively new to over 30% in high income countries such as Australia.[18] Figure 5 shows the proportion of TDR across different regions.

ART has been available in high-income countries for longer than elsewhere and they have highest prevalence of TDR, with rates over 25% in some regions.[15] A study conducted in Victoria, another Australian state, found a TDR prevalence rate of 16% in 2009.[10] Because of these high rates, genotypic drug resistance testing at time of diagnosis and prior to treatment has been introduced. In high-income countries patients have regular viral load checks, and resistance testing is conducted if increases in viral load indicate the treatment is failing. This allows a rapid switch to an alternative drug regimen which, in turn, reduces the risk of TDR. These approaches have led to stability of and even decrease in TDR in developed countries in recent times.[18] However, there are high costs associated with viral load and resistance testing, which require patient care and qualified medical staff, and therefore access is limited. Rather than treatment based on individual patient testing, many LMICs rely on WHO guidelines for best practice, including use of standard first-line and second-line regimens at treatment initiation and later switching. Where possible, treatment is also guided by clinical disease progression and monitoring of CD4+ cell counts.

Although ART has increased rapidly in LMICs over the last 10 years, especially sub-Saharan Africa,[128] roll-out is relatively recent compared with high-income countries, and the majority of patients are beginning first-line treatment using highly efficacious triple combination therapies. Therefore, rates of TDR may not follow the pattern seen in high-

income countries where early monotherapy and one-class dual therapy led to widespread resistance.[128] Indeed, mathematical modelling of expanding ART suggests that the prevalence of TDR will remain less than 5% of newly diagnosed HIV infections until treatment in a region is widespread and has been available for many years.[128] For instance, in African countries that began implementing rapid widespread ART between 2001 and 2003, TDR prevalence rose from 2.8% to 5.3%.[129,130] Surveillance of TDR is therefore crucial in order to monitor the efficacy of first-line ART, especially in LMICs where regimens are based on guidelines rather than individualized patient care and treatment options are limited.[129]

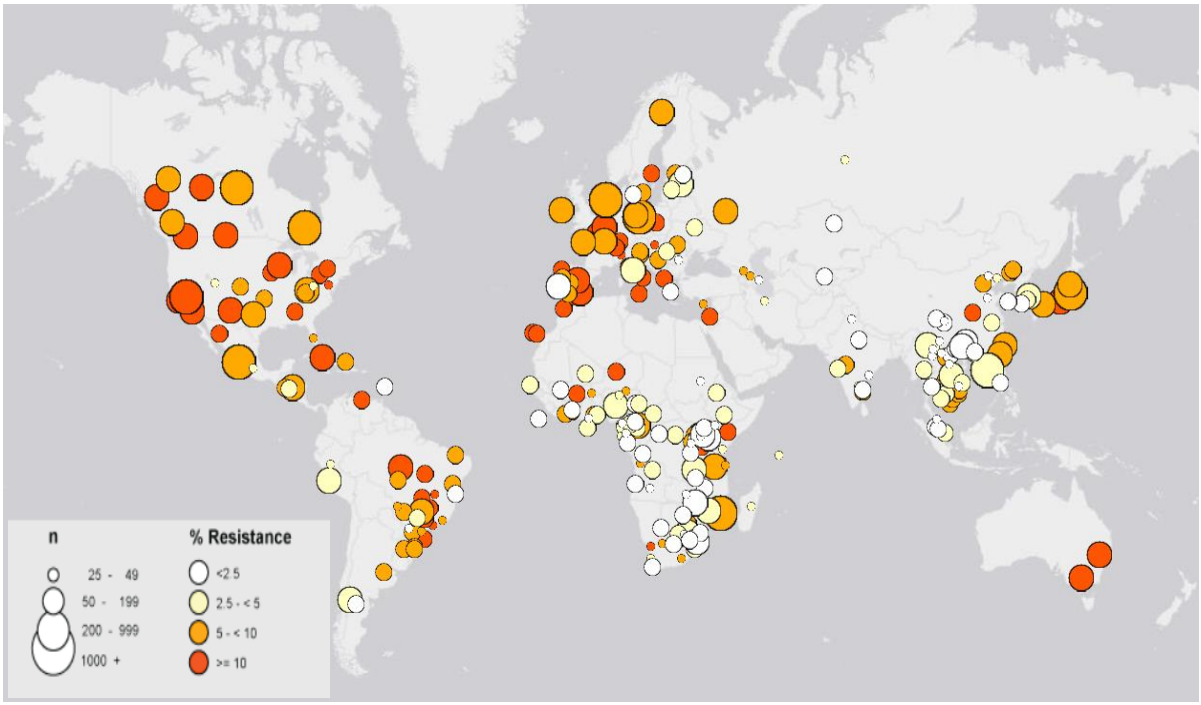


Figure 5. Drug resistance in treatment-naïve populations by geographic region. (Stanford University, 2014).[131]

Other factors also determine whether TDR levels increase. These include length of time patients remain on failing regimens which allow ongoing replication and transmission of resistant strains. Non-adherence to treatment can lead to drug resistance and this may occur

because of irregular drug supply, cost, travel time to clinics, family pressure, or the use of traditional medicines. Other factors to consider are rates of treatment failure among people with TDR and viral rebound, viral fitness of resistant versus wild-type strains and transmission risk.[128] Therefore caution is warranted when increasing ART in LMICs and it is imperative that regular surveillance of TDR is conducted to ensure continued efficacy of the very limited ART regimens available in these countries and to inform treatment guidelines on the success of HIV prevention and viral load suppression.[128,132,133]

2.12 Drug resistance and subtype

There are mixed reports about the effects subtype diversity has on drug susceptibility and resistance,[134] although from the research it appears that subtypes and CRFs are equally sensitive to initial treatment.[14,81] However, transmitted drug-associated mutations that are present before treatment may affect subtype-specific pathways of secondary resistance to treatment.[81,82,134] That is, resistance may evolve differently in response to ART, depending on subtype.[135] Subtype-specific transmitted resistance may affect treatment efficacy, particularly in countries where treatment options are limited or suboptimal,[17,82] however Langs-Barlow and Paintsil (2014) suggest that drug susceptibility and treatment outcome may be influenced more by quality of care because of geographic and economic status.[134]

Understanding genetic diversity is still very important in the treatment of non-B subtypes. Current drug regimens targeted against B may not be equally effective over the long term in non-B subtypes and may lead to faster drug resistance.[82] Only recently have non-B subtypes been the focus of research evaluating ART effectiveness and the evolution of mutations.[134] A number of studies have found differences in immune response to treatment between subtypes, and a propensity for certain mutations to develop in some subtypes and not others.[134] This is discussed further in section 2.13 below.

Understanding subtype diversity may provide important information about antiretroviral

resistance profiles, such as the biological features of each subtype and the ways they relate to pathogenicity and transmission. It may also be relevant for clinical management, by ensuring the validity of diagnostic tests and laboratory markers of infection during patient follow-up. Studies have shown the increasing importance of recombinant strains of HIV-1 and their capacity for transmission, which is possibly greater than that of pre-existing strains such as pure subtypes.[75]

Historically, the understanding of drug resistance mutations has been based upon subtype B sequences in high income countries such as Australia, where ART and drug resistance testing has been available for longer.[15,20] Drug resistance data on non-B subtypes is growing but still relatively limited because of the later introduction of ART and the limited availability of testing and collection of surveillance data in LMICs.

2.13 Surveillance mutations

As already mentioned, in order to improve treatment outcomes and increase prevention of TDR, accurate population-based global surveillance must be conducted. To do this, there needs to be a consensus definition of genotypic resistance, so prevalence over time and within different regions can be compared.[136] Deciding upon criteria for a standardized list of surveillance drug resistance mutations (SDRM) is difficult, because new treatments are being introduced and new resistance mutations identified. In 2007, four broad criteria to classify SDRM were specified: the mutation causes or contributes to drug resistance as defined by experts, the mutation is non-polymorphic and does not occur at highly polymorphic positions, the mutation list is applicable to the eight most common HIV-1 subtypes, and the list should exclude extremely rare mutations that result from drug pressure.[136]

An initial list of 80 RT and PR mutations was created, however, the addition of new drugs and new resistance mutations over time resulted in a new list in 2009 consisting of

93 mutations, with 34 NRTI mutations at 15 RT positions, 19 NNRTI mutations at 10 RT positions and 40 PI mutations at 18 PR positions.[136]

Subtype-dependent differences in drug resistance have been found for all three drug classes. There are 19 common NRTI mutations associated with subtype B at 15 different positions: M41L, A62V, D67N, K70R/E, K65R, L74V, V75I, F77L, Y115F, F116Y, Q151M, M184V/I, L210W, T215Y/F, K219Q/E and insertions at position 69.[15] The insertion complex at position 69 confers resistance to all NRTIs when in the presence of one or more mutations associated with NRTI regimens containing thymidine (TAMs) at codons 41, 210, or 215.[15] TAMs are associated with failure of AZT and d4T, and enhance resistance to all NRTIs and include mutations M41L, D67N, K70R, L210W, T215Y/F and K219Q/E.[15,137] Abacavir (ABC) treatment failure most commonly leads to K65R, L74V, Y115F, and M184V mutations, and M184V/I mutations are also resistant to FTC and 3TC. Treatment failure of ddI is associated with the K65R and L74V mutations and the K65R mutation is also resistant to FTC, 3TC, and TDF.[15]

Mutations at positions K103, V106, E138, G190, Y181, Y188 and N348 are common among people treated with NNRTIs, with a very high prevalence of K103N among people with subtype B, C and G viruses,[138] and Y181C among people with 01_AE followed by subtype F viruses.[138]

Subtype B viruses most commonly carry the NRTI mutations V and I at position M184, followed by TAMs at positions M41, K219, and T215,[136] while the most common mutations found in non-B subtypes are the NNRTI mutations K103N, Y181C/I, and Y188C/H/L.[139–141] Non-B subtypes exposed to 3TC and FTC have also been reported to carry the M184V mutation.[141]

Subtype C virus exposed to AZT/ddI treatment have been found to develop a unique mutation pathway comprised of the 67N, 70R, and 215Y mutations associated with the use

of TAMs, but this pathway is rarely seen in subtype B virus exposed to the same treatment.[142] Other studies in Africa and India have found a high prevalence of the K65R mutation in patients infected with subtype C virus that have been exposed to d4T/3TC regimens,[143–147] while K65R is least likely to appear in subtype A virus.[148]

Subtype-dependent resistance patterns have also been seen among people on NNRTI treatment such as single-dose NVP used for PMTCT in LMICs.[15] Studies have found that subtype C virus harbors more resistance mutations after this monotherapy compared with subtypes D, A and CRF02_AG, [149,150] while another study found Y181C significantly faded from detection in subtype A infections eight weeks after the single-dose NVP compared with subtype D.[149–151] Eshleman *et al*. (2001) also found the K103N mutation most frequently appeared in mothers while the Y181C mutation was predominant in children.[150]

The PI L90M mutation is common across the major subtypes and CRFs, with a high prevalence in subtypes D and G, while M46L is most prevalent among subtypes B and F, and L76V is highest among 02_AG.[152] A study found differences in PR mutations between subtype B and CRF01_AE viruses exposed to nelfinavir (NFV), with 01_AE viruses carrying mutations L10F, K20I, L33I and N88S more frequently, while D30N, A71V and N88D were found in subtype B viruses only.[153]

Some mutation differences are genetically subtype-dependent, such as the RNA template mechanism whereby RT pauses at different codons according to subtype.[15,20] However regional differences also affect mutation prevalence among subtypes, including treatment regimen used, stage of disease at diagnosis and access to regular viral load testing. As access to triple combination therapy expands in LMICs, a clearer picture of drug resistance and susceptibility amongst subtypes and CRFs will emerge. This information will influence the formulation of new treatments and drug combinations that may be targeted at specific subtypes or recombinant viruses to maximize efficacy and adherence.[15]

2.14 Genotypic drug resistance testing and surveillance

As with any antiretroviral treatment, the emergence of drug resistant strains is expected but HIV therapy is of particular concern because of the requirement for long-term treatment, the error-prone nature of HIV and its propensity to mutate and replicate at very high rates.[19] As noted earlier, increasing proportions of HIV infections are becoming complex recombinants of multiple strains. This genetic diversity may affect resistance pathways to treatment including cross-resistance to multiple drugs within a single class.[15]

The key factors driving drug resistant strains are: availability of treatment, need for lifelong treatment, treatment side effects and adherence, length of time treatment programs have been in place in a geographic location, prevalence of acquired resistance, relative fitness of AR resistant strains, decreasing transmissibility of drug sensitive strains from treated patients, quality of care including hospital care, access to regular viral load and CD4+ T cell count checks, sexual risk behaviors, intravenous drug use and financial cost.[134] In resource-limited settings, additional factors such as limited treatment options with low genetic threshold for resistance, beginning treatment without genotypic drug resistance testing, use of contaminated blood and instruments in medical settings, and the use of single-dose prevention protocols for MTCT makes them a prime target for accelerated drug resistance, both acquired and transmitted.[19]

The consequences of TDR are significant. The absence of genotype and drug resistance testing before treatment will decrease the effectiveness of first line therapy[72,111] for any treatment in which there is a TDR in the virus and will lead to treatment failure. Other mutations may then emerge which further restrict treatment options. This has ramifications for individual morbidity and mortality,[154] and at the wider population level. Increased resistance can add to the population viral load and hence increase the force of transmission at the population level, greatly affecting public health outcomes and funding.[127,155]

There are still limitations in knowledge about drug resistant strains of HIV-1, ART usage and viral subtype frequency within the transmitting population.[12] Inadequate education and support to observe, measure and report on treatment adherence, and non-adherence to long-term ART are likely to inflate current prevalence rates for both acquired and primary resistance and create an increased burden financially and economically.

The incorporation of primary and secondary drug resistance into routine surveillance is crucial for understanding the many factors involved in TDR, such as rate of treatment failure occurring in a population, the prevalence of forward TDR from undiagnosed or newly diagnosed individuals, identification of mutations that are most commonly selected, and differences in TDR between subtypes/CRFs. It is therefore imperative that we understand transmitted resistance, both at an individual and global public health level to create targeted prevention and control strategies.[117]

The second aim of the first study was to identify the prevalence of TDR among B and non-B subtype cases newly diagnosed between 2000 and 2013.

Part 3

2.15 Phylogeny

Molecular sequences hold much information about the evolution of life on Earth, including the dynamics of populations and diseases.[156] Innovative sequencing techniques, huge increases in production and availability of digitized molecular sequence databases, and the ever growing number of mathematical and computational analysis tools mean that the quantitative potential and resolution of evolutionary phylogeny is growing rapidly.[156–158]

Phylogeny is the study and predicted ordering of relationships among living organisms based on shared, derived similarities.[159] Phy assesses how biological sequences are

genetically related and estimates hypothetical common ancestors,[158,160] mostly based on DNA or protein sequences. The phylogenetic method assumes differences between sequences arise from mutations and not through other evolutionary processes such as recombination of two or more different strains.[158] Recombinant strains have different genomic regions originating from more than one genetic background and this can affect the phylogenetic structure when calculating the evolution of a tree.[158] The original purpose of phy was to assess and classify relationships between different species, however phylogenetic tools have greatly expanded over the years and now can be used to assess diversity within species.[160] This provides invaluable insight into how the life form, in this study HIV, has evolved over time, where it originated from, where it has travelled to and by what route.[23]

Phy has become the gold standard method for investigating viral transmission, however it is often under-used both clinically and in research, because of the time-consuming nature of alignment and analysis and its perceived complexity. In reality, phy is a powerful tool that can be used to investigate how viruses change within an individual and within a population, viral recombination and subtyping, and transmission dynamics (whether viruses share a common genetic background that arises from the same or similar source), and can answer clinical questions such as whether viral genetics influence infection severity.[158]

HIV phy has largely been based on *pol* sequences collected as part of routine drug resistance testing at the time of diagnosis or before treatment begins. The *pol* gene is a targeted drug treatment site and is therefore under intense selection pressure. This was initially thought to affect phy by grouping unrelated sequences together that carry drug resistance mutations at the same sites.[161] However, a study in 2004 found no difference between phylogenetic trees whether known drug resistance sites were included or excluded from sequences.[162]

Phy provides a method for identifying high risk populations or sexual networks so

prevention, intervention and containment strategies can be developed quickly.[163] By comparing viral diversity of patient sequences, it is possible to identify HIV transmission clusters, defined as groups of individuals who each have viral populations in them that share such a high genetic similarity with the rest of the group that they are likely related by transmission.[164,165] This allows quick identification of transmission occurring regularly, rapidly, or through particular networks. Current phylogenetic methods now have the capability to identify large transmission clusters even in the absence of intermediary sequence samples, for example, several people may sit within one cluster who have been infected by the same person, but that person is yet to be diagnosed (and is therefore not part of the tree). A common example of this is seen when an undiagnosed person in the early stages of infection rapidly transmits the infection to multiple individuals.[161]

An example of an intermediary sample is shown below. Figure 6, adapted from Hartfield, Murall and Alizon (2014),[158] illustrates a viral phylogeny transmission cluster (Figure 7 explains a phylogenetic tree). The top section shows an actual transmission network, with each horizontal line representing an individual patient (p) and each broken vertical line representing a transfer of infection. In the bottom section a phylogenetic tree is displayed with only four of the seven patient samples from the actual transmission network. Each node (dot) represents at least one transmission event which led to the creation of a new virus. In this figure, intermediate infections have not been sampled and therefore the phylogenetic tree is created with the limited information available.[158]

Other considerations when constructing phylogenetic trees include the possibility of large transmission clusters where co-infection is occurring between people who are already HIV infected, and viruses that diversify rapidly and impact on the relatedness to other viruses over time. Therefore, while phy is an important and useful tool to identify transmission dynamics, it can only approximate transmission networks. It cannot prove that any one

patient directly infected another, and cannot infer direction of transmission by distinguishing donor from recipient.[158]

Combining phylogenetic methods with epidemiological, demographic, and behavioral data can identify the spread of viruses over time, geography, transmission route, and population groups. These characteristics can be combined with genetic data in a way that protects individual privacy while providing valuable information about local and global transmission of HIV that can be used for research and clinical care. An example is the reporting of a decrease in the proportion of subtype B infections in developed countries and a corresponding increase in heterosexually acquired non-B infections from overseas.[73,117] Identification of this change can ensure targeted education, prevention and containment strategies, and increase the scope for research of non-B subtypes and their effect on treatment and patient care. Phy can also be useful for contact tracing of undiagnosed infections. Contact tracing is fraught with complications because of the length of time between infection and diagnosis, global population mobility, overseas travel, unknown partners, multiple concurrent partners, and other high risk behaviors that make it difficult to identify the origin of infection and onward transmissions.[161] Patient histories can be inaccurate and phy is a very useful tool to validate or refute epidemiological linkages.[161,163]

To use phy, RNA or DNA sequence data need to be generated. This involves numerous steps such as the design and testing of primers, polymerase chain reaction (PCR) amplification of the chosen genome regions, sequencing of amplified products and the uploading and editing of sequence data into various analysis programs.[157] As mentioned earlier in the chapter, analysis of nucleotide sequence variations of the subgenomic *pol* region encompassing RT and PR is routinely conducted to determine resistance to antiretroviral drugs. Classification of viruses into phylogenetically distinct subtypes and CRFs to track HIV evolution and diversity has largely been based on this region.[3,4,75,166]

However, with the rapidly expanding number of global viral sequence data available in databanks online, phy of HIV subtypes can be based on multiple subgenomic regions, most commonly the *gag*, *pol* and *env* regions.

Improved sequencing methods allow full sequencing across the regions chosen or multiple subgenomic sequencing can be used.[158,167] The use of multiple gene regions has revealed greater genomic subtype diversity than previously reported using one gene alone.[168]

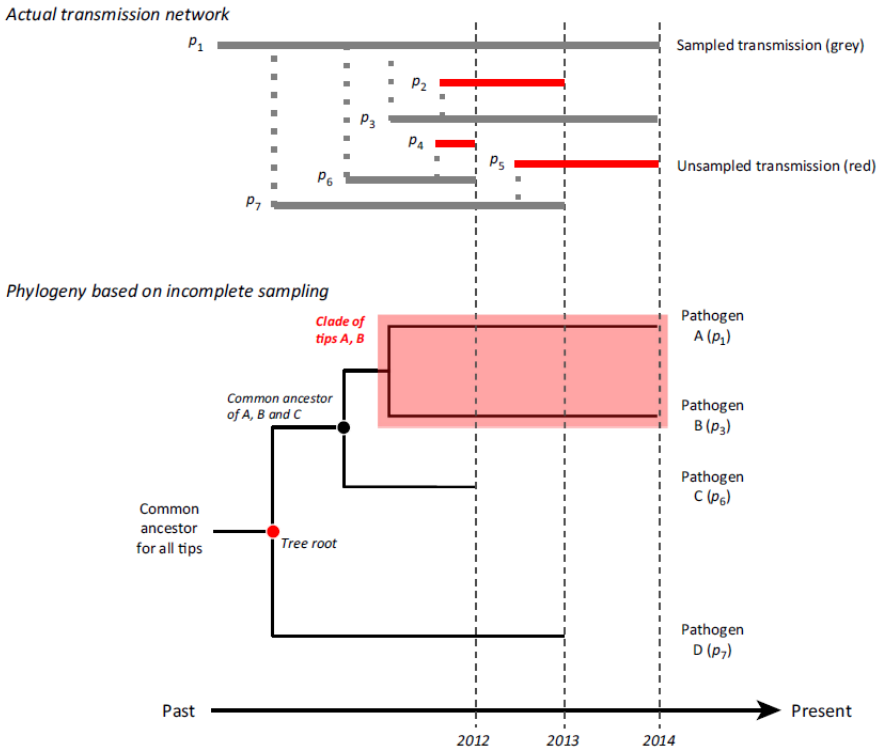


Figure 6. An actual transmission network and an inferred transmission tree. Top: a schematic describing a transmission network, with horizontal lines representing individual patients, and vertical broken lines showing infection. Bottom: how sampling from the transmission network determines the layout of a clinical phylogenetic tree. Patients are denoted p1 to p7; in this example, patients 1, 3, 6, and 7 are sampled to produce the phylogeny, represented as viruses A to D on the tips. Patients 2, 4, and 5 (red lines in the transmission network) are not sampled. The nodes of each branch represent the common ancestor to them. A collection of tips and their nodes is called a clade; the clade for tips A and B is highlighted in red. The root represents the common ancestor for all samples. The horizontal axis is scaled by the degree of genetic change (for example, nucleotide substitutions per site) over time. The vertical axis has no scale; spacing is included purely for clarity. Note that not all tips line up exactly with the present; this is because isolates can be collected at different times, and this information determines the phylogenetic shape. (Hartfield, Murall & Alizon, 2014).[158]

2.16 Steps in phylogeny

Phy can be time-consuming because each step may require a different program with its own file format. Files are often interconverted as they are moved from one program to the next. However, there are some programs that can perform multiple steps, such as the Molecular Evolutionary Genetics Analysis program (MEGA).[160] The four main steps of phy are listed below:

2.16.1 Step 1

The starting point is to identify and acquire a homologous (descended from a common ancestor, in this case, HIV-1 group M virus) RNA or DNA or protein sequence set. Pre-aligned reference sequence sets of whole genomes or partial fragments can be downloaded from online repositories such as GenBank (National Center for Biotechnology Information, NCBI) or LANL BLAST (Los Alamos National Laboratory), or sequences can be downloaded from BLAST (Basic Local Alignment Search Tool) by submitting each query sequence to the tool which then returns the top hits for sequences in the repository that are most similar. The databases can be searched a number of ways; by genomic region, country of origin, subtype, year the sample was taken, or similarity to the query sequence. Once a complete reference set is obtained, it must be combined with the query sequences in a format ready for alignment.

2.16.2 Step 2

Phylogenetic trees are determined based on the number of differences in nucleotide substitutions between them. Therefore the reference and query sequences must be aligned in order to compare them.[158] Each site along the sequence is considered a character, which can be one of the four possible nucleotides: adenine (A), guanine (G), cytosine (C) or thymine (T), or a gap in the sequence resulting from an insertion or deletion.[159] Online programs such as the Cyberinfrastructure for Phylogenetic Research (CIPRES, ≤https://www.phylo.org/≥

or free installation programs such as Bioedit ≤http://www.mbio.ncsu.edu/Bioedit/bioedit.html≥ can be used to align sequences.[160]

The most common method used is a progressive alignment such Clustal W or Clustal X,[169] which aligns sequences in pairs to create a distance matrix based on alignment scores. The matrix is then used to create a Neighbor Joining guide tree that clusters the sequences during the stepwise alignment. Only variable sites are phylogenetically informative so it is imperative that the alignment is correct. Manual analysis should be undertaken to improve the alignment, such as by removing incorrect gaps or shifting specific sequences that are slightly out of alignment.

2.16.3 Step 3

Once the sequence set is aligned, a substitution model that explains how mutations arise between related samples over time must be selected and applied to the sequences, followed by application of a model of evolution. Software such as MEGA have the capability to determine the substitution model and estimate the evolutionary tree, or stand-alone programs such as jModelTest ≤http://jmodeltest.org/user/dashboard≥ can be used to infer the substitution model.[158] There are many programs available to estimate the tree, such as Phylogeny Inference Package (PHYLIP, ≤http://evolution.genetics.washington.edu/phylip.html≥, and MEGA ≤http://www.megasoftware.net/≥.

Sequences on their own are not very informative about viral evolution, rather, the individual site variation between two or more sequences is examined.[170] Viral diversity occurs through mutations, including nucleotide substitutions, insertions, deletions and recombination events that are carried forward into new viruses and which then spread through viral populations by genetic drift or natural selection.[171] To understand viral evolution, assumptions are made about the rate of these mutations and applied to

mathematical models.[169] For example, point mutations may occur either by a *transversion*, when a purine nucleic base (A or G) replaces a pyrimidine base (C, T) or *vice versa*, or a *transition*, when a purine base replaces another purine base or a pyrimidine base replaces another pyrimidine base.[170,171] For HIV sequences the number of transitions is often higher than the rate of transversions.[157] Synonymous and non-synonymous substitutions must also be considered, the former being nucleotide substitutions that do not change the encoded amino acid (often a change at the third codon position) and the latter being substitutions that do change the amino acid, often with a deleterious effect on protein function leading to cell death.[159]

Non-synonymous substitutions are far less frequent and the third codon position is the predominant site of substitution. However, substitution rates can also vary by genomic region, depending on the tolerance level of amino acid change in that particular region. In critical regions such as active protein coding sites, mutations are severely restricted (functional constraint), while in non-critical regions there may be a much higher tolerance. This difference in substitution rates along the genome can lead to major bias in phylogenetic inference, and measures have been developed such as the ***gamma* distribution (γ)** to account for substitution distribution bias.[159,170]

2.16.3.1 Mathematical models of sequence evolution

The first mathematical model was developed in the late 1960s and named the Jukes and Cantor model.[172] It assumes an equal frequency of the four nucleotides in any given sequence and substitutions between all four being equally likely.[170] The most commonly used model for the study of HIV sequences was the Kimura 2 parameter model (K2P) published in 1980,[173] however this model was not found to reflect the reality of nucleotide substitutions so other models such as the Hasegawa, Kishino and Yano (HKY) model

published in 1985[174] and the General time-reversible (GTR) model published in 1986 were developed.[175]

2.16.3.1.1 K2P model

Transitions generally occur at a much higher rate than transversions, even though for any given nucleotide site there are three possible changes: one transition from purine to purine or pyrimidine to pyrimidine, and two possible transversions (either of the purines to a pyrimidine or *vice versa*).[170] In 1980, Kimura developed a nucleotide substitution algorithm that factored in this higher rate of transitions $\alpha + 2\beta$, [173] where α represents the rate of transition at any given site and β the rate of transversion.

2.16.3.1.2 HKY model

This model was first published in 1985 by Hasegawa and colleagues.[174] It combined the K2P model with another model by Felsentein that was published in 1981 (F81 model).[176] The F81 model went a step further than previous models by taking into consideration that certain bases naturally occur more frequently in some organisms than others, such as A and T in insect mitochondrial DNA. Other organisms have a higher proportion of G and C, which creates a more thermodynamically stable DNA structure because of the higher number of molecular triple hydrogen bonds.[157] The F81 algorithm allows for different frequencies of the four nucleotides when calculating the substitution rate.[170] The HKY model used both the K2P and F81 models to create an algorithm that accounts for differences in base frequencies and differences in transition and transversion rates.[170]

2.16.3.1.3 GTR model

The GTR model is the most commonly used model for large datasets. First published by Tavare in 1986, it is regarded as the most general, least biased model.[175] It works on a probability matrix constrained by six parameters in which each possible substitution has its

own probability (A ↔ C, A ↔ G, A ↔ T, C ↔ G, C ↔ T, G ↔ T).[170,175] The GTR model allows for different base frequencies and different rates of substitutions for all six possibilities. This model also allows for preferential change of substitutions in both directions over time,[157] and most frequently reflects the parameters found in sequence data sets.

Distribution of substitution rates such as *gamma* (γ) mentioned earlier, or invariant site distribution (I) can be added to these evolutionary models, the latter takes into consideration the probability of sites that are constrained to be invariant compared with those that are free to vary.[170] For very large datasets such as whole HIV genome sequences, the GTR+I+G model is often the most appropriate.

2.16.3.2 Estimating phylogenetic tress

There are two main methods of converting sequence information into an estimated phylogenetic tree: the distance matrix (clustering) method, and the discrete character based (tree searching) method.[159,170,177]

2.16.3.2.1 Distance based method

Distance based (clustering) methods are based on the proposition that the evolutionary history of a set of sequences can be reconstructed if actual evolutionary distances between the sequences are known. The method works by aligning sequences into a pairwise distance matrix, looking for differences between each pair, and then using that matrix to construct an optimal tree.[159,170]

There are a number of distance measures available that account for bias such as unequal or saturated substitution rates.[159] The most simple distance method is the Unweighted Pair Group Method with Arithmetic Averages (UPGMA), which works by linking the most similar pairs of sequences to form a single cluster that is then treated as a single sequence

and used to calculate distances to other sequences. This is then repeated stepwise with the next pair of similar sequences and so on. Eventually a dendrogram is created and sequences with the closest distances are clustered together on the inferred tree.[157] The scale may be expressed as distance or percent similarity. The tree is rooted at the point where the last two clusters are joined.[159]

Neighbor Joining (NJ) is the most widely used distance based method and it sequentially locates the closest neighbors based on the internal branch lengths of the tree. It is based on cluster analysis like UPGMA but identifies tree nodes rather than sequences or clusters of sequences. It also allows for unequal rates of molecular change among the branches.[159]

The clustering method is fast and easy to use, but does not optimize a criterion of fit between the tree and data. The inferred tree depends on the order in which sequences are added to the growing tree, and the method does not allow for competing hypotheses, that is, it does not allow for multiple trees to be produced that may explain the data equally well.[170]

2.16.3.2.2 Character based method

The character based method works by examining each site column separately and then searching for the tree that best explains the sequence data. The most common character based methods are maximum parsimony (MP), maximum likelihood (ML) and Bayesian. This type of analysis is time consuming but information rich, because a hypothesis is created for every single sequence alignment column (each site in a sequence for all sequences combined) and therefore the evolution of specific sites can be analysed.[157,170]

The MP and ML methods both choose an inferred tree based on the minimum number of changes required to explain the data.[178] The MP and ML methods both add branches in succession at different levels on the tree. These levels are evaluated and the best-fitting tree is chosen before the next branch is added. This is called stepwise addition. Branches can

also be rearranged, called branch swapping. There are a number of techniques to do this: tree bisection and reconstruction, nearest-neighbor interchange (NNI), and subtree pruning and regrafting.[169] However ML and MP methods differ in that the ML method estimates the tree by choosing the tree that is the most likely to occur given the data available, using either prior mutation rate estimates or using the data itself to estimate these parameters. Conversely, MP is a non-parametric method that ignores evolutionary models and identifies the tree with the least number of mutations between sequences.[179]

The Bayesian approach is similar to ML but instead of seeking the tree that maximizes the likelihood of the arrangement of the sequences, this method identifies the tree with the greatest likelihood, combining the prior probability of a tree based on the chosen evolutionary model with the likelihood of the data based on the alignment. The tree that best represents a phylogeny is chosen based on the highest posterior probability distribution.[157]

The MP method ignores branch lengths in building trees and therefore, if there are branches that diverge much more rapidly than others, the parsimony method may lead to incorrect topologies. However, unlike MP, only the best tree is produced from the ML method, so any uncertainty between sequence relationships that might be explained by another likely tree will not be shown.[158,178] Although the ML method is also the most time intensive and requires a model of evolution to reconstruct a phylogenetic tree (i.e. the GTR model) it is the most appealing because of its capacity to incorporate explicit models of sequence evolution and calculate statistical tests of evolutionary hypotheses.[170] Online portals like CIPRES dramatically reduce the computational requirements of sequence alignment and ML analysis through a simple browser interface that allows access to large computational resources ≤https://www.phylo.org≥.

2.16.3.3 Reliability – bootstrapping

To test the robustness of a tree topology, the bootstrapping technique is used. It estimates

the reliability of each interior branch of a tree, that is, how consistently a node is supported and whether the entire dataset supports the inferred tree or the inferred tree is one of many similar alternatives.[177] Bootstrapping works by taking random columns (same site across multiple aligned sequences) and creating a subsample of the dataset with these columns that is the same size as the original sequence length. The same columns can be used more than once in a subsample (random sampling with replacement). These subsamples are used to build trees, and an algorithm is used to calculate the frequency with which various parts of the tree are reproduced from each subsample. For example, if sequences X, Y and Z cluster together in every single subsample tree, the bootstrap value for that cluster would be 100%, in other words, the informative sites agree that this is a group.[180] A number of bootstrap iterations (subsamples) can be run, most commonly between 100 and 1000.[177]

Bootstrapping is considered to be a useful and dependable measure of tree accuracy. Significance of cluster support is usually expressed as a percentage, and 70% is widely considered to imply the cluster is well supported.[159,177] For clusters likely related by transmission, a bootstrap value of 98% or more is widely considered to represent a true linkage between sequences, as does an intra-cluster genetic distance of 1.5% or less from the nearest neighbor in the cluster.[161]

2.16.4 Step 4

Tree output needs to be displayed in a way that clearly shows the information to an audience. There are several programs available to draw trees for analysis and publication, including MEGA and Phylodendron ≤http://iubio.bio.indiana.edu/treeapp/treeprint-form.html≥.[160] A dendrogram consists of branches and nodes. The nodes reflect a common ancestor between two or more sequences which are represented as branches that diverge from the nodes. A phylogenetic tree is also known as an additive tree and is usually displayed so the length of each branch corresponds to the amount of divergence from the common

ancestor node. The shorter the branches, the less divergent the sequences are from the other sequences connected to it by that node.

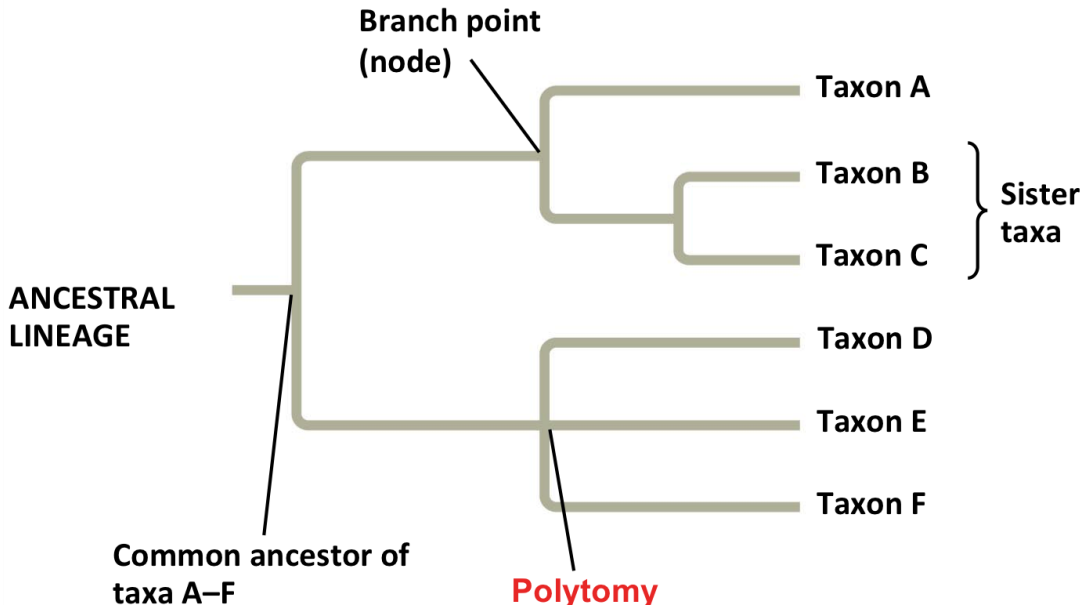


Figure 7. A phylogenetic tree. The branch point or node represents the common inferred ancestor of the branches it connects, the further back (to the left) in the tree the older the inferred ancestral strain. Taxon refers to a group of one or more populations of an organism, for HIV trees each taxa represents a DNA sequence from one person. Two sequences that split from the same node are called sister taxa. Polytomy occurs when more than two branches split off from a common node. (Pearson Education Inc).[181]

Phylogenetic trees can be rooted, with the root representing the oldest common ancestor within that tree, often as an outgroup or an external point of reference rather than as a member of the sequence group. For example, for HIV-1 M group sequences, a HIV-2 outgroup may be used to root the tree.[169] However rooting the tree with an outgroup can be problematic. If the outgroup is too distantly related the sequences can become randomized and cause 'long branch effects' which can lead to artificial rooting along ingroup branches.[159] Often tree rooting is used to ensure all ingroup sequences (query and reference)

cluster together with the outgroup sitting separately on the tree. The outgroup can then be deleted and the analysis performed with only the ingroup sequences.

Through understanding the way sequences are related to each other and using phy to infer the evolution of HIV at a population level, preventative and intervention efforts can be improved to target specific regions, groups and virus types.[165] With phy we can identify viruses that are similar and discover the level of variance between viruses within subtype/CRF groups, identify possible transmission clusters including demographic and epidemiological characteristics, and gain greater understanding of the overall genetic history of HIV.

2.17 Online subtyping tools

As mentioned previously, phy is the gold standard for determining sequence subtype or CRF.[182] Phylogeny is quite complex, however, and is not widely implemented globally. Rapid online subtyping tools are being developed and used as an alternative. These are often free to use and provide very fast results.[182]

A number of mathematical models underpin the capability of rapid online tools to classify viral strains into subtypes, CRFs or more complex recombinant strains,[157] although there is no universally agreed upon approach. The models are broadly categorized into those that do and do not subtype using phylogeny, whether or not the model performs or requires a sequence alignment, and the automation process (full, partial, none).[183]

2.17.1 Fully automated tools

Automated online tools approximate a subtype assignment because identification of a subtype reveals the clade, which is identified through phylogeny. These approximation tools have been developed because of the growing need for fast results and the issues that arise with increasing viral diversity which make it difficult to identify recombinants using

a phylogenetic tree.[183] However, these automated methods are not without issue, and often these tools cannot identify unique or complex mosaics, and may over- or under-identify recombinants. Some tools also have high levels of discordance with results from phylogenetic analyses, which is cause for concern, Figure 8.

There are three types of fully automated tools: (1) phylogenetic-based which use a phylogeny and alignment based algorithm such as REGA[184] and SCUEAL,[183] that allow detection and identification of recombinant strains with added bootstrap support (REGA) or the phylogenetic likelihood of a mosaic (SCUEAL),[185] (2) similarity-based which are phylogeny and/or alignment free, such as the sliding window analysis of the LANL BLAST tool, or the Stanford CPR tool,[186,187] or (3) statistical-based tools that use partial matching compression algorithms such as the COMET tool[188] or employ a probabilistic jumping alignment approach that uses a Hidden Markov Model to align the query sequence to the most similar reference sequence, such as jpHMM.[189] These online tools are described in detail in section 3.3.14 of Chapter Three.

Of the automated tools that rely on an initial alignment step to determine similarity with the reference set, some can also identify complex recombinant forms by conducting phylogenetic analyses, either by using a sliding window approach with bootstrap support or phylogenetic likelihood of a mosaic.[188] An initial phylogenetic bootstrap analysis is run to assess the query sequence against a given reference set to infer breakpoint locations that are then confirmed by detailed phy of pure subtype and known CRF reference sequences.[183] The best match returned is designated as the putative subtype or CRF.[188] These tools often enable the user to adjust parameters such as the sliding window size and step size, reference sequences, and alignment parameters. While this can improve detection of recombination events, it can also over-identify events or lead to ambiguous results.

One of the issues with automated tools requiring alignment is that the one alignment chosen may not be statistically distinct from other alternative alignments. Alignment is also affected by the choice of algorithm and the parameters set (default or user defined). Alignment confidence is measurable but this can be time consuming, especially when the query sequence is a complex mosaic. A conserved gene region such as *pol* may not be so affected by these factors as a more variable region such as *env*.[190]

Other online subtyping methods do not require an initial alignment because they work on the premise that sequence similarity can be measured by data compression, which identifies the amount of information shared between two sequences.[188] Examples of alignment-free models include the sliding window analysis used by LANL BLAST[186], and Markov models that identify the probability of a given nucleotide at the next site of a sequence by genomic context by looking at the last *k* nucleotides in the sequence. The two sequences are then compared by measuring the context-specific base frequencies in the second sequence to the Markov model built from the first sequence.[188,189] Two alignment-free tools that use Markov models are jpHMM[189] and the COMET tool, which is based on the Prediction by Partial Matching compression algorithm (PPM). All models are discussed in section 3.3.14, Chapter Three.[191]

2.17.2 Reliability of rapid subtyping tools

There are limitations when using any rapid online tool. Some tools have limited capability in which genomic regions they can analyze, with many only analyzing the *pol* fragment.[182] The major limitation is the level of subtype discordance found between tools, especially those with different algorithms underpinning the method, Figure 8. There is general consensus in the literature that rapid automated tools are highly sensitive in identifying subtype B sequences[182] but false subtype B can be identified in non-B sequences. Sensitivity decreases dramatically for the majority of rapid online tools for

identification of CRFs other than 01_AE and 02_AG, and decreases even further for more complex mosaics.[182,192,193] When comparing tools, the overall subtype distribution for a cohort can be quite different according to which tool is used. This makes it very difficult to ascertain the molecular epidemiology of HIV globally when there is no gold standard tool uniformly used worldwide.[182,192,193] In 2009, Kosakovsky Pond *et al*.[183] described a new algorithm, asserting adoption of a phylogeny-based method was imperative to subtype sequences accurately - SCUEAL. It is considered a reliable and robust subtyping tool that not only automatically subtypes HIV-1 *pol* sequences but which can also map recombination breakpoints and assign parental sequences in inter- and intra-recombinant strains (with confidence levels), using the most current LANL BLAST reference sequences, including pure subtypes and CRFs. The method is specifically designed to include up-to-date CRF reference strains for identification of not only CRFs, but also more complex recombinant mosaics.[183] SCUEAL is discussed in section 3.3.14.6, Chapter Three.

HIV-1 Variant (no.)	PHY	Stanford	Rega
Subtype B (587)	87.6	87.3	75.2
"Pure" non-B subtypes (13)	1.9	2.2	3.9
CRF02_AG (31)	4.6	4.3	3.3
CRF non-02 (16)	2.4	0.6	0.4
URF (23)	3.4	5.5	-
Not Determined	-	-	17.2

Figure 8. Distribution of HIV-1 variants according to subtype method. Phy is considered the gold standard method to classify HIV-1. **Key**: no. (number of sequences), CRF (circulating recombinant form), URF (unique recombinant forms). The automated tools included were: Stanford CPR v6.0.5 and REGA v2.0. High concordance between phy and Stanford CPR for identifying subtype B, but low concordance across all three methods for non-B pure subtypes and CRFs. REGA could not identify almost one-fifth of sequences. (Adapted from Yebra et al. 2011).[182]

The growing number of HIV recombinant viruses has made correct identification of strains increasingly complicated.[194] Phy is currently the gold standard for identifying subtypes and CRFs but is time consuming and complex. Online subtyping tools are used increasingly to identify strains, and although they are fast and easy to use, there are limitations, especially with identifying non-B variants and complex recombinant forms.[182,195] The phylogenetic and online subtyping tool study aimed to extract epidemiological information about the evolutionary relationships of newly diagnosed virus in South Australia between 2000 and 2012, and to assess the usefulness of phy and rapid subtyping tools for identifying subtype B, non-B and complex recombinant sequences from two separate regions of the HIV genome, given viral diversity is increasing in number and complexity worldwide.[164,192]

AIMS OF THE STUDY

A. Investigate the molecular and social epidemiology of HIV-1 in South Australia between 2000 and 2013
B. Characterize changes in HIV-1 genetic diversity and TDR mutations in South Australia between 2000 and 2013
C. Investigate and characterize phylogenetic characteristics of HIV-1 *pol* and *env* gene sequences
D. Characterize sequences that form part of high reliability clusters and transmission clusters
E. Compare phylogenetic and online subtyping analysis of *pol* and *env* gene sequences to determine level of concordance between different tools, and if there is a higher degree of complexity in subtype distribution than previously reported.
F. Assess the prevalence of intergene and intragene variation within the study population

CHAPTER 3: METHODOLOGY & RESEARCH DESIGN

3.1 Overview

The methodology and research design are described in this chapter. The background for each study is given, followed by the design, materials and sampling methodology, and the experimental procedures used for epidemiological subtype surveillance, drug resistance surveillance, and phylogenetic characterization of subgenomic sequences.

3.2 Study 1 - Molecular Epidemiology and Drug Resistance Surveillance

3.2.1 Background

HIV strains are evolving rapidly, with more and more recombinant strains emerging globally.[66] There is growing interest in understanding the evolution of HIV strains, and how subtype distribution patterns are changing over time.[9,70,72–74] Research on the importance of TDR surveillance and the increasing prevalence of TDR globally is expanding.[12,72,111,196] There are still significant gaps, however, in knowledge and understanding of HIV-1 strain variation, subtype and TDR surveillance in Australia.

Subtype and drug resistance surveillance of HIV-1 has been routinely conducted in the US, parts of Europe, and more recently in some Australian states.[66] However, while there was a wealth of analysis and discussion of overseas surveillance data, until 2010 only two molecular epidemiology studies had been conducted in Australia to analyze surveillance data.[71,197] The rationale for this was that until recently, the Australian epidemic had been largely composed of one particular strain (subtype B), and predominantly transmitted through one route (MSM).[66]

In 2000, South Australia became the first state to conduct routine genotypic and drug resistance testing as part of an enhanced surveillance system. However, these data have not been used for surveillance research and this provided the opportunity to assess 11 years of HIV-1 data to analyze subtype and drug resistance trends. The belief behind this was that

Australia's HIV-1 epidemic may be changing, mirroring global patterns of increased HIV strain variation because of an increase in Australian born people travelling overseas, steady growth of migrant populations, and increased sexual contact between different population groups. It was also thought that the pattern of drug resistance mutations may be changing, with more people transmitting resistant virus to treatment naïve populations.

3.2.2 Accessing HIV-1 notification and subtype data

Access to the required data was granted by SA Health, which maintains the South Australian HIV notification and subtype databases. The notification database contains all HIV notifications in South Australia including newly diagnosed cases and cases diagnosed outside South Australia. Demographic and clinical information is stored for each patient. The subtype database contains surveillance information from routine drug resistance testing, which is used to guide clinical management. Subtypes for the PR and RT regions of the *pol* gene are recorded, determined by submitting a 1098bp *pol* sequence to the Stanford CPR Drug Resistance online genotyping tool ≤http://sierra2.Stanford CPR.edu/sierra/servlet/JSierra≥, which also lists drug resistance mutations as determined by the 2009 Stanford surveillance resistance mutation list. Some demographic information is also recorded in this database.

In 2010, a detailed ethics submission was sent to SA Health and Flinders University seeking to access these data, which were provided by an independent custodian for the purpose of surveillance analysis.

3.2.3 Ethics

The ethics submission dealt in detail with confidentiality issues arising from using patient data and the ways in which the information would be used. The main issues were ethical considerations around Aboriginal/Torres Strait Islander people, protection of patient confidentiality including how information would be presented to protect patient

confidentiality, the security of the server on which the information would be housed, and the nature of public health messages that would be conveyed about different population groups with HIV. All questions raised by the ethics committees were responded to and the application was re-submitted. The submission required a full literature review and research proposal that was independently peer reviewed. Following a favorable peer review, ethics approval was granted in 2011, allowing the study, analysis and publication of 11 years of subtype and drug resistance surveillance data in South Australia.

3.2.4 Study population and design

Routine AIDS and HIV notification commenced in South Australia in 1985 and 1991, respectively, to facilitate contact tracing and treatment intervention. A standard form containing demographic, epidemiological, and clinical information is completed for each person newly diagnosed in South Australia,. Where possible, an in-depth interview is also conducted.

As mentioned previously, routine subtype and drug resistance testing commenced in 2000, conducted by SA Pathology. Sequences and subtype information are housed at SA Pathology and sent to SA Health where they are added to the subtype database and linked to the notification system by the patient number. Once ethics approval was granted, the independent custodian combined the notification and subtype databases to create a new dataset that linked patient subtype, demographic and clinical information. Case identifiers were removed by the custodian but limited demographic, epidemiological, and clinical data were retained, shown in Table 1 below. Initially, the combined dataset contained all new diagnoses between 2000 and 2010, but after an amendment to the ethics approval this was expanded to include the years 2011–2013, and the dataset was updated in January 2014 to include all new diagnoses between 2000 and 2013.

The Stanford HIV Drug Resistance Database houses the most up-to-date information on

resistance mutations and subtypes ≤http://hivdb.Stanford.edu/pages/surveillance.html≥. The 2009 surveillance resistance mutation list (the most current at the time) was used for the initial dataset analysis in 2011 to create a catalogue system within the newly created database, and it was also used for continuity when the database was updated with 2011–2013 data. To create the catalogue system each case record was checked manually, and each mutation was checked for removal (mutations reclassified by Stanford as having little to no effect on drug susceptibility, polymorphic mutations commonly found in untreated people) or reclassification (mutations that had been major mutations that Stanford reclassified to minor mutations). New variables were added to define how many resistance mutations each person carried and to which drug classes, and these were used for further analysis. The new classification system was later used by the custodian to update the state HIV subtype database.

The entire dataset was then manually checked for inconsistencies to ensure no data were missing and case demographic data were correct. There were 150 case records with missing subtype information and for 90 of these the information was located manually through SA Pathology and the database updated. This audit of the database took considerable time because of the number of individual variables, the time taken to request and receive individual case records (minus identifiers) to cross-reference with the database, and the time taken to check multiple drug resistance mutations for each individual with resistance, against the Stanford surveillance mutation list.

Table 1. Case information housed on the surveillance database

Full Description	Stata Code	N	Missing	Full Description	Stata Code	N	Missing
HIV NOTIFICATION INFORMATION				PI major 4	pimajor4	9	504
Patient No	patientno	513	0	PI major 5	pimajor5	6	507
Date of most recent rest	lasttested	512	1	PI major 6	pimajor6	2	511
Most recent viral load	vload	497	16	Protease minor	piminor	513	0
HIV status	status	513	0	PI minor multiple	mi_multi	513	0
Date of specimen collection	testdate	513	0	PI minor 1	piminor1	**156**	357
New case notified in SA only	newcase	513	0	PI minor 2	piminor2	108	405
EPIDEMIOLOGICAL INFORMATION				PI minor 3	piminor3	55	458
Marital status	ms	513	0	PI minor 4	piminor4	33	480
Racial origin	ro	513	0	PI minor 5	piminor5	13	500
Birthdate	birthdate	512	1	PI minor 6	piminor6	3	510
Age groups	age25	513	0	PI minor 7	piminor7	2	511
Age groups	age30	513	0	PI minor 8	piminor8	1	512
Age groups	agrp	513	0	PI minor 9	piminor9	0	513
Age groups	agrqu	513	0	PI minor 10	piminor10	0	513
Age groups	age50	513	0	NRTI resistance mutations	nrti	513	0
Age groups	agrp1	513	0	NRTI multiple	nr_multi	70	443
Age groups	agrp2	513	0	NRTI 1	nrti1	70	443
Post code	postcode	462	51	NRTI 2	nrti2	41	472
Suburb	suburb	448	65	NRTI 3	nrti3	22	491
Gender	sex	513	0	NRTI 4	nrti4	15	498
HIV country of birth	hcob	423	90	NRTI 5	nrti5	7	506
Location infection acquired	location	513	0	NRTI 6	nrti6	4	509
Previous test	ptest	513	0	NRTI 7	nrti7	3	510
Date of previous test	ptestdate	264	249	NRTI 8	nrti8	2	511
Testing history	th	513	0	NRTI 9	nrti9	1	512
Likely mode of infection	mode	513	0	NRTI 10	nrti10	1	512
Sex of the source	sourcesex	513	0	NRTI 11	nrti11	1	512
CLINICAL INFORMATION				NRTI 12	nrti12	0	513
Stage of infection at diagnosis	stage	513	0	NRTI 13	nrti13	0	513
CD4 count at diagnosis (grouped)	cd42010	513	0	NRTI 14	nrti14	0	513
Viral load at diagnosis	vloaddx	512	1	NRTI 15	nrti15	0	513
Current or past blood donor	donor	407	106	NNRTI resistance mutation	nnrti	513	0
PEP 04/07	pep	513	0	NNRTI multiple	nn_multi	513	0
Datepep	datepep	177	336	NNRTI 1	nnrti1	49	464
CD4 count (grouped)	cd42010	513	0	NNRTI 2	nnrti2	18	495
RISK FACTOR INFORMATION				NNRTI 3	nnrti3	4	509
Risk factor	risk	513	0	NNRTI 4	nnrti4	3	510
Heterosexual overseas	hetoseas	99	414	NNRTI 5	nnrti5	3	510
Immigration status	istat	513	0	NNRTI 6	nnrti6	3	510
Sero conversion illness	sero	477	36	NNRTI 7	nnrti7	2	511
Sero conversion date	serodate	510	3	NNRTI 8	nnrti8	1	512
Notifying source	ns	513	0	NNRTI 9	nnrti9	0	513
RISK EXPOSURE INFORMATION				Genotyping recent	rgt	382	131
Investigation status	is	511	2	Date of recent genotyping	recentdate	51	462
Sexual partners exposure	partner	513	0	Sequence ID	idseq	51	462
Anal sex	analsex	513	0	Recent viral load	rvl	382	131
Known partner	knownpartn	451	62	Recent PI major (Y/N) ?	rpima	371	142
History of IDU	drugs	478	35	Recent PI minor (Y/N) ?	rpimi	371	142
Shared injecting equipment	shared	470	43	Recent NRTI	rnrti	368	145
Blood transfusion	bloodtrans	477	36	Recent nnrti	rnnrtic	14	499
Date of blood transfusion	btdate	513	0	Recent PI major (list of mutations)	rpimac	6	507
GENOTYPE				Recent PI minor (list of mutations)	rpimic	8	505
Lab Number	labn	382	131	***NEW VARIABLES CREATED FOR STUDY***			
Date tested	gtest	382	131	Treatment status	treat	513	0
Viral load	genovl	382	131	Been genotyped	genodb	513	0
Hiv diagnosis	dx	382	131	Age at diagnosis	ageyears	512	1
Diagnosis	dxwhere	513	0	Year HIV (same as var yr)	yrhiv	513	0
Protease resistance (clade)	pr	513	0	Month HIV	mthiv	513	0
Reverse transcriptase	rt	513	0	Quarter HIV	qtr	513	0
Protease major	pimajor	513	0	Aids diagnosis	adx	64	449
PI mjaor multiple	ma_multi	513	0	hiv death?	hdth	20	493
PI major 1	pimajor1	77	436	aids death?	adth	9	504
PI major 2	pimajor2	53	460	Notifying source	notsr	513	0
PI major 3	pimajor3	22	491	Region from	region	513	0
				Year diagnosed with HIV	yr	513	0

Key: Stata code (code used in statistical program Stata that is linked to variable name), N (number of cases available), Missing (number of cases with information missing for that variable).

This study was a retrospective molecular epidemiological analysis of HIV-1 in South Australia. Inclusion criteria for selection were: newly diagnosed case in South Australia between 2000 and 2013 with a plasma-derived RNA *pol* sequence available for genotyping

and drug resistance profiling taken within 12 months of diagnosis. There was an additional inclusion criterion for the drug resistance surveillance, namely people had to have a 'treatment naïve' status at the time the blood specimen was taken for genotyping/drug resistance profiling. To reduce bias all people meeting the criteria were included in the analyses.

A total of 656 people newly diagnosed with HIV between 2000 and 2013 was identified from the South Australian HIV notification database and 569 were retrospectively selected from this dataset according to the inclusion criteria regarding subtype. Of these 569 cases, 496 were selected for TDR analysis because of their treatment naïve status.

Subtype was previously determined and stored in the database, using the calibrated population resistance (CPR) tool linked to the HIV-1 Drug Resistance Database (Stanford University, Palo Alto, CA), available at ≤http://hivdb.Stanford.edu≥.[198] This tool also provides a genotypic drug resistance profile. It analyses a contiguous 1098bp sequence spanning the PR and RT regions of the *pol* gene and classifies subtype of the PR and RT regions separately, using all the pure subtype reference sequences and CRF01_AE and 02_AG. It does not accurately detect recombinants except 01_AE and 02_AG.[199]

If the PR and RT sequences were phylogenetically concordant, the case was assigned as that particular pure subtype or CRF. If the PR and RT sequences were phylogenetically discordant, the case was assigned as a *pol* intersubtype recombinant (*pol*-ISR). If a PR or RT region was missing, the case was assigned as the subtype or CRF of the region available.

The CPR was also used to determined drug resistance mutations by reading the sequences and comparing them to an updated list of surveillance drug resistance mutations (SDRMs) to compute the prevalence of resistance to each of the three main classes of antiretroviral drug: protease inhibitors (PIs), nucleoside reverse transcriptase inhibitors (NRTIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs).[199]

3.3 Study 2 - Phylogenetic Characterization of Partial *pol* and *env* Sequences

3.3.1 Background

During the course of the first study, 20 cases were found to have phylogenetically discordant PR and RT sequences. These cases were mainly acquired overseas and the people had been born overseas. These divergent sequences within one gene sequence (*pol*) led to the hypothesis that HIV infections could contain more complex recombinants than previously shown through the Stanford genotyping capabilities of one gene. An international study confirmed that sequencing and phy of more than one gene showed a higher degree of recombination.[168] Phylogenetic tree construction and the use of multiple online genotyping tools can provide greater precision when identifying subtypes.

It was also apparent from the first study that epidemiological data collected at time of diagnosis and during follow-up appointments can provide valuable insights about the way HIV is transmitted in the population, and can give an indication of the population groups most affected. However, most epidemiological data are collected through self-reporting and case studies, and may not always be accurate. Phylogenetic tree construction of HIV sequence data can provide biological evidence to compare with epidemiological data, such as similarity between sequences and whether they may be part of a transmission pair or cluster.

The aim of this second study therefore was to sequence a second gene on the HIV-1 genome (*env*), and conduct phy of the *pol* and *env* regions to

a) assess the number of highly reliable sequence clusters and transmission clusters and determine whether these data matched corresponding epidemiological data, and
b) compare phylogenetic tree construction using a variety of online tools including Stanford CPR, to assess whether there was a higher degree of recombination complexity in South Australian diagnosed cases than identified by Stanford CPR

alone, and whether any unique recombinant forms were present in the cohort.

3.3.2 Ethics

A second ethics proposal was submitted in 2012, requesting permission to access stored plasma samples held by the custodian at SA Pathology to conduct further gene sequencing and phy of the gp41 region of the *env* gene and compare the previously collected *pol*-PR/RT data with the newly created *env*-gp41 data. This was approved in 2012. A request for variation to ethics approval was then submitted in 2013 seeking permission to combine the surveillance and demographic data from the HIV notification database with the *pol* and *env* data obtained from SA Pathology, to conduct phy of HIV strain variation over time. This was granted in 2013.

3.3.3 Study population and design

Beginning in March 2012, plasma samples stored at SA Pathology were located within the archived and current sample storage systems. This required identification of eligible cases so that the availability and approximate location of all samples associated with that case name could be compiled into a list. Archived samples were frequently located by section or box only, and only identifiable by sample number and name, so all stored samples in that region or box had to be individually assessed to locate the relevant samples for the cases of interest. Where possible, multiple samples taken from each case at different time points were collected to verify accuracy of phylogenetic tree construction and online subtyping tools. The inclusion criterion was a new South Australian diagnosis between 2000 and 2013 inclusive. Where possible we collected the original plasma sample that was used to subtype the *pol* region at time of diagnosis, or a sample obtained as close to that time as possible. In total, 332 plasma samples were found from 293 cases who were diagnosed as HIV positive between 2000 and 2013, and for whom *pol* genetic sequence data were available. Of these samples, 233 (from 221 cases) were successfully used to amplify and sequence the *env*-gp41

gene sequence.

In brief, samples were extracted then a subgenomic region of the *env* gene incorporating gp41 (7816 to 8344 on the HXB2 reference sequence ≤http://www.hiv.LANL BLAST.gov/content/sequence/HIV/mainpage.html≥ was amplified by PCR and sequenced. Multiple alignments were performed for both the *pol* and *env* query sequences datasets, using two methods:

1. **Alignment using LANL BLAST reference sequences**
    a. Input each query sequence in LANL BLAST ≤http://www.hiv.LANL BLAST.gov/content/sequence/BASIC_BLAST/≥
    b. Download the 10 most similar reference sequences for each query sequence, add to FASTA file.
    c. Run an alignment for the separate *pol* and *env* datasets.
    d. Cut each reference sequence to the size of the 1098 *pol* sequence and 530 *env* sequence respectively. Re-run alignment.
2. **Alignment using the 2012 LANL BLAST RIP reference sequence set**
    a. Download complete 2012 RIP reference sequence set ≤http://www.hiv.LANL BLAST.gov/content/sequence/NEWALIGN/align.html≥ for *pol* and *env* sequences, input the HXB2 start and end coordinates that match the query sequence location.
    b. Add to query sequences. Perform an alignment.

In total, four alignments were created, two *pol* alignments (one with LANL BLAST reference sequences, one with RIP reference sequences) and two *env* alignments. A maximum likelihood phylogenetic (phy) tree was constructed from *pol* and *env* alignments (both LANL BLAST reference sequence alignments, and RIP reference sequence alignment) using general time reversal substitution model with inverse *gamma* distribution (GTR+G+I) and 1000 bootstrapped datasets, using the Molecular Evolutionary Genetics Analysis software version 6 (MEGA 6).[160] Genotyping and cluster analysis was performed

using the phylogenetic trees, and genotyping was also done using multiple online subtyping tools (Stanford CPR, jpHMM, REGA, SCUEAL, NCBI, and COMET), and compared. Any *pol* or *env* samples found to be recombinant were further analyzed with SCUEAL to determine recombination breakpoints. Clusters were defined as sequences that had a common bootstrap value at the shared ancestral node ≥70% (high reliability clusters) or ≥98% and average genetic distance less than or equal to 0.015 (1.5%) nucleotide substitutions per site (transmission clusters).

3.3.4 Chemicals, materials, equipment and programs

Chemicals, primers, materials, equipment and computer programs used are listed in Tables 2–5. All chemicals were of analytical grade or higher. The symbols ® and TM indicate that the particular product names are either registered trademarks or trademarks of suppliers.

Table 2. Chemicals and commercial products used

Product	**Dilution**	**Supplier**	**Location**
QIA Amp Viral RNA Mini Kit		QIAGEN Australia	Victoria, Australia
SuperScript™ III One-Step RT-PCR System with Platinum® Taq High Fidelity		Invitrogen Life Technologies,	California, USA
Ultra-Pure molecular grade Agarose		Life Technologies	Victoria, Australia
100bp DNA ladder	1:10	New England Biolabs	Massachusetts, USA
EZ vision 3 loading dye (N313-1ML)	1:10	Amresco	Ohio, USA
QIA quick purification kit		QIAGEN Australia	Victoria, Australia
Big Dye® Xterminator purification kit		Life Technologies	Victoria, Australia
Sequencing primer	3.2pmol/ reaction	Invitrogen Life Technologies,	California, USA

Table 3. Primers used for amplification and sequencing of *pol*-PR/RT & *env*-gp41

Gene fragment	Primers	Oligonucleotide sequences	F/R
PR	GF1	5'-AAGAAGGGGGGCACATAGC-3'	F
	GF2	5'-CTAGRAAAAARGGYTGTTGGAAATGTG-3'	F
	GF2m	5'-GCTAATTTTTTAGGGAARATYTGGCCTTCC-3'	F
	PF6	5'-CAGACCAGAGCCAACAGCC-3'	F
	PF3	5'-AGCAGGAGCCGATAGACAAG-3'	F
	PF2	5'-CTCCCYCTCAGAAGCAG-3'	F
	PR4	5'-GCAAATACTGGAGTATTGTATGG-3'	R
	PR4m2	5'-ARTCYTGAGTTCTYTTATTRAGYTCYCTRAAATC-3'	R
	PR2	5'-GGCCATCCATTCCTG-3'	R
RT	RF1	5'-TGATAGGGGGAATTGGAGG-3'	F
	RF1m	5'-CCAAAAATGATAGGGGGAATTGGAG-3'	F
	RF3	5'-TTAAAGCCAGGAATGGATG -3'	F
	RF6	5'-CCATAYAATACTCCAGTATTTGC-3'	F
	RF2	5'-GAAAGGATCACCAGCAATATTCC-3'	F
	RR4	5'-AGCTGTCTTTTTCTGGCAG-3'	R
	RR2	5'-CCTGTTTTCTGCCARTTC-3'	R
	RR5	5'-GTGCTTTGGYYCCCCTRAGGAGTTTAC-3'	R
	RR8	5'-GAATCCAGGTGGCTTG-3'	R
	RR10a	5'-GTTTTCTGCTAGTTCTAGCTCTGCTTC-3'	R
	RR10b	5'-GTTCTCTGCCAATTCTAATTCTGCTTC-3'	R
gp41	*ENV*F2	5'-GAGCAGCAGGAAGCACTATG-3'	F
	*ENV*R1M	5'-GTGAGTATCCCTGCCTAAC-3'	R

Key: F: Forward primer, R: Reverse primer, *env*: *Env*elope, PR: Protease, RT: Reverse transcriptase. *Pol* and *env* primer design was based on diagnostic needs and performed by the SA Pathology laboratory.

Table 4. Equipment

Equipment	**Supplier**	**Location**
Micro Centrifuge 5424	Eppendorf	Hamburg, Germany
MJ Research PTC-200 thermal cycler	BIO-RAD	California, USA
Electrophoretic Power pack	BIO-RAD	California, USA
Heidolph Reax-top vortex mixer	Heidolph	Schwabach, Germany
48- capillary 3730 DNA Analyser	Life Technologies	Victoria, Australia
Nikon digital camera	Nikon	Tokyo, Japan
Corbett CAS-2400 Liquid Handling System	Corbett Life Science/QIAGEN	Victoria, Australia

Table 5. Software packages

Software package	Licensed company	Location
Stata v.11.1	StataCorp	Texas, USA
Kodon 2.3 & 3.6	Applied Maths	Sint-Martens-Latem, Belgium
Molecular Evolutionary Genetics Analysis (MEGA) v.5 & v.6	Pennsylvania State University	Pennsylvania, USA
Recombination Information Program (RIP) v.3	Los Alamos National Laboratory (LANL BLAST)	Los Alamos, USA
Stanford HIVdb v.7	Stanford University	California, USA
Bioedit v.7	Ibis Biosciences	Illinois, USA
CIPRES	CIPRES	≤http://www.phylo.org≥
NCBI Genotyping	NCBI	Maryland, USA
SCUEAL v.1	Datamonkey, University of California	California, USA
COMET v.1	Max Planck Institute for Informatics	Saarbrucken, Germany
REGA v.2	REGA group	Oxford, UK
jpHMM v.1	University of Göttingen	Göttingen, Germany

3.3.5 Specimen collection and storage

Specimens used for this study had previously been collected by medical practitioners and stored at SA Pathology. Blood was collected in sterile tubes using EDTA as the anticoagulant and stored at 2–25°C for no longer than 24 hours. Plasma was separated from whole blood within 24 hours of collection by centrifugation at 800–1600g for 20 minutes at room temperature. Plasma was transferred to sterile polypropylene tubes and stored in 0.5–1.0mL aliquots at room temperature for up to one day, or 2–8°C for up to 5 days, or frozen at –20°C.

3.3.6 Extraction of HIV for gene sequencing

Samples were handled in a biohazard hood, using sterile filtered tips and powder free gloves. Samples with a viral load below 1000 copies/mL were not tested because previous attempts for routine genotyping had proved problematic and cost ineffective. For samples

with a viral load between 1000 and 10,000 copies/ml, 500µL was taken from each thawed sample and centrifuged at 4°C at 21,000–25,000g (16,000 rpm), for 60–75 mins. The upper 300 µL of serum was removed and discarded, and the remaining 200µL kept for extraction.

For concentrated samples (above) or samples with a viral load above10,000 copies/mL, 200µL plasma was placed into a labelled 2ml Sarstedt tube and extracted using the QIA Amp Viral RNA Mini Kit, following the QIAGEN manufacturer's method ≤http://www.euresist.org/c/document_library/get_file?uuid=82975600-9030-40ab-b298-32a7198e1671&groupId=85965≧. Briefly, the viral lysis buffer (AVL) was placed in a 56°C water bath for 10 minutes until resuspended and then cooled to room temperature. Next, 800µL of AVL buffer was added to 200µL of sample, vortexed for 10 seconds then incubated at room temperature for 10 minutes. Then, 800µL of 100% ethanol was then added to each tube and vortexed at maximum speed for 3–5 seconds, 600µL of the plasma/ethanol mix was added to the top of a labelled spin column, and centrifuged at room temp at 8000 rpm for 1 minute. The collection tube was changed, and this process was repeated twice, until all the plasma/ethanol mix had been spun through the column.

Once complete, 500µL of wash buffer 1 (AW1) was added to the spin column and centrifuged at 8000 rpm for 1 minute, and the collection tube changed again. Next, 500µL of wash buffer 2 (AW2) was added to the spin column and centrifuged at maximum speed (8000 rpm) for 3 minutes, changing collection tube after and centrifuging again at maximum speed (8000 rpm) for a further minute.

The spin column was then placed into a labelled, 1.5mL Sarstedt tube and 50µL of elution buffer added directly onto the spin column membrane, left at room temperature for 1 minute to dissolve the RNA, then centrifuged at 8000 rpm for 1 minute. The spin column was then discarded and the RNA sample stored at –20°C if not used immediately.

3.3.7 PCR amplification and sequencing of partial pol and env fragments

Pol gene PCR amplification and gene sequencing was previously performed by SA Pathology, spanning the HXB2 region 2253 to 3350, incorporating the PR and RT regions. *Env*-gp41 PCR and gene sequencing was performed as part of the PhD study, the 530bp sequence incorporated the *env*-gp41 region, spanning HXB2 7816 to 8344. Primers used for *pol* and *env* sequences and other relevant information is summarized in Tables 3 and 4. All PCR experiments were performed on the MJ Research PTC-200 thermal cycler (Bio-Rad, California, USA) using SuperScript™ III One-Step RT-PCR System with Platinum® Taq High Fidelity (Invitrogen Life Technologies, California, USA).

3.3.7.1 Single round PCR, or step 1 of nested PCR

A single round RT-PCR was attempted initially with specific primers. If this failed a nested approach was attempted (for *pol* only). Positive and negative controls were included for each PCR experiment. Specifically designated laboratories or rooms were used for master mix preparation, PCR and all post PCR manipulations, RNA extraction and inoculation. Reagents and primers for PCR were thawed, except the enzymes. The 200µL micro-tubes were labelled for each reaction, and 1.5mL tubes labelled for the forward and reverse primer mixes. Pre-nested and nested PCR reactions for the *pol* and *env* regions contained 2x buffer (made up of 0.4mM dNTP, 2.4mM MgSo4), RT/Taq *pol*ymerase, 100mM of DTT, and 20µM of each primer, in a total volume of 25µL. The PCR mix was made up as shown in Table 6.

The master mix was pulse spun and 23µl added to the bottom of each 200µL tube, then 2µL of nuclease-free water was added to the negative control in the clean room. The tubes were transferred to a biohazard hood, and 2µL of known HIV positive sample added to the positive control, and 2µL of case sample RNA added to the remaining tubes.

Table 6. PCR master mix

	x1 conc.	x1 vol. μL)
Nuclease-free H_2O		7.10
2x Buffer		12.50
DTT	100mM	1.90
Forward Primer (F2)	20μm	0.50
Reverse Primer (R1M)	20μm	0.50
RT/Taq mix	2 units	0.50
RNA, HIV-1 positive (or negative control)		2.00
Total		25.00

The PCR microtubes were briefly microfuged and placed into the thermal cycler, which was set for a 25μL volume and heated lid. The following cycling conditions were used for both *pol* and *env* PCRs: one cycle at 50°C for 30 minutes to perform reverse transcription, one cycle at 94°C for 10 minutes to deactivate reverse transcription and activate the Taq, followed by amplification of 60 cycles of: denaturing at 94°C for 30 seconds, primer annealing at 50°C for 1 minute, and extension at 72°C for 2 minutes. There was one final step of elongation at 72°C for 5 minutes then the samples were cooled to 11°C until the PCR tubes were removed and stored at 4°C.

3.3.7.2 Step 2 nested PCR

Nested PCR was attempted for the *pol* regions that could not be amplified using normal PCR. The PCR mixes were made up as shown in Tables 7 and 8. Methods were the same as for first round PCR, except 1μL of the pre-nested product, including the negative and positive controls, was carried over to the nested reaction. Cycling conditions were the same as above minus the first step (50°C for 30 minutes to activate RT).

Table 7. PCR pre-nested master mix

Pre-nested master mix	x1 conc.	x1 vol. (μL)
Nuclease-free H_2O		6.10
2x Buffer		12.50
DTT	100mM	1.90
Forward Primer F3	20μm	0.50
Forward Primer F4	20μm	0.50
Reverse Primer R3	20μm	0.50
Reverse Primer R4	20μm	0.50
RT/Taq mix	2 units	0.50
RNA		2.00
Total		25.00

Table 8. PCR nested master mix

Nested master mix	x1 conc.	x1 vol. (μL)
Nuclease-free H_2O		9.10
2x Buffer		12.50
DTT	100mM	1.90
Forward Primer F2	20μm	0.50
Reverse Primer R1M	20μm	0.50
RT/Taq mix	2 units	0.50
RNA		1.00
Total		25.00

3.3.8 Gel electrophoresis and clean-up of PCR sequences

3.3.8.1 Visualization of PCR products

PCR products were visualized on 1.5% agarose gel, prepared by mixing 1.3g agarose in 90mL of 0.5X TBE to make a 2 x 20 well gel. Next, 2μL EZ vision loading buffer and 8μL PCR product for each sample, including controls, were mixed and loaded onto the gel and 8μL of 100bp DNA Marker, (1μL marker, 2μL EZ vision loading buffer and 5μL water) was loaded in the first well of each of the two lanes. The gel underwent electrophoresis at constant 80mA until dye had migrated at least two-thirds of the distance down the gel (~50 minutes for a fully loaded gel). The gel was then examined under UV light and photographed. If bands were found at the correct size (~1100bp for *pol*, ~530bp for *env*) and

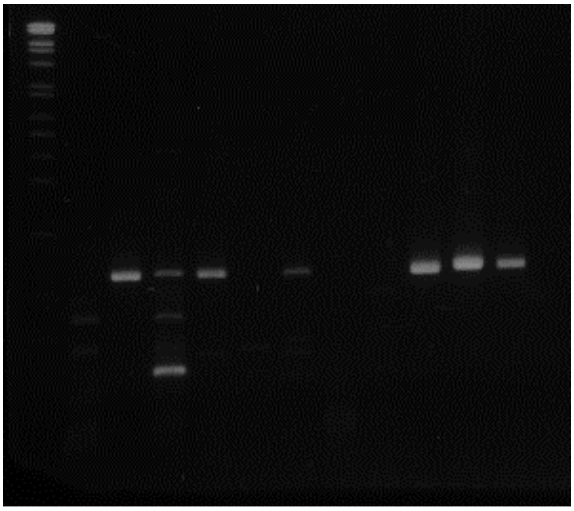
the negative control was clean, the PCR product was purified ready for sequencing.

3.3.8.2 Column purification of PCR products

The QIA-quick PCR purification kit (QIAGEN, Victoria, Australia) was used to purify PCR products following manufacturer's recommendation ≤www.qiagen.com/au/resources/download≥. For each reaction to be purified, a 1.5mL Eppendorf tube was labelled with case number and date and the spin column was labelled with the same number as the tube from the PCR template. Each PCR product had 5x volume of PB buffer added to it and mixed by pipetting, then the total volume was directly applied to the membrane of the spin column. The sample was centrifuged at 13,000 rpm for 1 minute, and the collection tube emptied. Next, 750μL of PE buffer was added to the spin column, centrifuged at 13,000 rpm for 1 minute, and collection tube emptied, followed by further centrifuging for 1 minute. The spin column was then placed in the labelled Eppendorf tube, and 30μL of elution buffer directly pipetted onto the membrane of the spin column. This was incubated for 1 minute at room temperature then centrifuged at 13,000 rpm for 1 minute. The column was discarded and the sample stored at 4°C until needed for sequencing.

3.3.9 Sequencing of partial env fragments

Amplified *pol*-PR/RT and *env*-gp41 fragments were directly sequenced in both directions using the respective amplification primers shown in Table 3. Every sequencing reaction contained approximately 1μL of purified PCR product, 1.6μM of sequencing primer, 1μL of Big Dye terminator enzyme mix, and 5x sequencing buffer diluted to 1x (made up of 400mM Tris HCl pH 9.0 and 10mM $MgCl_2$). Nuclease-free water was added to the reaction mix to give a final volume of 10μL. The size, purity and concentration of each PCR product was assessed from the gel and 1, 2 or 4μl was used in the sequencing reaction based on this assessment. Figure 9 illustrates this.



1 2 3 4 5 6 7 8 9 10 11 12

Figure 9. Agarose gel electrophoresis, *env*-gp41 sequences. From left to right, lane 1: 100bp ladder, lanes 2 6, 8-9: empty, lane 3: 1μL of purified product would be used, lane 4: product would not be used because of artefact present, lane 5: 2μL to be used, lane 7: 4μL, lanes 10–12: 1μL.

A 200μl tube was labelled for each sequencing reaction, and a master mix made up separately for forward and reverse primers as shown in Table 9.

Table 9. Sequencing master mix

Sequencing master mix	x1 conc.	x1 vol. (μL)	Multiplier
Big Dye Mix		1.00	(template 1, 2, 3 or 4μL. H_2O adjusted accordingly)
5x Buffer		1.50	
Primer F2 or R1M	1.6μm	1.00	
Nuclease-free H_2O		5.50	
Template		1.00	
Total		10.00	

The Corbett CAS-42 (QIAGEN, Victoria, Australia) was used to dispense the master mix and templates into a 96-well plate. The required volume of master mix from both the forward and reverse mixes was placed into separate wells. The required volume of each purified PCR product was placed into one forward and one reverse primer well. The plate was sealed and placed into the thermo-cycler, which was set for a 10μL volume and heated

lid. The following cycling conditions were used for both *pol* and *env* sequencing: 30 cycles of denaturation at 96°C for 10 seconds, primer annealing at 50°C for 5 seconds and elongation at 60°C for 4 minues. The samples were then cooled to 11°C until the tubes were removed and stored at 4°C. Samples were then taken to the onsite sequence analysis facility to be cleaned up and run on the 48-capillary 3730 DNA Analyser (Victoria, Australia).

3.3.10 Contiguous assembly and analysis of env sequences

Forward and reverse sequence chromatographic files were imported into Kodon v.2.4 and 3.61 (Applied Maths, Sint-Martens-Latem, Belgium), and aligned to create one contiguous 530bp sequence. Both primer sequences were excised, ambiguities corrected, and SNPs annotated.

3.3.11 Multiple alignments of partial pol and env reference and query sequences

Multiple aligments of *pol* and *env* query sequences were done by two reference sequence methods.

3.3.11.1 Method 1 - Query sequences submitted for LANL BLAST analysis

Reference sequences were obtained from the LANL BLAST database ≤http://www.hiv.LANL BLAST.gov/content/sequence/BASIC_BLAST/basic_blast.html≥ by testing each query sequence and downloading the 10 most closely related reference sequences for each query sequence. Reference sequences differed in size, ranging from partial gene sequences spanning either the *pol* or *env* region, to whole genome sequences.

3.3.11.2 Method 2 - Query sequences aligned with RIP reference sequence set.

A complete reference sequence set was obtained from the LANL BLAST database ≤http://www.hiv.LANL BLAST.gov/content/sequence/NEWALIGN/align.html≥. This provided a complete alignment of HIV-1 group M genome sequences including a consensus for each subtype and reference sequences for all CRFs (current as at 2012).

All reference sequences in the LANL BLAST database are specifically classified and annotated. For example for strain M.CM.95.YBF30.AJ006022, the M stands for Group M, CM the country of origin (Cameroon), 95 the year the sample was collected (1995), YBF30 is the name of the virus strain and AJ006022 is the GenBank Accession number.

Four separate FASTA sequence files were compiled, two *env* files and two *pol* files. One *env* and one *pol* file contained all the query sequences and LANL BLAST reference sequences. The other *env* and *pol* files contained all the query sequences and 2012 LANL BLAST RIP reference sequences. The FASTA files were submitted to CIPRES online ≤http://www.phylo.org≥ to undergo multiple alignment using ClustalW and default parameters ≤http://www.phylo.org/tools/clustalw.html≥.

The ClustalW output files were transferred to Bioedit v.7 for manual editing. The 9719bp HBX2 whole genome reference sequence (GenBank accession number K03455) was included in both *pol* and the *env* datasets, to determine length and location of the query *pol*-PR/RT and *env*-gp41 PCR sequences. The *pol*- PR/RT sequence stretched from 2253–3350 (1098bp) and *env*-gp41 sequence stretched from 7816–8344 (530bp) relative to HXB2 respectively.

Once manual editing was complete, the reference sequences (including HXB2) in each of the four datasets were trimmed to the same length as the query sequences.

3.3.12 Construction of NJ and ML trees using Kodon and MEGA

An initial neighbor joining phylogenetic tree was created in Kodon for each of the four files to ascertain which reference sequences were sufficiently unrelated to query sequences to be excluded. Reference sequences, including BLAST sequences of other cases in LANL BLAST, were excluded from the dataset if they were not clustered with any query sequences, or if they clustered but had a bootstrap value less than 50%. This excluded some pure subtype and CRF reference sequences and LANL BLAST sequences. The isolate SIVcpzCAM5

(Genbank accession number AJ271369) was included as an out-group in the *pol* and *env* alignments to ensure the tree structure was properly rooted. For both *pol* and *env* trees the out-group lay outside the rest of the tree and was also removed so as to not bias subsequent ML analysis.

The bioedit files were too large for Modeltest to ascertain the model of best fit. The files were imported into MEGA v.6, converted to MEGA format (.meg) then used to find the model of best fit, using the 'find best DNA/Protein models (ML)' tool. The model of best fit for both *pol* and *env* datasets was the general time reversible model. Each pair of nucleotide substitutions has a different rate, each of the nucleotides can occur at different frequencies, and the model assumes a symmetrical substitution matrix (time reversible), that is, for example, that C changes into T at the same rate that T changes into C.[162]

The .meg files were then used to create maximum likelihood trees, using the following parameters:

Statistical method: Maximum Likelihood
Test of phylogeny: Bootstrap method, 1000 replications.
Substitutions type: Nucleotide
Model/Method: General Time Reversible Model
Number of discrete *gamma* categories: 5
Gaps/Missing Data treatment: Use all sites
Select codon positions: 1st, 2nd, 3rd, & Noncoding sites
ML Heuristic method: Nearest Neighbour Interchange (NNI)
Initial tree for ML: NJ/BioNJ
Branch swap filter: Very strong

In total, four trees were created, one *env* and one *pol* tree with query sequences and the LANL BLAST sequences, and one *env* and one *pol* tree with query sequences and the LANL BLAST 2012 RIP reference sequences. Only the LANL BLAST 2012 RIP reference trees were used to determine subtype and conduct statistical analysis. The LANL BLAST reference trees were created as comparison trees, to see whether query sequences were clustered in the same way whichever reference sequences were added. Both sets of trees

showed a similar topology, and the LANL BLAST reference trees were excluded from further analysis.

3.3.13 Phylogenetic tree analysis

Phylogenetic analyses were performed using *pol* and *env* sequences of 221 cases newly diagnosed in South Australia between 2000 and 2012, to ascertain subtypes and to explore subtype and demographic characteristics of cases with sequences that cluster together with high reliability (≥70% bootstrap value from common ancestral node) and compare these with characteristics of cases whose sequences did not cluster together with high reliability. Demographic information for cases that formed part of transmission clusters was also examined, transmission clusters are defined below. The 2012 HIV-1 RIP custom background reference sequence dataset used in the analyses was obtained from the LANL BLAST database ≤http://www.hiv/LANL BLAST.gov/content/sequence/NEWALIGN/align.html≥. The *pol* sequence was 1096 base pairs long, beginning at HXB2 position 2253 and ending at 3350, which spans the entire protease gene and the first 800bp of the RT gene (RT gene is 1319bp long).[162] The *pol* gene is under intense selective pressure by antiretroviral therapy, which may affect phylogenetic reconstruction. However, we did not remove codons associated with drug resistance as per results by Hue et al.,[162] who found congruence between trees with and without target sites removed.

The *env* sequence was ~530 base pairs long, beginning at HXB2 position 7816 and ending at 8344, spanning part of the gp41 gene, including the entire transmembrane domain (gp41 is 1034bp long, beginning at position 7758 and ending at 8792). *Env*-gp41 is involved in the fusion of the virus particle with host cells.[200] The region of gp41 sequenced also spanned the T-20 drug HR-1 and HR-2 sites, which are under selective pressure by fusion inhibitors. However studies have found that the gp41 gene remains highly conserved even throughout long-term therapy so these sites were retained in the sequence.[201]

Subtype/CRF classification was based on proximity to the closest reference sequence/s within the tree. Query sequences were considered to be of the subtype or CRF of the reference sequence clustered with or near it.

3.3.13.1 Cluster analysis

Phylogenetic tree bootstrap values reflect the frequency with which a sequence sits in a certain position on the tree after a number of bootstrap replications, for this study, 1000 bootstrap replications. A 70% cutoff value at the common ancestral node is considered highly reliable for determining sequences that cluster together and have evolved from the same clade. They may be directly or distantly related, reflected by the branch lengths. For the purpose of this study, sequences that clustered together with an ancestral node bootstrap value of ≥70% were called 'high reliability clusters' and sub-clusters were defined as sequences with an ancestral node bootstrap value of ≥70% located within the larger ≥70% clusters.

Transmission cluster membership was defined as two or more sequences with bootstrap values of 98% or higher, and each sequence had a genetic distance of ≤1.5% from at least one other sequence in the cluster.[163] These sequences may be directly related, related through shared transmission (two people infected by the same person), or related by intermediary transmission (person A infects person B, person B infects person C).

3.3.14 HIV-1 online subtyping tools

3.3.14.1 Overview

As mentioned previously, phylogeny is the gold standard for determining sequence subtype or CRF[182] however it is quite complex, and not widely implemented globally. Rapid online subtyping tools are increasingly being developed and used as an alternative. These are often free to use, and provide very fast results.[182] Automated online tools

approximate a subtype assignment, as to define a subtype is to know the clade, which is identified through phylogeny. These approximation tools have been developed due to the growing need for fast results, and the issues that arise with increasing viral diversity which makes it difficult to identify recombinants using a phylogenetic tree.[183] However, these automated methods are not without issue, often these tools cannot identify unique or complex mosaics, and may over or under identify recombinants. Some tools also have a high level of discordance with results from phylogenetic analyses, which is cause for concern.

A number of mathematical models underpin the capability of rapid online tools to classify viral strains into subtypes, CRFs or more complex recombinant strains,[157] though there is no universally agreed upon approach. The models are broadly categorised into those that do and do not subtype using phylogeny, whether or not the model performs or requires a sequence alignment, and the automation process (full, partial, none).[183] There are three main types of tools: 1) similarity-based tools which are alignment free and work on sliding window analysis, such as Stanford CPR and LANL BLAST,[186,187] 2) statistical-based tools that use prediction by partial matching compression algorithms such as COMET[188] or jpHMM;[189] both tools measure sequence similarity by data compression schemes that do not require sequence alignment, and 3) phylogenetic-based tools such as REGA[184] and SCUEAL,[183] which work on phylogeny and alignment based algorithms[185] that allow detection and identification of recombinant strains with added bootstrap support (REGA) or the phylogenetic likelihood of a mosaic (SCUEAL).[188]

Currently, the REGA, SCUEAL and Stanford CPR online subtyping tools do not have the capability to analyse *env* sequences, and REGA cannot detect recombination events for *env* sequences ≤800bp.

To identify recombination, an initial phylogenetic bootstrap analysis is run to assess the query sequence against a given reference set in order to infer breakpoint locations which are then confirmed by detailed phy of pure subtype and known CRF reference sequences.[183] The best match returned is designated as the putative subtype or CRF.[188] These tools often enable the user to adjust parameters such as the sliding window size and step size, reference sequences, and alignment parameters. While this can improve detection of recombination events, it can also over-identify events or lead to ambiguous results.

One of the issues with automated tools requiring alignment, is that the one alignment chosen may not be statistically distinct from other alternative alignments. Alignment is also affected by the choice of algorithm and the parameters set (default or user defined). Alignment confidence is measurable but this can be time consuming, especially when the query sequence is a complex mosaic. A conserved gene region such as *pol* may not be as impacted by these factors as a more variable region such as *env*.[190]

Other online subtyping methods do not require an initial alignment, they work on the premise that sequence similarity can be measured by data compression, which identifies the amount of information shared between two sequences.[188] Examples of alignment-free models include the sliding window analysis used by the Basic Local Alignment Search Tool (BLAST)[186], and Markov Models that identify the probability of a given nucleotide at the next site of a sequence by genomic context; looking at the last *k* nucleotides in the sequence. The two sequences are then compared by measuring the context-specific base frequencies in the second sequence to the Markov model built from the first sequence.[188,189] Two alignment free tools which use Markov models are jpHMM[189] and COMET, the latter of which is based on the Prediction by Partial Matching compression algorithm (PPM).[191]

Six HIV-1 online genotyping tools were used to identify and evaluate the proportion of

pure subtypes, CRFs and possible unique ISRs, (jpHMM, REGA, Stanford CPR, SCUEAL, COMET and LANL BLAST). The use of more than one tool enabled analysis of the degree of similarity/difference between the tools in assigning subtypes, and limited the number of incorrect subtype identifications. The subtype decided upon using the online tool parameters (see below) will be referred to hereafter as the '**inferred subtype**'.

Five of the six online tools provided percentage support values with the subtype: bootstrap model averaged support (SCUEAL), % similarity to reference sequence (LANL BLAST and Stanford CPR tool), and bootstrap support (REGA and jpHMM). Subtypes with a support value of 70% or over were defined as reliable.

The online tool parameters for assigning subtypes were as follows:

- All six online tools were used for *pol* sequences. If five of six tools assigned the same subtype then the *pol* sequence was assigned as that subtype.
- Four online tools were used for *env* sequences (jpHMM, REGA, LANL BLAST and COMET). If three of four tools assigned the same subtype then the *env* sequence was assigned as that subtype. SCUEAL and Stanford CPR do not have the capability to subtype the *env* gene.

It should be noted that REGA has the capability to perform recombination analysis on sequences ≥800bp, but currently only uses up to reference CRF47_BF (there are currently 60+ CRFs listed on LANL BLAST). REGA does not have the capability to perform recombination analyses on sequences ≤800bp, so the *env* sequences were only analyzed for pure subtypes by this tool. JpHMM is capable of pure subtype and recombinant analysis for the *env* region, but may incorrectly classify short fragments located near the 3' end. The results for jpHMM assigned *env* subtypes were therefore interpreted with caution. Subtypes assigned by online tools were then compared to phy to assess the degree of similarity.

3.3.14.2 *Stanford HIVdb – calibrated population resistance (CPR) tool*

The CPR tool is a free online tool linked to the HIV-1 Drug Resistance Database (Stanford University, Palo Alto, CA) available at ≤http://hivdb.Stanford.edu≥.[198] It estimates genotypic drug resistance and can subtype sequences using algorithms to match reference sequences to query sequences. It currently only accepts *pol* sequences, and assigns subtypes by constructing pairwise alignments using the PR and RT regions of the *pol* query sequence and each of the reference sequences (subtypes A-D, F-H, J and K, CRFs 01_AE and 02_AG). Each gene region is assigned to one of these subtypes/CRFs according to the highest shared percentage identity. Separate assignments are reported for PR and RT.[198] It does not accurately detect other recombinants besides 01_AE and 02_AG.[199]

3.3.14.3 *jpHMM*

The jpHMM method from the University of Göttingen ≤http://jphmm.gobics.de/≥ is a generalisation of the Markov model and the jumping alignment (JALI) model, the latter works by aligning each position in a query sequence to one sequence of a multiple alignment set using a scoring matrix.[189] The reference sequence can change within an alignment (a jump), and each jump incurs a penalty *jumpcost*, similar to a gap penalty in standard sequence aligment.[189,202] The jpHMM tool therefore works by aligning each position to a column of the multiple sequence alignment, or to a certain reference sequence within the set. This model incorporates both horizontal and vertical information in the sequence alignment.[189]

It uses a probabilistic generalization of the jumping alignment approach to subtype sequences and identify recombination breakpoints. The query sequence is compared and aligned to individual sequences from a large reference alignment. A sliding window moves over the alignment, and compares regions of X with different reference sequences, finding breakpoints between different subtypes.[189] It creates a report of assigned subtype and bootstrap support value. The limitations of jpHMM are that it may incorrectly classify short

sequence fragments located near the 3' end of the genome, and may not be sensitive enough to detect subtypes H, J, and K.[189]

3.3.14.4 *REGA v.3*

The Rega tool aligns query sequences with pure subtype and CRF reference sequences up to 47_BF using clustalW, and constructs a phylogeny using Phylogenetic Analysis Using Parsimony (PAUP) and the Neighbor Joining (NJ) method (100 bootstrap iterations, 400bp sliding window and 20bp step size), then conducts likelihood mapping with TreePuzzle. It creates a report of assigned subtype and bootstrap support value, however the efficacy of Rega is limited by a threshold which prevents the identification of a subtype or CRF when there is not strong statistical support, sequences that do not meet the criteria for confident assignment (≥70% similarity) are unclassified.[184,194] Another limitation of REGA v.3 is that it only includes reference CRF sequences up to 47_BF.[203]

3.3.14.5 COMET v.1

COMET v.1 is an alignment-free tool that detects subtype and recombination events using the Prediction by Partial Matching compression algorithm. It was used to subtype both the *pol* and *env* regions, and uses the complete 2010 subtype/CRF reference set from LANL BLAST, which includes CRFs up to 44_BF.[188]

Markov models are built for a set of reference sequences, then when a query sequence is entered, COMET uses the models to create a matrix in which the rows correspond to the reference subtype, and the columns reflected the estimated log likelihood of observing a given nucleotide in the query sequence for each of the reference subtypes. A decision tree is then applied to the matrix which determines the final subtype assignment; either a reference type (pure subtype or CRF), a pure reference type (which may come from a non-recombinant region of a CRF), or unassigned (novel recombinant form or ambiguous query sequence).[188]

3.3.14.6 SCUEAL

The SCUEAL tool is another phylogeny-based model, run within Datamonkey ≤http://www.datamonkey.org≥. The method uses a maximum likelihood multi-model inference approach to assess *pol* sequences. It is a completely automated tool with a complex algorithm that allows quick analysis of large datasets, including those that potentially contain mosaic structures. It does this by screening single query sequences against a fixed reference alignment that is updated regularly. It assigns a predicted subtype, CRF or more complex variant, but also identifies detailed information about recombination, including specific breakpoint locations and the percent likelihood of the presence of inter- and intra-subtype recombination, including assignment of parental/sister lineages so the user can objectively evaluate the robustness of the estimates.[183] This tool was used to classify unique recombination events in the present study.

3.3.14.7 LANL BLAST

The LANL BLAST (hereafter referred to as LANL BLAST) program is used widely to search the LANL BLAST HIV database. This is comprised principally of sequences submitted to GenBank ≤https://www.ncbi.nlm.nih.gov/genbank/≥ and is used to find protein and DNA sequences most similar to the submitted query sequence/s. The BLAST program uses a sophisticated algorithm to maximize sensitivity to weak similarities while minimizing execution time. A query sequence is submitted as a FASTA format, search parameters are entered (output style: pairwise; number of BLAST matches to display: 25; Run BLAST against: all subtyped sequences; show location of match in genome: tick yes) then the BLAST program searches through all database sequences using the algorithm and finds the matches most similar to the query. Bulk query sequences can be submitted, and no alignment is needed.[186] BLAST returns the best hits from the database (number decided by the user), including accession numbers for each sequence, name of the sequence, subtype, country the

sequence came from, the sampling year, the GenBank sequence description, and a score (bits). The score is calculated by the BLAST algorithm and takes into consideration the length of the alignment and percentage of matching bases, E value (likelihood of this match occurring by chance), identities (the number and percent identity between the submitted query and the matching query across the longest continual alignment), and location of match in genome.[186]

The Blast searching tool is very good at matching a query sequence with the most similar sequences in large reference sequence repositories, and uses a sliding window and step increment along the query sequence. However depending on the size of the sliding window and step size, this can lead to an over or under representation of recombination, which is further complicated when multiple highly similar references sequences are included, especially some CRFs which are very similar and difficult to discriminate in the particular region of the sequence being queried.[182]

3.3.15 Detection of recombinant viruses

Pol or *env* sequences were considered to be recombinants if at least one online tool detected a recombinant virus and there was discordance between the other online tools for that *pol* or *env* sequence. Sequences were considered to be unique recombinants if SCUEAL detected an ISR with breakpoints that did not correspond with any known CRFs, with evidence of recombination support values of more than 50%. Cases were considered to have intergene recombination if there was discordance between the **inferred** *pol* and *env* subtypes (decided by using the online tool parameters). Cases were considered to have intrasubtype recombination if SCUEAL detected recombination of multiple strains of the same pure subtype with an intrasubtype recombination support vale of more than 50%.

3.4 Statistical analysis

For the molecular epidemiological subtype study, HIV-1 data were analyzed using

subtype as the dependent variable, and year of diagnosis, country of origin, location of acquisition, reported risk exposure, and age as explanatory variables. Categorical variables were analyzed using chi-squared or Fisher exact test to identify subtype-specific characteristics. Multivariate analysis was performed using logistic regression.

For the drug resistance study, data were analyzed using resistance mutations as the dependent variable, and the same explanatory variables but with the addition of subtype. Notification data were aggregated by year of diagnosis (2000–2004, 2005–2009, and 2010–2013) into three time periods of relatively equal numbers and sufficient size to conduct statistical tests. Categorical variables were analyzed using the Fisher exact test.

For the phylogenetic and online subtyping tool study, data were analyzed by comparing clustered and non-clustered sequences. Explanatory variables compared were subtype, year of diagnosis, country of origin, location of acquisition, reported risk exposure, sex, and age. Sequences were aggregated into two time periods by year of diagnosis (2000–2006 and 2007–2013) to ensure there were adequate numbers to conduct statistical tests. Fisher exact test was used to compare clustered vs. non-clustered cases.

Significance levels were set at $p \leq 0.05$. All data were analyzed using the software package Stata 10.1 (StataCorp LP, College Station, TX) or VassarStats ≤http://vassarstats.net/≥.

CHAPTER 4: MOLECULAR EPIDEMIOLOGICAL ANALYSIS

4.1 Overview

The monitoring of HIV subtype distribution is important for understanding transmission dynamics. Historically, HIV subtypes and CRFs have been broadly linked with geographic location and risk group,[70] with subtype B dominating the Australian epidemic. However, subtype distribution in the global HIV epidemic has diversified extensively through mutation and recombination, partly driven by a combination of population mobility, diversity of sexual contacts through travel and migration, and the impact of antiretroviral therapies.[6,52]

Surveillance systems have been in place since the beginning of the global epidemic and in recent years these have incorporated molecular epidemiology as a tool for surveillance of HIV-1 genetic diversity and to monitor transmission and geographic pathways of genetic variants.[6,73,108,109]

In 2000, South Australia became the first state to integrate drug resistance testing into routine HIV reporting and surveillance. The resulting molecular epidemiological analysis shows the changing HIV-1 subtype distribution in South Australia and risk factors associated with different subtypes.

In the following four results chapters, the word 'cases' is used to refer to both the affected person and the infection. For example, 'female cases' refers to persons who are female, whereas 'the pol-ISR AE/B case' refers to the virus.

4.2 South Australian HIV population: subtype and clinical notification data.

Between 2000 and 2013 there were 656 newly diagnosed HIV-1 cases. For 569 of these (87%; 482 males and 87 females) subtypes were determined by analyzing two separate sequences from the *pol* region, the PR and RT genes.

The PR and RT genes were phylogenetically concordant for 96% (549) of cases. There were 75% (413) subtyped as pure subtype B and 25% (136) non-B cases; the latter comprised

43% (58) pure subtypes (5 subtype A, 49 subtype C, 2 subtype D and 2 subtype G) and the remaining 57% (78) were CRFs, 59 01_AE and 19 02_AG.

In the remaining 3.5% (20) of cases there was phylogenetic divergence between the PR and RT sequences. The Stanford CPR subtyping tool ≤http://sierra2.Stanford.edu/sierra/servlet/JSierra≥ detected 12 intersubtype/CRF patterns, Table 10. These cases will be referred to hereafter as *pol* intersubtype recombinants (*pol*-ISRs).

The annual number of new diagnoses (including subtyped and non-subtyped cases) has more than doubled since 2000, from 23 new cases in 2000 to 56 cases in 2013 (mean = 47/year). Demographic and other characteristics of those for whom subtypes were obtained (n = 569) were very similar to those for the total diagnosed population (n = 656, Table 11). Only subtyped cases will be discussed in the remainder of this chapter.

There has been a significant change in subtype distribution in newly diagnosed individuals. The proportion of non-B cases and *pol*-ISR cases combined increased from 19% (33/171; 2000–2004) to 24% (52/216; 2005–2009) and 39% (71/182; 2010–2013; $p \leq 0.0001$). Figure 10 shows the proportional increases split by non-B subtypes and *pol*-ISRs.

In 2010, the proportion of non-B/*pol*-ISR cases rose to 47% but it has decreased each year thereafter, to 39% in 2011 and 34% in 2013, Figure 11. Subtype distribution can be seen in Figure 12. The proportion of locally transmitted cases that were non-B or *pol*-ISR increased significantly from 2% (3/128) in the first time period, 6% (9/155) in the second and 12% (12/104) in the third ($p = 0.01$), Figure 11. Just over half (13/24) of these non-B locally acquired cases were female, all of whom reported heterosexual transmission (one with IDU risk and one overseas-born partner risk). Of the 11 males, seven reported MSM transmission, three reported heterosexual transmission (one IDU risk) and one occurred through MTCT.

Table 10. Proportion of newly diagnosed cases by pure subtypes, CRFs and mixed pattern ISRs in South Australia, 2000–2013

	PR Sequence	RT Sequence	N	%
Pure subtypes	A	A	5	0.88
	B	B	413	72.58
	C	C	49	8.61
	D	D	2	0.35
	G	G	2	0.35
CRF's	01_AE	01_AE	59	10.37
	02_AG	02_AG	19	3.34
Mixed pattern ISRs	A	01_AE	5	0.88
	B	A	1	0.18
	B	01_AE	1	0.18
	B	02_AG	3	0.53
	B	C	2	0.35
	B	D	1	0.18
	B	G	1	0.18
	C	B	1	0.18
	D	A	1	0.18
	01_AE	A	1	0.18
	02_AG	B	2	0.35
	02_AG	01_AE	1	0.18
			569	**100.00**

Table 11. Characteristics of newly diagnosed HIV-infected cases in South Australia 2000–2013

Characteristics	Total n=656 (%)	With subtype n=569 (%)	Female subtyped n=87 (%)	Male subtyped n=482 (%)	M/F Ratio	Proportion B n=413 (%)
Year Diagnosed						
2000–2004	198 (30)	171 (30)	22	149	6.8	138 (81)
2005–2009	266 (41)	216 (38)	27	189	7.0	164 (76)
2010–2013	192 (29)	182 (32)	38	144	3.8	111 (61)
Gender						
Male	547 (83)	482 (85)	-	-		384 (79)
Female	109 (17)	87 (15)	-	-		29 (33)
Age at diagnosis (yrs)						
24 and under	71 (11)	62 (11)	20 (23)	42 (9)	2.1	36 (58)
25-50	470 (72)	408 (72)	58 (67)	350 (73)	6.0	307 (75)
51 and over	112 (17)	97 (17)	8 (9)	89 (18)	11.1	69 (71)
N/A	3 (.45)	2 (.35)	1 (1)	1 (.2)	1.0	1 (50)
Region of birth						
Australia	371 (57)	335 (59)	32	303	9.5	298 (89)
Sub-Saharan Africa	74 (11)	64 (11)	27	37	1.4	4 (6)
Asia	56 (8)	42 (7)	17	25	1.5	12 (29)
Central Asia/Middle East	11 (2)	11 (2)	3	8	2.7	1 (9)
America	10 (1.5)	9 (2)	2	7	3.5	8 (89)
Europe	50 (8)	36 (6)	0	36	n/a	28 (78)
Unknown	84 (13)	72 (13)	6	66	11.0	62 (86)
Risk exposure						
Heterosexual	169 (26)	138 (24)	56	82	1.5	50 (36)
Heterosexual/IDU	42 (6)	39 (7)	9	30	3.3	33 (85)
Heterosexual (o/seas born partner)	21 (3)	21 (4)	10	11	1.1	1 (5)
MSM	355 (54)	311 (55)	-	313	n/a	294 (95)
MSM/IDU	36 (6)	29 (5)	-	29	n/a	28 (97)
Blood/Medical Procedure	15 (2)	15 (2)	5	10	2.0	3 (20)
MTCT	8 (1.5)	6 (1)	4	2	0.5	1 (17)
Unknown	10 (1.5)	10 (2)	3	7	2.3	4 (40)
Location acquired						
Australia/Australia	425 (65)	387 (68)	39	348	8.9	364 (94)
Overseas	219 (33)	174 (31)	48	126	2.6	44 (25)
Not reported	12 (2)	8 (1)	-	8	n/a	6 (75)

Key: Data represents number (%) of cases within each category. Percentages in the last column are the proportion of B cases compared with non-B cases (total subtyped/total B cases for each individual variable within a category). Subtyped sample was representative of the total population diagnosed. MSM (men who have sex with men), IDU (intravenous drug use), MTCT (mother to child transmission), Unknown (Insufficient information in database).

These 24 non-B cases acquired in Australia comprised 19 non-B infections, predominantly 01_AE (n=13), followed by 02_AG (n=3), subtype C (n=2) and subtype D (n=1). Seven people carrying 01_AE were born in Asia, four born in Australia, and one each in Europe and Africa. Three people carried 02_AG viruses, two of whom were born in Africa and one in Australia. Both subtype C viruses were carried by Australian-born people and one Australian-born person carried subtype D. The remaining five people, all male, carried *pol*-ISR viruses. Three were born in Australia and two in Africa, only one person was diagnosed in 2008, the other four were diagnosed between 2010 and 2013.

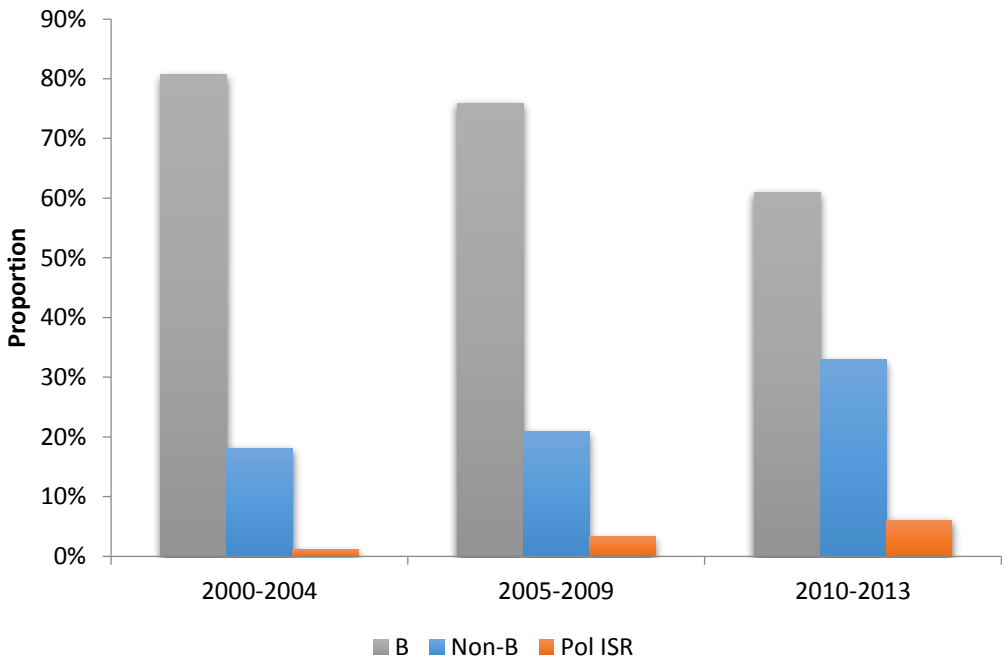


Figure 10. Proportion of newly diagnosed non-B and *pol*-ISR cases in South Australia, 2000–2013.

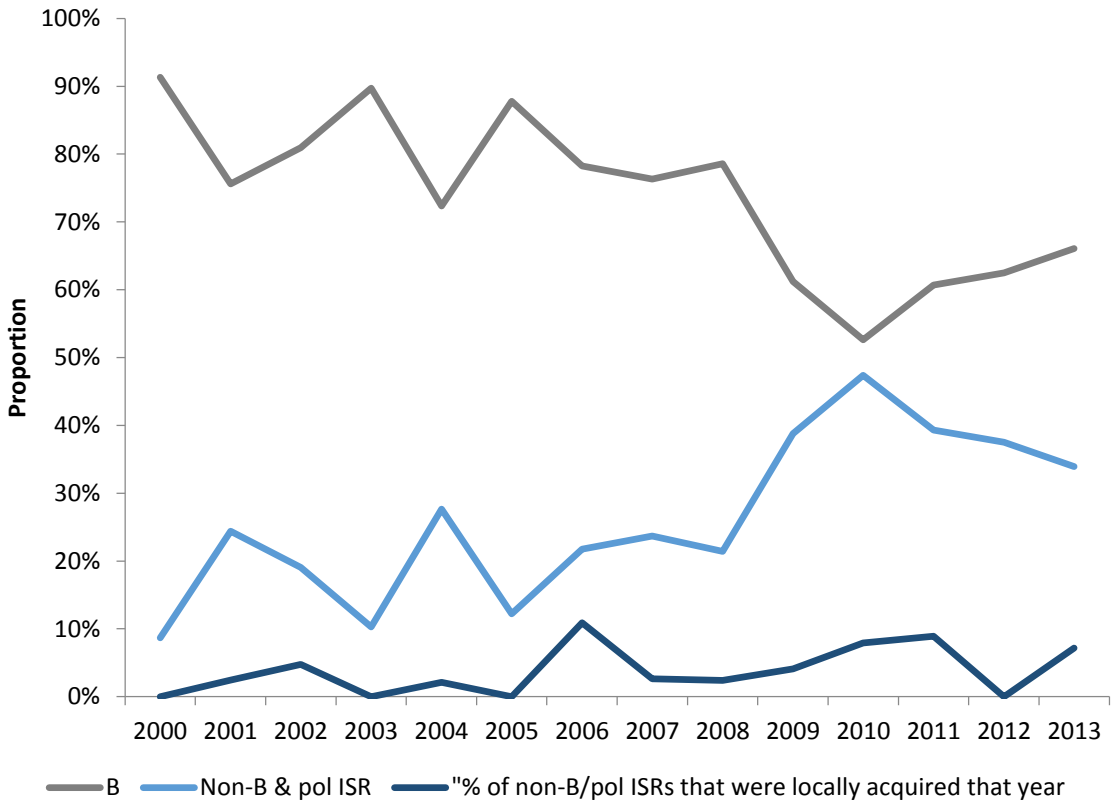


Figure 11. Proportion of newly diagnosed B versus non-B cases in South Australia 2000–2013, including local transmission of non-B infection.

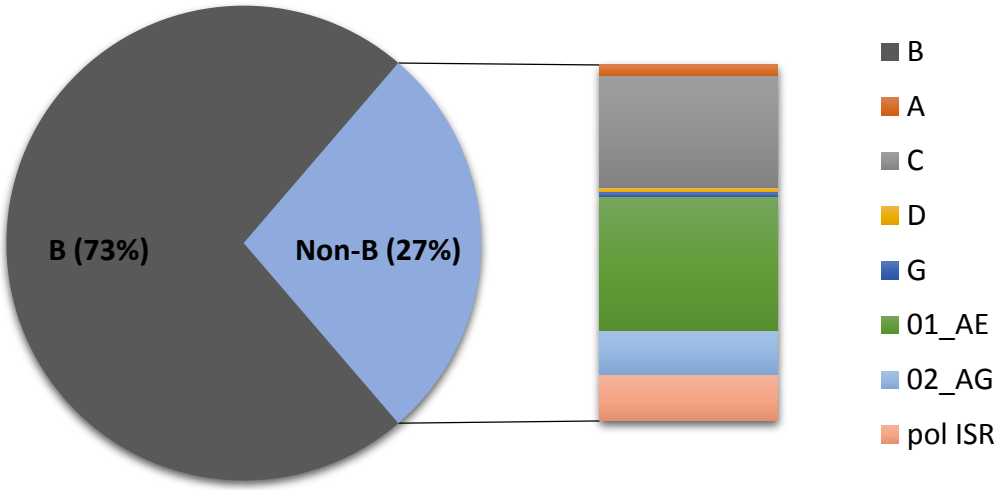


Figure 12. Subtype distribution, 2000–2013.

4.3 Male to female ratio

There has been a change in the ratio of male to female HIV diagnoses. Between 2000 and 2004 there were 6.8 males newly diagnosed with HIV for every female, but between 2010 and 2013, there were only 3.8 males to every female newly diagnosed, Table 11. This reflects the increase in the proportion of newly diagnosed cases being females, from 13% (22/171; 2000–2004) to 21% (38/182; 2010–2013), though the finding was not significant using the Yates chi square corrected for continuity ($p = 0.06$). Female cases outnumbered or were almost equal to male cases in some population groups: children, adolescents and young adults, people born in Sub-Saharan Africa and Asia, and those with heterosexual and MTCT transmission risk, Table 11.

There was a correlation between the proportional increase in female infections and the increase in the proportion of non-B and *pol*-ISR infections, with the proportion of non-B cases who were female being significantly higher (37%, 58/156) than B cases (7%, 29/413; $p \leq 0.0001$). Females with a non-B infection predominantly acquired their infection overseas (78%, 45/58), whereas 26 of the 29 subtype B infected females acquired the infection in Australia (90%).

Transmission in the female cohort was predominantly heterosexual (86%, 75/87), including nine females who listed heterosexual contact with IDU risk, and 10 females who reported infection by an overseas-born partner. Five females had a medical procedure listed as transmission risk, four were reported as MTCT and three were unknown. In contrast, the male population mostly consisted of MSM transmission (71%, 342/482), including 30 males who listed MSM with IDU risk, while 26% of males reported heterosexual transmission, including 30 who reported IDU risk, and 11 who reported being infected by an overseas-born partner. Ten males reported direct blood exposure, two male children were infected through MTCT, and seven had unlisted transmission risk.

4.4 Infections acquired in Australia or overseas

Almost one-third of cases were reported as being acquired overseas (31%, 174/569), with the proportion increasing from 24% (41/171) in 2000–2004, 28% (61/216) in 2005–2009 and 40% (72/182) in 2010–2013 ($p \leq 0.005$), Figure 13a. Subtype B was predominant in Australian-acquired cases over the entire time period, (94%, 363/387), while the majority (75%) of those acquired overseas were non-B or *pol*-ISR (130/174). The subtype profile between the two groups was significantly different ($p \leq 0.0001$).

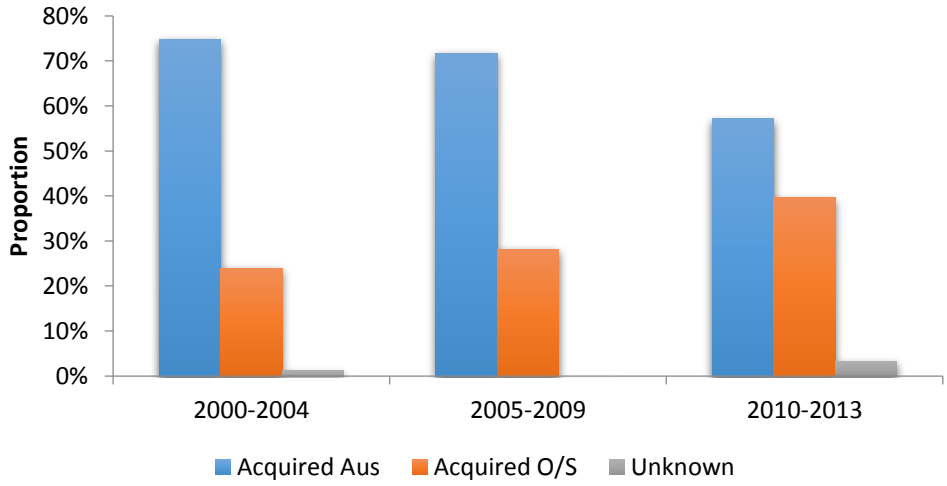


Figure 13a. Proportion of cases by place of acquisition, 2000–2013.

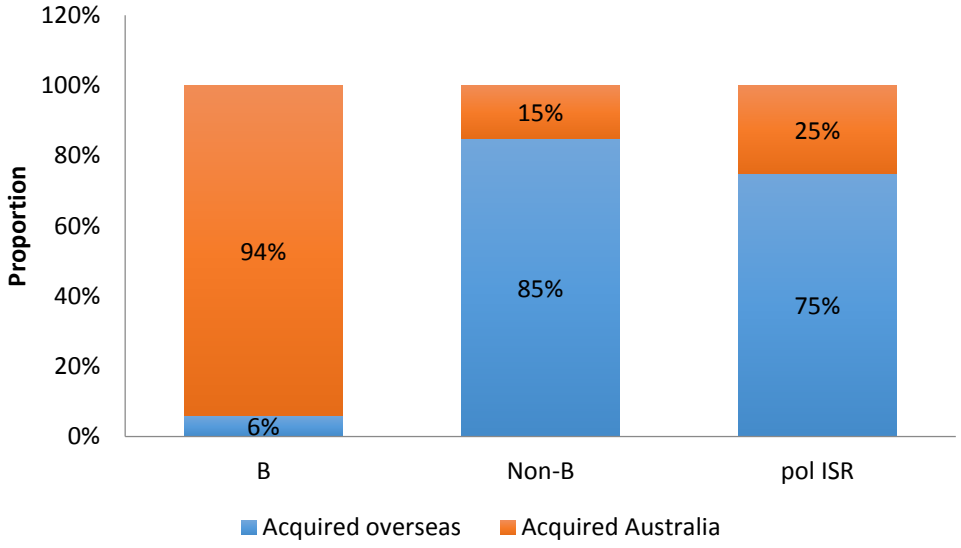


Figure 13b. Proportion of B, non-B and *pol*-ISR cases acquired overseas or in Australia.

There were six main subset groups, shown in Figure 13b. They are listed below in order of prevalence from the total cohort of people who had a location listed for infection acquisition (n=561):

1. People with a B virus who acquired their infection in Australia (n=363, 65%)
2. People with a non-B virus who acquired their infection overseas (n=115, 20%)
3. People with a B virus who acquired their infection overseas (n=44, 8%)
4. People with a non-B virus who acquired their infection in Australia (n=19, 3.4%)
5. People with a *pol*-ISR virus who acquired their infection overseas (n=15, 2.8%)
6. People with a *pol*-ISR virus who acquired their infection in Australia (n=5, 0.8%)

4.5 Subtype distribution by reported transmission risk exposure

There were 60% of subtyped cases transmitted through MSM exposure (340/569) with 8.5% also reporting IDU risk. However, MSM cases declined from 67% in 2000–2004 (114/171) to 61% in 2005–2009 (132/216) and 52% in 2010–2013 (94/182) (p = 0.01). Correspondingly, the proportion of reported heterosexual transmission cases increased from 31% (53/171) to 35% (75/216) and 38% (70/182) respectively. MTCT and direct blood exposure cases also increased over the same time periods, from 0.60% (1/171) to 3% (7/216) and 7% (13/182) (p = 0.004).

Subtype distribution was uneven among different risk exposure groups. Subtype B infections predominated in the MSM group (94%, 321/340), while there were less numbers of B compared to non-B and *pol*-ISR cases in the heterosexual population (B 42%, 84/198; non-B 52%, 103; *pol*-ISR 6%, 11) and mostly non-B/*pol*-ISR cases in the MTCT/direct blood exposure populations (B 19%, 4/21; non-B 67% 14; *pol*-ISR 14% 3).

Each risk exposure group is discussed further below and shown in Figure 14a.

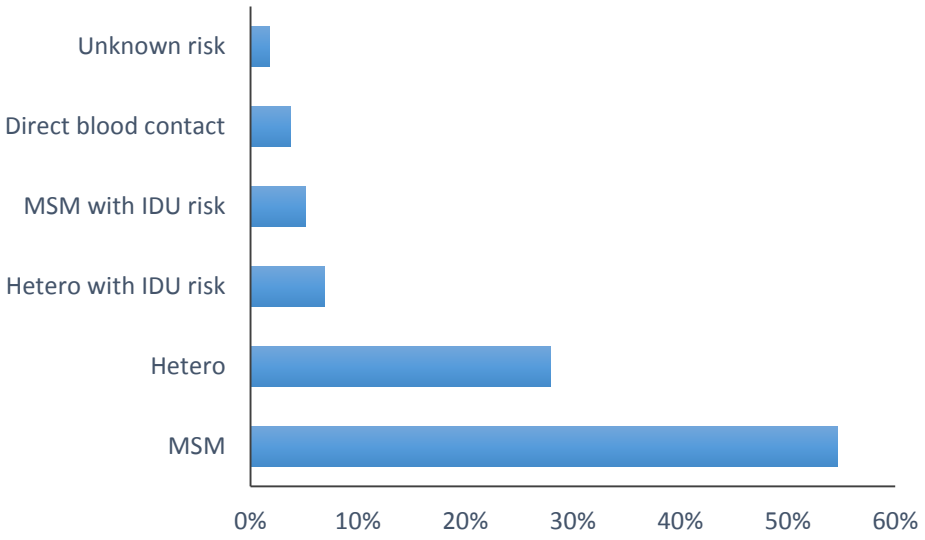


Figure 14a. Proportion of infections acquired by reported transmission route.

4.5.1 MSM population

Sixty percent (340/569) of all subtyped cases were MSM exposures, including 29 cases with IDU risk. The majority of MSM cases were acquired in Australia (86%, 294), and were Australian-born (69%, 236), or had region of birth unlisted (16%, 54). Almost all MSM cases with IDU risk were acquired in Australia (93%, 27/29), and 23 of these MSM/IDU cases were Australian-born. All MSM with IDU risk cases carried a subtype B virus except one, an African-born male with a subtype C virus.

Only 15 people in the MSM group carried a non-B virus. Eleven MSM carried 01_AE virus, three carried subtype C and one carried subtype G. Four MSM also had a *pol*-ISR virus – all subtype B recombinants (2 AG/B, 1 B/D, 1 B/G). Eleven of the total 19 infections were carried by Australian-born males; four acquired in Australia and seven overseas. Seven were carried by overseas-born males, three acquired in Australia, three overseas, and one

unknown. There was also a male of unknown origin who acquired the infection overseas.

4.5.2 Heterosexual population

Just over one-third of the cases were transmitted through heterosexual contact (35%, 198), including 20% (39) cases with IDU risk, and 21 cases in which transmission by an overseas-born partner was reported. Just over half of the heterosexual population carried a non-B virus (52%, 103), 6% carried *pol*-ISRs and 42% carried subtype B viruses.

4.5.2.1 Heterosexual sex with IDU risk

IDU risk was reported for 39 cases, with the majority acquired in Australia (82%, 34). Of the five that were reported as being acquired overseas, three were Asian-born people, one was Australian-born and one had region of birth unlisted. The predominant subtype was B (85%, 33/39). Four of the six non-B/*pol*-ISR cases were among the five acquired overseas.

4.5.2.2 Heterosexual transmission from an overseas-born partner

A subset of 21 heterosexual cases was reported as being transmitted by partners who were born overseas in areas of high HIV prevalence. Only one of the 21 cases was reported as being acquired in Australia, an African-born male infected with a 01_AE virus. The remaining 20 people reported acquiring the infection overseas, and 17 of these were born overseas themselves. All 17 of these overseas-born people had a non-B or *pol*-ISR virus, predominantly 01_AE or 02_AG (n=7) and subtype C (n=6). Three Australian-born males reported being infected overseas; one had subtype B, the other two had 01_AE.

4.5.2.3 Australian- and overseas-born heterosexual populations

The heterosexual population was split into three sub-groups: Australian-born, overseas-born, and those without a listed region of birth (8% of total heterosexual infections). Almost half the heterosexual cases were people born in Australia (47%, 94/198), of whom 66% (62) were male. Within this Australian-born heterosexual cohort, (73%, 68) acquired the infection

within Australia, of whom 92% carried a subtype B virus. Of those who acquired the infection overseas (27%, 25), 28% carried a subtype B virus. Two people did not report where they had acquired the infection.

Forty-five percent of the heterosexual population was born overseas (90/198), and 84% (75) of these acquired the infection overseas. Of these, 84% (63) carried a non-B virus and 10% (9) carried a *pol*-ISR virus. Sixteen percent (14) acquired the infection within Australia, of whom 64% (9) carried a non-B virus and 7% (1) carried a *pol*-ISR virus. One person did not report where the infection has been acquired, Figure 14b).

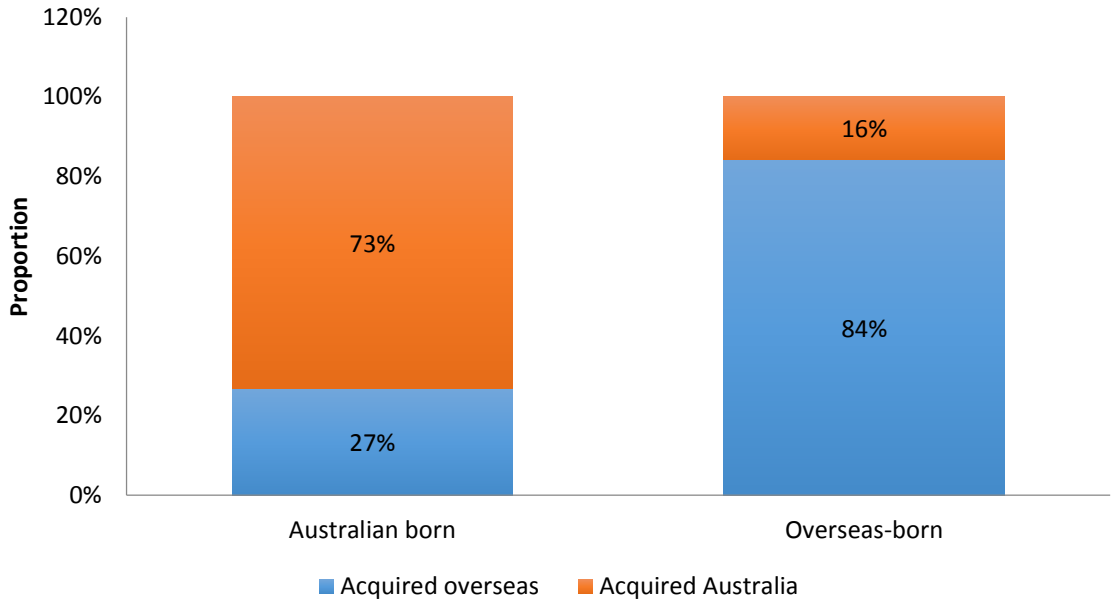


Figure 14b. Proportion of heterosexually transmitted infections acquired overseas or in Australia, by person's region of birth.

4.5.3 Population infected by direct blood contact including MTCT

Only 4% (21) of infections were transmitted through direct blood contact, including MTCT. Three of these were acquired in Australia. One was an A/01_AE MTCT case diagnosed in 2008 at the age of 2, and the other two were adult males with subtype B who reported direct blood contact, both diagnosed in 2010. The remaining 86% (18/21) were

cases acquired overseas and all were overseas-born except two with unknown region of birth. Sixteen of the 18 overseas-born cases carried a non-B subtype or *pol*-ISR.

Of note were three cases reported as medical procedure exposures, diagnosed in 2009, 2010 and 2012 respectively. They were all subtyped as 02_AG and occurred in children aged 10 years at time of diagnosis who were born in a part of Central Asia where there was a known hospital-based transmission epidemic at the time.[62]

4.6 Subtype distribution by age at diagnosis

The average age at diagnosis was 39 years old, being 40 years for males and 32 years for females. Average age at diagnosis varied by subtype, as seen in Figure 15a. The 01_AE and subtype B groups had the highest mean age at diagnosis, while the *pol*-ISR and 02_AG groups had the lowest. The age range for each category is as follows: pure subtype B (18–80 years), pure subtypes non-B (4–70 years), *pol*-ISRs (3–55 years), 01_AE (20–72 years), and 02_AG (3–43 years). Proportions of diagnoses by age groups are reported below.

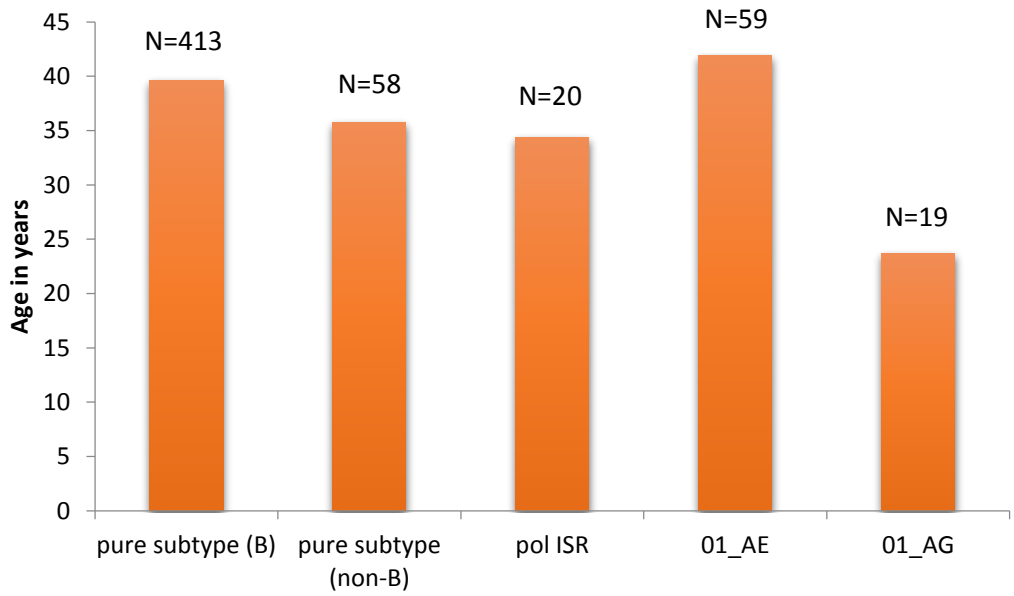


Figure 15a. Average age at diagnosis by subtype.

4.6.1 Children and adolescents/young adults infected with HIV-1

There were 62 child, adolescent and young adult cases, defined as any person aged 24 or under at time of diagnosis. They comprised 11% of the subtyped cohort. The proportion of child and adolescent cases increased slightly over time, from 11% (19/171) in 2000–2004, to 9% (20/216) in 2005–2009 and 13% (23/182) in 2010–2013. There were significantly more females in this young cohort (32%, 20/62) compared with those diagnosed at ages 25–50 (14%, 58/408), or over 50 years of age (8%, 8/97), ($\chi^2 = 17.41$ $p \leq 0.005$), Figure 15b.

Twenty-six cases were of a non-B or *pol*-ISR subtype (42%). The overall observed frequency of non-B infections in the under 25 years cohort was over double the expected frequency ($p \leq 0.005$).

Child and adolescent cases were further broken down into the following groups.

4.6.1.1 Characterization of child cases (birth–7 years)

This age group represented 0.9% of total cases (5/569) consisting of three females (all born overseas) and two males (one born in Australia, one overseas). All were MTCT cases. The only case acquired in Australia was the one male born in Australia. Four cases had non-B viruses (2 02_AG and 2 C) and one case had *pol*-ISR (A/AE).

4.6.1.2 Characterization of child cases (8–14 years)

This age group represented 0.7% of total cases (4/569) and there were three males and one female. All were born overseas and acquired a 02_AG virus overseas, three through medical procedures in Central Asia and the fourth was an unknown transmission from Africa.

4.6.1.3 Characterization of adolescent and young adult cases (15–24 years)

This age group represented 9% of total cases (53/569) and 70% (37) were male. All but one of the 16 females reported heterosexual exposure (including three with IDU risk), while

one reported a direct blood exposure. Ten acquired their infection in Australia (nine born in Australia, one in sub-Saharan Africa), and six acquired it overseas (four born in sub-Saharan Africa, two born in Asia).

Eight of the males reported heterosexual exposure (including one with IDU risk), the remaining 29 males reported MSM exposure, including five who reported dual IDU risk. Thirty males had a subtype B infection and 25 males acquired their subtype B or non-B infection in Australia.

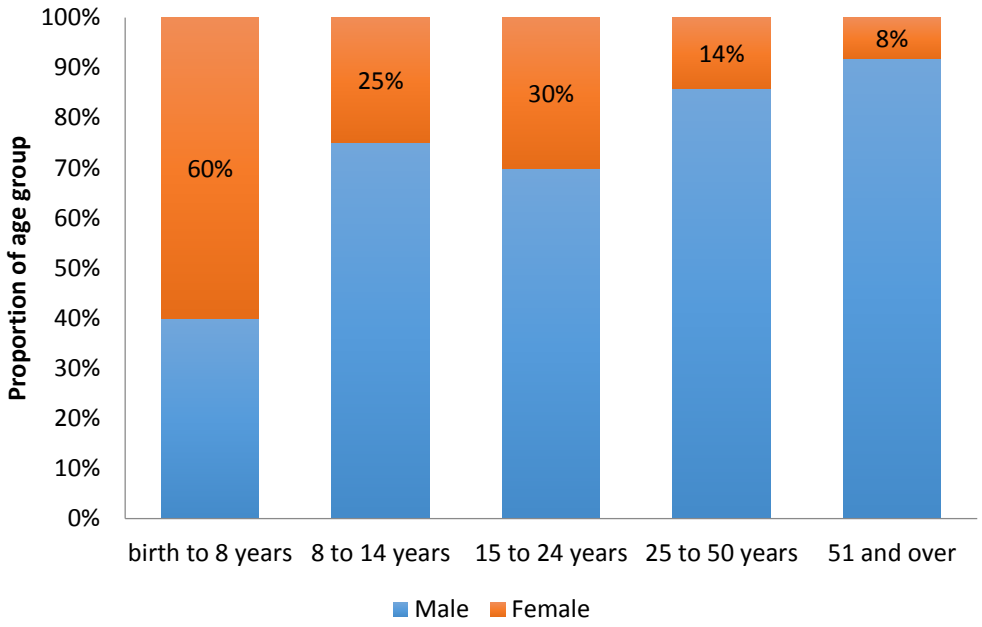


Figure 15b. Proportion male and female by age group.

4.6.2 Adult cases (25–50 years)

This age group represented 72% of total cases (408/569) and 86% (350) were male. Most cases were subtype B (75%, 307) of whom 93% (287) were male. Of those males with subtype B, 75% (214/287) reported MSM exposure within Australia, and 76% (164/214) of them were also Australian-born.

The majority of non-B cases in the entire cohort were in this age bracket, 65% (101/156),

which included 75% of total *pol*-ISR cases (15/20). Thirty-eight percent (38) of these non-B cases were female, and all but five females reported heterosexual transmission (two were unknown and three were from direct blood exposure). A large proportion of females (82%; 31/38) with non-B viruses acquired them overseas, while 85% (17/20) of females with subtype B virus acquired the infection within Australia.

Of the 63 males with a non-B or *pol*-ISR virus, only 16% (10) reported MSM exposure. The majority (76%, 48) reported heterosexual exposure. Almost all (87%, 55) males with a non-B/*pol*-ISR virus acquired the infection overseas and 75% of all males with a non-B/*pol*-ISR virus were born overseas (47/63).

4.6.3 Middle aged cases and over (51 years and over)

This age group represented 17% of total cases (97/569) and 92% (89) were male. The predominant subtype in this age category was subtype B (71%, 69), and of these subtype B cases, 78% (54) were transmitted through MSM risk exposure, 19% (13) through heterosexual exposure, and one through direct blood contact. The transmission risk was not reported for one person. Nearly all of these subtype B cases were acquired in Australia (93%, 64), mostly by people who were born in Australia (67%, 43/64), but there was also a subset of overseas-born people who acquired a subtype B infection in Australia (22%, 14). Nine people did not have birth region listed.

This age bracket accounted for 18% of total non-B cases in the entire study cohort (25/136) and 15% of total *pol*-ISR cases (3/20). Of these 28 non-B/*pol*-ISR cases 82% (23) were acquired overseas; 54% (15) were carried by Australian-born people, 43% (12) were carried by overseas-born people, and one person did not have birth region listed. Five of the eight women in this age group carried non-B/*pol*-ISR virus.

People with non-B or *pol*-ISR cases in this age group reported transmission as follows: 68% (19) heterosexual, 25% (7) MSM, 4% (1) direct blood exposure, and 4% (1) unknown

exposure.

4.7 Subtype distribution by region of birth

People born in Australia constituted 59% (335/569) of total cases, 28% (162) were born overseas while 13% (72) had no recorded country of birth. The ratio of male to female cases differed by region of birth, see Table 11. There were no females in the European-born cohort, the highest male to female ratio was seen in the cohort of unknown birth region, with a ratio of 11.0 males to every female. This was followed by the Australian-born cohort (9.5:1), American-born (3.5:1) Central Asian/Middle Eastern (2.7:1), Asia (1.5:1), and sub-Saharan Africa (1.37:1).

4.7.1 Australian-born population

The majority of the Australian-born cohort had subtype B viruses (89%, 298/335) and 92% (275) of these were also acquired in Australia. Those in the Australian-born, Australian-acquired cohort were predominantly males who reported MSM transmission (76%, 210/275), including a subset of 22 males who reported IDU risk. The remaining 26% (65/275) were reported as follows: 23% (62) heterosexual transmissions (including 30 IDU risk), 0.4% (1) direct blood exposure, and 0.8% (2) unknown exposure.

Eleven percent (37) of Australian-born people had non-B or *pol*-ISR viruses, 68% acquired the infection overseas (25) and the remaining 32% (12) acquired the infection within Australia. Males comprised 78% (29) of non-B cases. The primary transmission exposure was heterosexual (65%, 24), followed by MSM (30%, 11). There was one MTCT exposure, and one unknown exposure. The eight female cases were all transmitted through heterosexual exposure, six in Australia and two overseas.

The majority of the 37 Australian-born people with a non-B virus had 01_AE (62%, 23) followed by subtype C (6 cases) 02_AG (2 cases) subtype D and G (1 case each) and *pol*-

ISR variants (4 cases).

4.7.2 Overseas-born population

The majority of the overseas-born population had a non-B or *pol*-ISR virus (67%, 109/162), and acquired it overseas (87%, 95/109). Only one-third of the overseas-born cohort had a subtype B virus (33%, 53/162) and 70% (37/53) of these had acquired it in Australia. Of those that acquired a subtype B virus overseas, the region of birth was as follows: US/Europe (n=9), Asia (n=3), sub-Saharan Africa (n=2) and Central Asia/Middle East (n=1).

4.7.2.1 Central Asia/Middle East

There were 11 cases originating from Central Asia or the Middle East, three following medical procedures, four heterosexual transmissions, two MSM and two unknown. Ten were acquired overseas and one stated unknown. There was only one subtype B and it was acquired overseas through MSM.

Three cases were 02_AG, all from medical procedures overseas and all aged 10 at diagnosis (central Asia cases). Two cases were 01_AE, both heterosexual transmissions, and one case was subtype C in MSM. The remaining four cases were *pol*-ISRs, diagnosed between 2009 and 2013; one A/D and one A/B (both unknown risks), one 02_AG/B and one A/AE (both heterosexual).

4.7.2.2 Asia

Over 70% (30/42) of Asian-born people had non-B/*pol*-ISR viruses. Of the 30 non-B cases 77% (23) were acquired overseas and the predominant non-B strain was CRF_AE (61%, 14), C (22%, 5), followed by CRF_AG (9%, 2), G (4.5%, 1), and *pol*-ISR B/C (4.5%, 1). The seven non-B cases acquired in Australia were all 01_AE, comprising five females through heterosexual transmission, and two males through MSM.

Twelve people had a subtype B virus, 11 through MSM transmission and one female through heterosexual transmission. Nine of these Asian-born people were infected in Australia and three infected overseas (including the female).

4.7.2.3 Europe/America

There were 45 people born in Europe or America, 80% (36) of these had a subtype B virus and 35 of these were male. The female acquired a subtype B infection overseas through heterosexual transmission and eight males also acquired the infection overseas, six through MSM, and one through blood exposure and heterosexual contact respectively. Twenty-six males acquired subtype B infection locally, predominantly through MSM exposure.

There were nine non-B cases, including a young American female with a subtype C virus who acquired the infection overseas through direct blood exposure, and eight older European males, seven of which reported heterosexual contact overseas. One male reported direct blood contact overseas and had an AE/B recombinant virus, the other male reported IDU risk within Australia.

4.7.2.4 Sub-Saharan Africa

Just over 10% (64) of all subtyped cases were people born in sub-Saharan Africa, 37 males and 27 females. Almost all (89%, 57) acquired the infection overseas. There were 80% (51) non-B cases, 14% (9) *pol*-ISR cases and 6% (4) subtype B cases, with the latter all contracted through MSM exposure.

The main route of transmission for African-born people was heterosexual (77%, 49), followed by MSM (9%, 6), direct blood contact (4.5%, 3) and MTCT (4.5%, 3), with the last two of these all acquired overseas.

4.8 Demographic analysis of subtype distribution

Case characteristics differed by subtype. Subtype B was predominantly an MSM infection, acquired in Australia by men born in Australia. Non-B infections were predominantly acquired overseas by people born overseas. There was overlap between the two. Each subtype/CRF is reported separately below.

4.8.1 Subtype A (n=5)

All five subtype A cases were African-born people who acquired the infection overseas through heterosexual transmission. Four were females, two aged 25 and under, and two aged in their 50s.

4.8.2 Subtype C (n=49)

Of the 49 C cases, 61% (30) were people born in sub-Saharan Africa, 12% (6) born in Australia, 10% (5) Asia, 4% (2) Europe, 2% (1) America, 2% (1) central Asia/Middle East and 8% (4) birthplace unknown. Only one of the 49 subtype C cases was acquired in Australia (Australian-born female through heterosexual transmission). The 48 overseas-acquired cases were transmitted primarily through heterosexual contact (77%, 37), followed by blood contact or medical procedure (8%, 4), MSM (5.5%, 3), MTCT (4%, 2) and unknown risk (5.5%, 3).

4.8.3 Subtype D (n=2)

There were two D cases, one an African-born male infected overseas and the other an Australian-born female infected in Australia. Both reported heterosexual transmission.

4.8.4 Subtype G (n=2)

There were two G cases, one a male born in Australia and infected overseas through MSM transmission, and one a female born in Asia and infected overseas through heterosexual transmission. The female was diagnosed in 2012, the male in 2013.

4.8.5 CRF01_AE (n=59)

Of the 59 01_AE cases, 39% (23) were people born in Australia, closely followed by 36% (21) born in Asia, then 8% (5) Europe, 5% (3) sub-Saharan Africa, 3% (2) central Asia/Middle East and 8% (5) of unknown origin.

Nearly one-quarter (22%, 13) of 01_AE cases were acquired in Australia, seven by Asian-born people, four Australian-born, one European-born and one African-born. Ten were acquired through heterosexual transmission and three through MSM.

The 78% (46) overseas-acquired cases were transmitted primarily through heterosexual contact (78%, 36), followed by MSM (17%, 8) and blood contact or medical procedure (4%, 2). Year of diagnosis for all 01_AE cases ranged from 2000 to 2013.

4.8.6 CRF02_AG (n=19)

Of the 19 02_AG cases, 63% (12) were people born in sub-Saharan Africa, 10.5% (2) born in Australia, 10.5% (2) Asia, and 16% (3) Central Asia/Middle East. Only three of the 19 cases were acquired in Australia, (one Australian-born female, one African-born female and one African-born male, all through heterosexual transmission).

The 16 overseas-acquired cases were transmitted primarily through heterosexual contact (56.5%, 9), followed by blood contact or medical procedure (25%, 4), MTCT (12.5%, 2) and unknown risk (6.5%, 1). Year of diagnosis for all 02_AG cases ranged from 2004 to 2013.

4.8.7 pol-ISR *cases (n=20)*

Of the 20 *pol*-ISR cases, 12 ISR patterns were detected, see Table 12. Three-quarters (15) of *pol*-ISR cases were reported as being acquired overseas while 75% (15) of persons were born overseas. Thirteen of these cases fell into both categories.

The 20 *pol*-ISRs accounted for 3.5% of all subtyped cases and 13% of all non-B cases. The proportion of *pol*-ISRs has increased across the time periods, from 1.2% (2/171) of all

subtyped cases diagnosed between 2000 and 2004 to 6% (11/182) between 2010 and 2013 (Figures 16 and 17). Of the non-B cases, the proportion of *pol*-ISRs increased from 6.1% (2/33) to 15.5% (11/71) during the same time periods, and accounted for 26.3% of non-B cases in 2013 alone (5/19). Correspondingly, the proportion of other non-B subtypes/CRFs has declined, Figure 16.

Of the 15 *pol*-ISR cases in persons born overseas, 45% (9/20) were from sub-Saharan Africa, 20% (4/20) from Central Asia/Middle East, and one each (2/20; 5%) from Europe and Asia. The remaining 25% (5/20) were either born in Australia (four cases) or were of unknown origin (one case). Three of the four Australia-born cases acquired the infection in Australia, one child through MTCT and two were MSM. One MSM acquired the infection overseas.

Five of the 20 *pol*-ISRs occurred in females; all these women were born overseas and contracted HIV overseas. All but one had A/AE variants, which accounted for four of the six *pol*-ISR A/AE cases in the dataset, Table 12. One diagnosis occurred in 2004, two in 2009 and one in 2013. The other female had the only A/D strain in the dataset. She was born in the Middle East and diagnosed in 2013, Table 13.

Of the 15 males with *pol*-ISRs, 14 were diagnosed between 2008 and 2013, and nine were diagnosed between 2010 and 2013. The complex variants in the male cohort were more diverse than in the female cohort.

Four males were born in Australia and three of these acquired the infection in Australia, one through MTCT (A/AE), one through MSM (B/D), and one through MSM (02_AG/B), Table 13. The other Australia-born male acquired the infection overseas (MSM, B/G variant).

Ten males were born overseas and one male did not list country of birth. Eight of the overseas-born males and the male of unknown origin acquired the infection overseas; two

reported blood exposure or medical procedure (01_AE/B and A/01_AE), four reported heterosexual transmission (two 02_AG/B, one B/C, one 02_AG/01_AE), two reported dual heterosexual and IDU risk (both B/C), and one did not report risk (A/B), Table 13. The male with the A/B recombinant originated from the same country in the Middle East as the female with the A/D recombinant, diagnosed one year apart.

Two remaining two overseas-born males originated from Africa but acquired the infection in Australia, both diagnosed with B/G variants; one in 2011 through heterosexual contact, and one in 2013 through MSM, Table 13.

Table 12. Subtype and CRF pattern variations of *pol*-ISRs.

Subtype and/or CRF	**No of different patterns**	**No of diagnosed cases**
A & 01_AE	2	6
B & C	2	3
B & 02_AG	2	5
01_AE & 02_AG	1	1
D & A	1	1
B & G	1	1
B & D	1	1
B & 01_AE	1	1
B & A	1	1
	12	**20**

Table 13. Characteristics of PR/RT discordant recombinant cases, newly diagnosed in South Australia 2000–2013.

Subtype: PR/RT	Sex	Yr dx	Age dx	Location infection acquired	Transmission Risk	Region of Birth
B/C	M	2001	35	Overseas	Het/IDU	N/A
A/01AE	F	2004	40	Overseas	Heterosexual	SSA
A/ 01AE	M	2008	3	Australia	MTCT	Australia
B/C	M	2008	26	Overseas	Het/IDU	Asia
A/ 01AE	F	2009	29	Overseas	Heterosexual	CA/ME
A/ 01AE	M	2009	51	Overseas	Blood/Procedure	SSA
A/ 01AE	F	2009	46	Overseas	Heterosexual	SSA
02_AG/B	M	2009	21	Overseas	Heterosexual	SSA
02_AG/01_AE	M	2009	35	Overseas	Heterosexual	SSA
02_AG/B	M	2010	39	Australia	MSM	Australia
B/01_AE	M	2011	38	Overseas	Blood/Procedure	Europe
B/02_AG	M	2011	26	Australia	Heterosexual	SSA
B/D	M	2011	53	Australia	MSM	Australia
B/A	M	2012	33	Overseas	Unknown	CA/ME
C/B	M	2012	29	Overseas	Heterosexual	SSA
B/02_AG	M	2013	28	Overseas	Heterosexual	CA/ME
B/02_AG	M	2013	27	Australia	MSM	SSA
B/G	M	2013	32	Overseas	MSM	Australia
01_AE/A	F	2013	55	Overseas	Heterosexual	SSA
D/A	F	2013	43	Overseas	Unknown	CA/ME

Key: Yr dx (year of diagnosis), Het/IDU: (heterosexual with IDU risk), N/A (not available), SSA (sub-Saharan Africa), CA/ME (Central Asia/Middle East), MTCT (mother to child transmission).

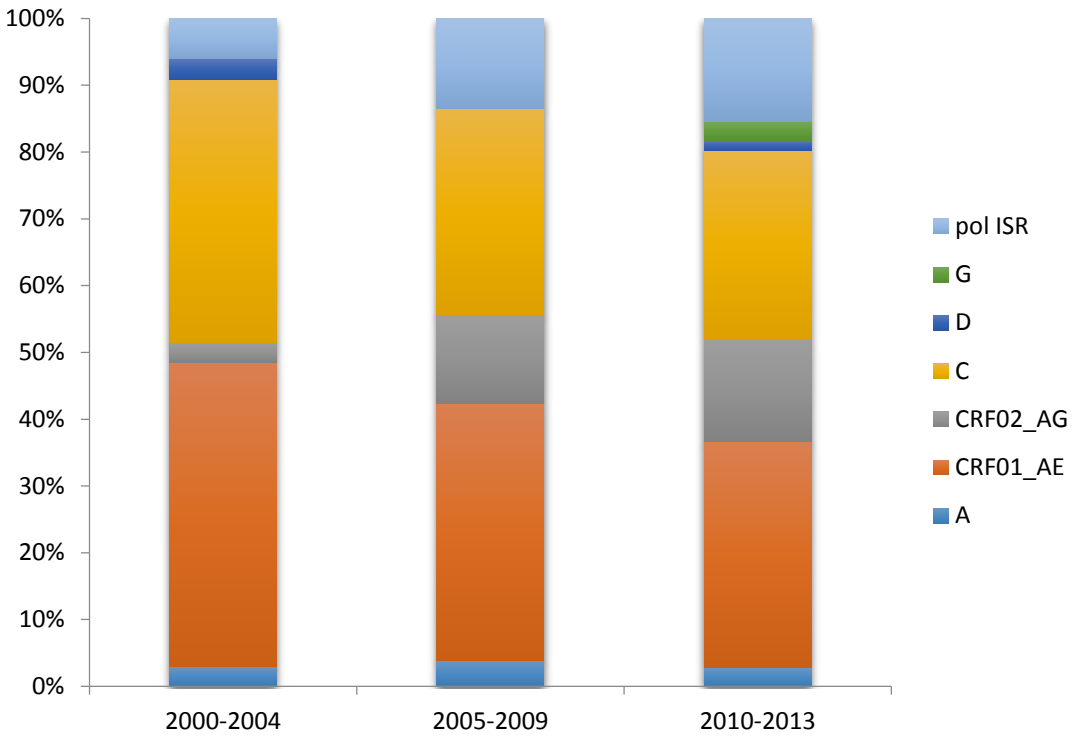


Figure 16. Proportion of non-B subtypes/CRFs/*pol*-ISRs, by year group.

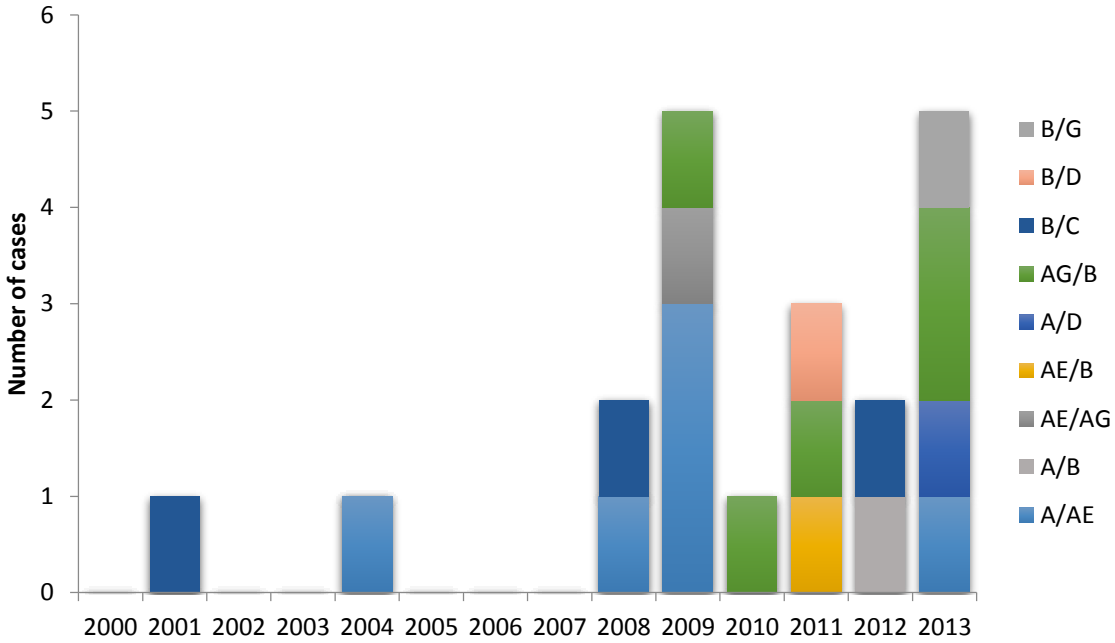


Figure 17. *pol*-ISR cases diagnosed in South Australia by year, 2000-2013. Each group may contain more than one pattern, i.e. A/01_AE contains some variants that have an A PR sequence and 01_AE RT sequence, and some that contain a 01_AE PR sequence and A RT sequence.

*4.8.7.1 Timeline of pol-*ISR *diagnosed cases*

Table 13 and Figure 17 show the first *pol*-ISR case was diagnosed in 2001 and was a B/C recombinant. The person was a male between 25-50 years who acquired the infection locally either by heterosexual contact or IDU, Table 13.

In 2004, the first of five *pol*-ISR females was diagnosed. She was also aged between 25-50 years, African-born and diagnosed with an A/AE strain through heterosexual contact overseas.

In 2008, a 3-year-old boy was diagnosed with an A/AE strain contracted through MTCT in Australia, this was the first reported *pol*-ISR local transmission. An Asian-born male aged between 25-50 years was also diagnosed in 2008, with a B/C strain reported as being acquired overseas through heterosexual contact or IDU.

In 2009 there were 5 diagnosed *pol*-ISR cases. Four were African-born people and one Central Asian-born person; all reported overseas transmissions. The Central Asian-born person was the second female *pol*-ISR case, aged between 25-50 years and diagnosed with an A/AE strain and reported heterosexual contact overseas. Of the four African-born people, one was a female also aged between 25-50 years, diagnosed with A/AE who reported heterosexual contact overseas. Two of the males were also in the 25-50 age bracket, while the other was under 24, diagnosed with A/AE, AG/AE and AG/B respectively. The A/AE male reported an overseas medical procedure as the transmission risk, while the AG/B and AG/AE males reported heterosexual contact.

In 2010 a second AG/B strain was diagnosed, infecting a male also in the 25-50 year age bracket who was born in Australia and reported local acquisition through MSM transmission.

In 2011 there were three variations of subtype B *pol*-ISR infection, two carried by males in the 25-50 year age bracket. A European-born male was diagnosed with a B/AE strain,

reported as acquired overseas through blood contact or a medical procedure. An African-born male was diagnosed with a B/AG strain, reported as a local acquisition by heterosexual contact, while an Australian-born male over 51-years was diagnosed with a B/D strain, acquired locally through MSM transmission.

There were a further two subtype B *pol*-ISR infections in 2012, again carried by males in the 25-50 year bracket. A Central Asian-born male acquired a B/A infection overseas, transmission risk undisclosed, and an African-born male acquired a C/B infection overseas through heterosexual contact.

There were a further five diagnoses of these complex *pol*-ISR strains in 2013. Only one was reported as being acquired in Australia, a B/AG strain carried by an African-born MSM male in his late 20's.

The remaining four cases were acquired overseas, two males and two females. The males were both between the age of 25-35 years, one was born in Australia and carried a B/G strain transmitted by MSM, and the other was born in the Middle East and carried a B/AG strain transmitted heterosexually through a partner born overseas.

The females were aged between 40 and 55 years, one an African-born woman with a 01_AE/A strain transmitted heterosexually via a partner born overseas, and the other a Central Asian-born woman with an undisclosed transmission risk carrying a D/A strain.

4.9 Multivariate regression: subtype associations with comparator variables

Multivariate logistic regression analyses were conducted to see if significant differences remained between the subtypes when other variables were controlled for. The subtype B group was used as the baseline for comparison and other independent variables (acquired Australia/acquired overseas, Heterosexual/MSM transmission, Australian-born/overseas-born, age < 25 years / 25 and over, and male/female) were controlled for.

As shown in Table 14, there was a strong association between subtype C and overseas acquisition, with subtype C cases 34.5 times more likely than subtype B cases to be acquired overseas (95% CI 6.3–188.8, $p \leq 0.001$). Non-B cases were 20 times more likely to be acquired overseas compared with B cases (95% CI 9.4–43.4, $p \leq 0.001$).

Non-B cases were eight times more likely than B cases to be acquired through reported heterosexual transmission (95% CI 3.6–18.2, $p \leq 0.001$), which rose to 36 times more likely for subtype C cases (95% CI 6.4–203.7, $p \leq 0.001$) and 37 times more likely for subtype A, D, G and 02_AG cases combined (95% CI 3.3–410.3, $p = 0.003$).

Non-B cases were almost six times more likely (95% CI 2.6–11.0, $p \leq 0.001$) among people born overseas, and this stayed significant at $p \leq 0.05$ for 01_AE, subtype C, *pol*-ISRs and subtype A, D, G and 02_AG cases combined.

Non-B cases were three times more likely (95% CI 1.2–8.5, $p = 0.02$) among females, this was just under statistical significance for 01_AE (95% CI 0.9–10.8, $p = 0.06$), and not statistically significant for the other categories.

The age at diagnosis of HIV infection was not different between B and non-B cases.

Table 14. Multivariate Odds Ratios in newly diagnosed HIV cases, 2000–2013, when compared to the proportion of subtype B infections.†

Characteristics (Comparator group)					
Clade	**Acquired overseas (Acquired Australia)**	**Heterosexual (MSM)**	**Overseas-born (Australian-born)**	**Age <25 (≥ 25)**	**Female (Male)**
Non-B	20.2	8.1	5.4	1.1	3.2
	(9.4–43.4)	(3.6–18.2)	(2.6–11.0)	(0.4–3.1)	(1.2–8.5)
CRF_AE	23.9	9.3	4.3	0.8	3.3
	(8.8–65.0)	(3.4–25.1)	(1.8–10.6)	(0.2-2.9)	(0.9–11.3)
C	34.5	36.1	10.6	0.9	1.8
	(6.3–188.8)	(6.4–203.7)	(2.5–44.3)	(0.1–5.7)	(0.3–9.2)
pol-ISR	6.5	6.6	10.4	0.6	1.0
	(1.5–27.8)	(1.5–28.7)	(2.3–47.5)	(0.05-6.5)	(0.2–5.7)
Other	11.7	37.0	15.6	4.3	4.5
	(2.2–62.8)	(3.3–410.3)	(2.8–88.5)	(0.7–25.3)	(0.8–24.2)

Key: †95% CIs are indicated in brackets. Proportion of subtype B infections was used as the baseline for comparison. All associations in the first three columns were statistically significant at $p \leq 0.05$. There were no statistically significant associations in the fourth column (age); age and non-B ($p = 0.87$), age and 01_AE ($p = 0.73$), age and C ($p = 0.89$), age and *pol*-ISR ($p = 0.66$), age and other ($p = 0.10$). The only association that was statistically significant in the fifth column (sex) was 'non-B' at $p \leq 0.05$. Sex and 01_AE ($p = 0.06$), sex and C ($p = 0.51$), sex and *pol*-ISR ($p = 0.96$) and sex and other ($p = 0.08$) were non-significant. 'Non-B' includes all non-B subtypes including *pol*-ISRs and CRFs. The category 'Other' consists of subtypes A, D, G and 02_AG.

4.10 Summary

People infected with B or non-B virus subtypes represented highly distinct populations, with subtype B virus predominantly acquired in Australia and non-B virus predominantly acquired overseas. These two groups combined comprised 85% of all newly diagnosed infections.

The majority of people with subtype B infections were MSM who acquired their infection within Australia, while non-B infections were mainly acquired heterosexually

overseas and almost 40% were in females. Although the majority of subtype B infections were acquired in Australia by Australian-born MSM, one-fifth of subtype B infections were heterosexually acquired by males and females who were predominantly born and infected in Australia.

The majority of non-B infections were acquired overseas by people born overseas or people traveling to areas of high HIV prevalence.[8,9,69] These figures are comparable with findings by Chalmet *et al*. (2010) in Belgium, in a similar sized cohort.[165] The number of newly diagnosed pol-ISR strains is increasing over time, reflecting growing global genetic variability, especially in LMICs.[112,204,205]

4.11 Discussion

This is the first Australian study to investigate genotypic diversity and trends in subtype distribution of the HIV epidemic spanning the past 14 years. At present, no formal national HIV-1 molecular surveillance program exists in Australia. In 2000, South Australia implemented routine baseline drug resistance testing, providing viral sequence data combined with notification information to yield enhanced comprehension of South Australian HIV molecular epidemiology.[14] As HIV continues to evolve, and migration patterns change and tourism increases, it is important to monitor this geographic diversity in order to understand and respond to transmission patterns.[8,71]

Despite ongoing programs and improved access to testing and treatment, the number of new HIV diagnoses has increased in South Australia, from 23 in 2000 to 56 in 2013. Further, there has been a significant change in subtype distribution among newly diagnosed people in South Australia, with the proportion of non-B infections increasing over the past 14 years from 9% in 2000 to 34% in 2013, including the emergence of a number of *pol*-ISR variants. These non-B infections are predominantly carried by individuals born overseas or Australian-born people who were infected overseas.

Differences were identified by place of acquisition, region of birth, gender and age at diagnosis. Non-B infections represented almost 30% of all cases in this study, including 4% that were *pol*-ISRs. This proportion rose to nearly half of all new diagnoses in 2010 but had dropped to just over one-third of all cases by 2013. This increase in non-B infections is similar to reports from other developed countries.[9,117]

4.11.1 Route of infection

Australian-born MSM with locally acquired subtype B continue to represent the largest HIV risk group in South Australia, similar to recent findings in Asia.[92] This proportion is decreasing over time, however, because of the increase in heterosexual infections acquired within Australia and overseas, supporting epidemiological trend analyses in Australia, Canada, Europe and the United States.[14,103,109,165,194,206] The number of HIV infections diagnosed in the MSM IDU community is growing and a small but increasing number of MSM males are acquiring non-B infections locally and overseas, reflecting findings in the UK[70] and the recent Victorian study.[69] Although national HIV incidence has remained stable in the MSM population over the past decade, the proportion of undiagnosed infections in the South Australian MSM community is estimated to be around 20%,[207] and this figure may actually be higher with the introduction of new infections from overseas and circulation of subtype B among intravenous drug users.

Globally, heterosexual transmission is the major route of HIV infection.[89] In South Australia we are seeing a shift in the epidemic toward this, with a decrease in the number of cases among MSM and an increase among heterosexual men and women, including a large proportion from sub-Saharan Africa. This shift has also been noted in Europe.[208] Since 2000, heterosexually acquired infections in South Australia have almost doubled from 20% of all new diagnoses to 39%, but this is not as high as the 4-fold increase seen in the UK since 1996.[14]

Corresponding with an increase in heterosexual infections, there has been a significant increase in non-B infections from overseas. These infections were acquired through unprotected heterosexual sex by people born overseas who migrated to Australia,[209] or by Australian-born people who travelled to countries where there is a high prevalence of HIV.

There has also been a relatively recent increase in newly diagnosed infections acquired through direct blood exposure including MTCT. These are mostly non-B subtypes transmitted overseas, including a rising number of infections transmitted by IDU in Asia,[62,210] and a high prevalence of iatrogenic and MTCT infections in children in central Asia and Africa.[209]

Twenty-one people were infected through direct blood exposure in this cohort and nearly all were diagnosed from 2008 onwards. Three infections occurred in South Australia. Two of these were in adult males with subtype B through direct blood contact with origin not reported, and the other in a young boy infected with non-B through MTCT. The child was diagnosed at three years of age and his mother was born overseas but newly diagnosed in Australia when the child was aged two. Genotypic resistance testing at time of diagnosis found no drug resistance in the mother and multiple drug resistance in the infant, suggesting that infection occurred through breastfeeding or that the mother had been previously diagnosed and treated overseas and the virus she carried had reverted back to wild-type at time of resistance testing.[211]

All other direct blood exposure infections were acquired overseas, predominantly non-B or *pol*-ISR and the majority were acquired in sub-Saharan Africa or Asia, including four MTCTs. It is well documented that HIV infection transmitted through intravenous drug use and unsafe medical practices including during birth is a major concern in both Africa and Asia.[62,209,210] There were three child cases from central Asia in children aged10 years at time of diagnosis, with medical procedures reported as the transmission risk. HIV infection in the

child population is still fairly rare in Central Asia, but of those that have been reported, the majority are nosocomial.[62]

4.11.2 Infection by sex

The proportion of females in the HIV-infected population has also changed and over 20% of newly diagnosed infections between 2010 and 2013 occurred in females. The relative proportion of females infected with subtype B was low, explained at least in part by the high representation of MSM in the B cohort. In addition, the majority of females in the study were born overseas and acquired their infection overseas where non-B subtypes are more frequently circulating. Australian-born females predominantly had subtype B infections acquired in Australia.

All but one of the female subtype B infections were heterosexually acquired (including seven dual IDU risk) while one woman reported direct blood exposure overseas. It is likely these women were infected through bisexual males who were infected through MSM contact or intravenous drug use.

The majority of female non-B infections were also heterosexually acquired (including two dual IDU risk). However, eight females from high prevalence countries[62,209–211] were infected overseas through direct blood exposure, which constituted approximately half of all direct blood transmissions overseas. The proportion of females compared with males was highest among children seven and under, young adults, and people born in Africa or Asia.

4.11.3 Multivariate analysis

When the other variables were controlled for, non-B cases were still eight times more likely to be transmitted through heterosexual contact than subtype B cases, and this increased to 37 times more likely for subtype C and 26 times more likely for all other non-B types excluding 01_AE. There was also a small number of non-B cases transmitted through male-

to-male sex, predominantly 01_AE infections acquired by Asian or Australian-born males in Australia and overseas.

Despite the increase in South Australia in heterosexually acquired non-B HIV infections, MSM are still a critical at-risk population. Targeted prevention, intervention and treatment strategies are needed for MSM and heterosexual populations, with continued surveillance and discussion about best practice and treatment options for direct blood exposure infections from overseas.[109]

As well as a growing number of imported non-B infections, there is evidence of local non-B transmission, including *pol*-ISR variants. These infections are predominantly 01_AE, but subtypes C, D and 02_AG have also been diagnosed, and there were five people with *pol*-ISR infections. The first diagnosis of a locally acquired non-B infection occurred in 2001, but the prevalence has increased from 2006 onward, predominantly transmitted to Australian, Asian and African-born people through MSM or heterosexually. Research suggests that people predominantly form sexual partnerships with those in their own ethnic group,[212] but evidence from Europe shows both migrant and non-migrant populations are infected with non-B subtypes.[212] The findings from the current study mirror both findings. HIV infection seems to occur between migrants within ethnic groups once they are settled in Australia, but there is also evidence of sexual contacts between migrants and Australian-born people.

The average age at diagnosis was 40 years, but females were diagnosed at a younger average age. Age at diagnosis also differed by subtype and a significantly higher proportion of non-B cases was diagnosed under the age of 25 compared with the B cohort. When location acquired, country of birth, and risk behavior were controlled for this was significant only for persons with subtype A, D, G or CRF_AG. These clades are common in Africa and Asia, where over one-third of these young people were born. Young people born in Africa

reported heterosexual contact or MTCT as the risk exposure while young Asian-born people reported heterosexual, MSM and direct blood exposure.

A significantly higher proportion of people under the age of 25 with subtype A, D, G or 02_AG was female. Just over half of these young females were born in sub-Saharan Africa, and all young females regardless of birth region were infected heterosexually. It is likely the young Australian-born women were infected with a non-B subtype by males born outside Australia. It is well documented that gender inequity is associated with increased rates of HIV in women, and regions like sub-Saharan Africa have a strongly patriarchal culture that celebrates and promotes male dominance. Men often have multiple sexual partners and women are not empowered to refuse sexual intercourse.[213] It is important to address and navigate these risk factors and cultural issues with sensitivity while caring for young women infected with HIV.

Transmission of non-B subtypes and CRFs is rapidly expanding geographically, and the rise in non-B diagnoses may be a marker of more recent transmission events—some attributable to tourism and some to importation by people born in high-prevalence countries where multiple subtypes and CRFs circulate.

Over half of the global HIV population is infected with subtype C,[4,214,215] which is still widely circulating in Africa and India. Subtypes A and B follow, then CRF_AG and CRF_AE, the latter predominating in Asia.[52,215] In our cohort, subtype C accounted for fewer than 10% of total infections but over one-third of non-B infections. All but four of these were reportedly acquired overseas, the majority in people who originated from Africa but also in people originating from Australia, South America, Asia, and Europe.

The predominant non-B infection in the current cohort was 01_AE, with prevalence twice that of the 5% global average, and almost all cases were reported as being acquired overseas. Almost 40% of 01_AE cases were people born in Australia and 36% in Asia.

01_AE infections were mostly acquired overseas through heterosexual and MSM transmission, but there was some evidence of local transmission in Australia including through dual IDU risk. Hemelaar *et al*. found global CRF infections increased by over 50% between 2000 and 2007,[52,215] and a more recent study by Ambrosioni *et al*. found CRF infections increased significantly between 1997 and 2012.[117] The current study reflects this temporal increase in CRF: 01_AE and 02_AG infections comprised 9% of all new infections between 2000 and 2004, and almost 20% of all new infections between 2010 and 2013. Unlike subtype C and 02_AG infections, however, almost one-third of CRF cases occurred in Australian-born people and were most likely to have been acquired during overseas travel to Asia.

Subtypes A, D, G and CRF_AG are predominantly found in Africa, with a combined global prevalence rate of 10%.[52] Although prevalence in our cohort was only 5%, the majority were diagnosed from 2007 onward and were acquired overseas by people of African origin, possibly reflecting the increase in Australian immigration from this region. All people with subtype A infections were African-born, and all were acquired overseas heterosexually. A number of Australian and Asian-born people also imported or acquired locally subtypes D, G and 02_AG through heterosexual transmission or direct blood exposure.

4.11.4 Intersubtype recombination

Consistent with epidemiological studies from other countries and within Australia, an increasing proportion of non-B cases were recombinant viral subtypes.[60,69,204,216–218] There were twelve different *pol* intersubtype patterns detected. People with *pol*-ISR infections were approximately six times more likely to have acquired the infection overseas, almost seven times more likely through heterosexual acquisition and were 10 times more likely to have been born overseas. One-quarter of all cases with *pol*-ISR infection were female.

The first *pol*-ISR was diagnosed in South Australia in 2001, and another in 2004, with

different variants. No further intersubtype variants were diagnosed until 2008, but in subsequent years there has been a small but steady increase in the proportion of these infections, with an increase in new variant patterns.

4.11.5 Strengths and Limitations

This study has a number of strengths. South Australia was the first state to implement routine drug resistance testing as part of an enhanced surveillance system for all new HIV diagnoses. It was also the first state to link subtype data with demographic information for each case. This gave the opportunity to analyze all new diagnoses over a 14-year time period and report accurate data on prevalence rates in the HIV infected population, making this the largest epidemiological study of the HIV-1 epidemic in South Australia.

There were, however, some limitations. Our analyses focused exclusively on subtypes from a section of the *pol* gene because the standard practice is to sequence only the PR and RT region for drug resistance testing. Therefore, more complex recombinant viruses may have been undetected and the proportion of pure subtypes and CRFs may be over-represented. Further subtype validation should be conducted with alternative HIV genes, such as *env*[219] or, even better, the whole genome should be determined if full genome sequencing is economically viable. Phylogenetic tree construction and online subtype analysis of multiple genes would be beneficial to explore whether there is a higher degree of viral complexity than is currently known. It would also allow us to verify the accuracy of self-reported demographic information.

4.11.6 Recommendations

The findings from the current study have public health implications, both for targeting specific at-risk populations and for assessing the potential increase of non-B subtypes including new recombinant variants within the domestic HIV-1 epidemic.[3] Targeted approaches are needed, including increased access to HIV testing upon entry to Australia

and subsidized access to treatment for people eligible for Medicare, including students, people on working visas, and those awaiting determination of permanent residency status, such as people born overseas with Australian-born partners.[220] Routine baseline subtype and drug resistance testing should be conducted for all newly diagnosed people Australia-wide, and for those who have been previously diagnosed overseas but are new cases in Australia.[221] Other targeted approaches should include education about prevention, HIV testing for Australians travelling overseas to high-risk areas, and targeted intervention strategies for MSM and heterosexual populations including specific interventions for people using intravenous drugs. Education about and use of pre- and post-exposure prophylaxis is another strategy that could be enhanced, together with revised policies and practices in LMICs for treatment and management of HIV during pregnancy, birth and breastfeeding.[222,223] A review of unsafe medical practices is needed to address the high rates of iatrogenic infection, especially in Asia and Africa.

There are limited data available on subtype differences and even fewer data available on non-B subtypes in countries where they are the major infection type. There is growing evidence, however, that suggests HIV strains differ in terms of virulence, transmission, and rate of progression.[4] A 10-year prospective study in Senegal found female sex workers with a non-A subtype had significantly shorter AIDS-free survival times.[78] A 2010 London study found the CD4+ cell decline was four times faster in subtype D cases and there was a higher virologic rebound at six months, after adjustment for baseline, gender, and ethnicity.[14] A study of Kenyan women found a greater than two-fold higher risk of mortality and faster rate of CD4+ cell decline in D cases compared with A after adjustment for viral load,[77] and in a Ugandan cohort, subtype D cases tended to develop AIDS earlier.[79] In Rakai, Uganda, the median time to onset and risk of progression to death were significantly shorter for cases with subtype D or CRF compared with A.[224] Each of these studies concluded that HIV

disease progression is affected by subtype and that this may have an impact on treatment decisions and policy in terms of initiation of therapy and future vaccine trials.[14,77,224]

Understanding genetic diversity is very important for the treatment of non-B subtypes. Many researchers now agree that although subtypes and CRFs appear equally sensitive to treatment, transmitted polymorphisms present before therapy may affect subtype-specific pathways of secondary resistance.[81] This, combined with suboptimal therapy and poor adherence in developing countries, makes them a prime target for accelerated drug resistance, both acquired and transmitted.[17] Current drug regimens targeted against subtype B may not be as effective in the long term for non-B subtypes and may lead to faster drug resistance.[89]

Reporting and interpretation of surveillance data can be problematic. Reporting of newly acquired infections does not necessarily mirror actual rates in the wider community because HIV diagnoses represent only the subgroup of people who have been tested and had an HIV-positive result. These are people who have relatively easy access to health services and feel confident to access those services.[225] In Australia and elsewhere, immigrants, visa holders, and refugees face major barriers when accessing health services for screening and treatment of HIV, arising from stigma, financial restrictions, limited support systems and English skills, and residence concerns.[208,225] Refugees in particular may be difficult to reach because of traumatic life experiences before arrival in Australia.

This major concern has prompted the UN to recognize migrants as one of the groups most vulnerable to HIV, and overseas-born people now comprise one-third of HIV notifications in Australia.[208,226] This figure has increased to 40% in South Australia as of 2010. Problems with access to testing and the steady influx of new arrivals from low and middle income countries with high HIV prevalence are likely to lead to an underestimate of HIV infections in these populations, a possible increase in local transmission of non-B

subtypes, and poor treatment adherence that could lead to TDR.[3,226]

The global spread of HIV diversity is highly dynamic with regard to epidemiological factors such as risk group and geographic location. The virus continually generates through mutation and recombination, and travel and migration assist in the transfer of this diversity between populations.[14] This study provides evidence that the HIV-1 epidemic in South Australia is changing from predominantly subtype B infections in the Australian-born MSM population to an increasing number of non-B infections within the heterosexual population, initially in those born overseas but increasingly with transmission to Australian-born persons. While 01_AE transmissions may relate to risk behaviors pursued during travel to high-prevalence areas, there appears to be an increasing tendency for transmission to occur within Australia, and within groups not previously the target of relevant and focused anti-AIDS education. It is also clear that over time the historical segregation between clades in terms of geography and risk group is becoming less distinct, with non-B infections occurring in the MSM population and subtype B infections occurring in overseas-born MSM males and Australian-born females.

4.11.7 Conclusion

The impact of the increasing number of non-B infections in the South Australian population on prevention efforts and treatment outcomes is as yet unclear. Ongoing surveillance and a deeper understanding of HIV variation, including factors and molecular mechanisms that affect transmission, replication, and resistance, are crucial for the development of appropriately targeted subtype-specific prevention and treatment options for populations most at risk.[8,14,52,214] Further evidence of subtype differences could drastically change the way we respond to the HIV epidemic.

CHAPTER 5: TRANSMITTED DRUG RESISTANCE

5.1 Overview

The increased availability of antiretroviral treatment including early HAART and later triple combination therapies and the problems associated with monitoring adherence to treatment has led to growing concern about TDR. Infection with a drug resistant strain of HIV can affect the number of treatment options offered and can also lead to further transmitted drug resistant strains being circulated in the population.

The immense genetic diversity within the HIV-1 virus may also present additional challenges because antiretroviral therapy has predominantly been developed and tested using subtype B virus.[17] There is scant research on subtype diversity and drug susceptibility/resistance,[227] although the research that has been conducted suggests that subtypes and CRFs are equally sensitive to treatment.[14,81] However, transmitted mutations present before treatment may affect subtype-specific pathways of secondary resistance to treatment.[81,82] Subtype-specific transmitted resistance could affect treatment efficacy, particularly in countries where treatment options are limited or suboptimal.[17,82]

5.2 South Australian HIV population

5.2.1 Current or past treatment

Of the 656 reported HIV-1 cases between 2000 and 2013, 569 (87%) had a drug resistance profile. Of these, only 2% of cases (nine males, four females) reported current or past treatment.

5.2.2 Treatment status unknown

There were 26% of cases (147; 134 males, 49 females) with a recorded treatment status of ‘unknown’. Almost half (49%; 72/147) were acquired overseas and of these only 27 had a reported subtype (19 were non-B). Of the 70 cases acquired locally, 32 (46%) had reported genotypes, 31 were subtype B.

5.2.3 Treatment naïve at time of genotype test

There were 72% of cases (496; 427 males, 69 females) with a reported treatment status of 'naïve' (never been treated for HIV infection) at the time of genotype testing. Of these, 70% (348/496) reported acquiring the infection in Australia, 28% (141) overseas and 2% (7) stated 'unknown'. Subtypes of treatment naïve cases are presented below.

Only treatment naïve cases are included in the remainder of this chapter.

This chapter reports on the prevalence of transmitted drug resistance using surveillance mutations from the 2009 Stanford University HIV drug resistance database and other non-polymorphic mutations that occur at the same positions as the surveillance mutations.

5.3 Resistance mutations in the treatment naïve cohort

5.3.1 Prevalence rates

Of the 496 treatment naïve cases, 77 displayed resistance mutations indicative of TDR according to the WHO list of surveillance drug resistance mutations ≤http://hivdb.Stanford.edu/pages/WHOResistanceList.html≥, with the inclusion of non-polymorphic mutations that also occur at the same sites as surveillance mutations. The overall TDR prevalence rate was 15%. Of the 77 TDR cases, 72% (56) had high-level resistance to at least one drug class, and four had high-level resistance to two drug classes, see Tables 15 and 17.

TDR prevalence for NNRTIs, NRTIs and PIs was 9% (46), 7% (35) and 2% (12) respectively, and overall TDR prevalence decreased significantly over time from 29% in 2000–2004 (43/150) to 4% in 2010–2013 (6/142, $p \leq 0.001$). TDR prevalence declined significantly for PIs from 6% (9/150; 2000–2004) to 0.7% (1/142; 2010–2013, $p = 0.02$), NNRTIs from 17% (25) to 2% (3, $p \leq 0.001$) and NRTIs from 13% (20) to 1.4% (2, $p \leq 0.001$).

5.4 Subtype/CRF distribution in the treatment naïve cohort

The subtype/CRF was concordant between the PR and RT regions in 481 of the 496 treatment naïve cases, with a distribution as follows: subtype B (74%; 365/496), 01_AE (10%; 51), subtype C (8%; 41), 02_AG (3%; 16), subtype A (0.8%; 4), subtype D and G (each 0.4%; 2).

There remaining 15 cases were *pol*-ISRs. The Stanford CPR tool ≤http://sierra2.Stanford.edu/sierra/servlet/JSierra≥ detected nine intersubtype/CRF patterns with a distribution as follows: A/01_AE variants (1%; 4/496), 02_AG/B variants (1%; 4), and one case each of B/A, B/C, B/D, B/G, 02_AG/ 01_AE and B/01_AE (0.2%; 1).

5.5 Resistance mutations by subtype/CRF

The prevalence of TDRMs was 19% (68/365) for subtype B cases, 6% (7/116) for non-B cases, and 13% (2/15) for *pol*-ISR cases. The proportion of TN subtype B cases harboring TDRMs decreased significantly from 33% between 2000 and 2004 (40/121) to 7% between 2010 and 2013 (6/88, $p \leq 0.001$), and TN non-B cases (including *pol*-ISRs) decreased significantly from 17% (3/29) to 0% (0/54, $p = 0.04$).

5.5.1 Subtype B cases

The following sections under 5.5.1 report single and multi-class drug resistance for all treatment naïve people carrying subtype B infections. Resistance to each drug class is then reported according to demographic characteristics.

5.5.1.1 Single and multi-class drug resistance

The majority of TN cases with TDRMs (88%, 68/77) were subtype B. Sixty percent (41/68) carried NNRTI mutations, 46% (31/68) carried NRTI mutations and 15% (10/68) carried PI mutations. Twelve subtype B cases carried mutations to multi-drug classes, Table 10.

Single class NNRTI resistance occurred in 31 subtype B cases. All but three had a single mutation (Table 15) and of those, 96% (27/28) carried the N mutation at position K103. Single class NRTI resistance occurred in 21 subtype B cases and 62% (13) had a single mutation. Single class PI resistance occurred in four subtype B cases, all with single mutations.

There were 12 subtype B cases (2% of all TN cases) with dual or triple class resistance. Dual NRTI/PI resistance and dual NNRTI/PI resistance occurred in two cases each, while dual NNRTI/NRTI resistance occurred in six cases. Two subtype B cases carried triple resistance, both of which had the M mutation at position L90 (PI), the N mutation as positions K103 (NNRTI) and D67 (NRTI), and the D mutation at position T69 (NRTI). One of the two triple resistance cases also carried the Q mutation at position K219 (NRTI) and the revertant L mutation at position T215, Table 15.

5.5.1.2 Demographic information by drug resistance class

5.5.1.2.1 NNRTIs

Of the 41 subtype B cases carrying NNRTI resistance 93% (38) were acquired in Australia, and 78% (32) of these subtype B cases were reported MSM transmissions. Eight cases were people reporting heterosexual contact and only two were female. One person reported an unknown exposure route.

Only six subtype B cases were people who reported being born overseas and all of them had NNRTI resistance. All were MSM and five reported acquiring the infection in Australia. Each of them carried the K103N mutation, including one person carrying triple class resistance with mutations D67N, T69D (NRTIs) and L90M (PI).

5.5.1.2.2 NRTIs

Of the 31 subtype B cases with NRTI resistance, 90% (28) were people who acquired

the infection in Australia, and 94% (29) of the subtype B cases were male. Twenty people were Australian-born, one person was born overseas and 10 cases had no listed country of birth. Twenty of the 31 cases (65%) reported MSM transmission.

5.5.1.2.3 PIs

Ten subtype B cases carried PI resistance, all were male, all reported MSM transmission, and eight reported acquisition in Australia. Five people were born in Australia, one person was born in Europe and the other four cases had an unlisted region of birth.

*5.5.2 Non-B cases (including pol-***ISR***s)*

The following sections under 5.5.2 report single and multi-class drug resistance for all treatment naïve people carrying non-B infections, including those carrying recombinant and complex recombinant forms. Resistance to each drug class is then reported according to demographic characteristics.

5.5.2.1 Single and multi-class drug resistance

Of the 131 non-B cases in this TN cohort, 7% (9) carried TDRMs (Table 16). Six cases carried a single mutation (two cases from each drug class), one case carried triple NNRTI mutations and a single NNRTI mutation, one case carried two NNRTI mutations and one NRTI mutation. One case carried four NNRTI mutations only. Of the eight different positions where NNRTI mutations occurred, five were affected in non-B cases, plus three of the 10 NRTI positions and two of the seven PI positions.

Single class resistance to each drug class occurred in the following cases: **NNRTI**: 1 subtype C, 1 01_AE and 1 A/AE *pol*-ISR case, **NRTI**: 1 subtype C and 1 subtype A case, **PI**: 2 01_AE cases.

Dual class resistance to NNRTIs and NRTIs occurred in two cases, a 02_AG case with

two NNRTI mutations and one NRTI mutation, and an A/AE *pol*-ISR case with two NNRTI mutations and one NRTI mutation (Table 15).

5.5.2.2 Demographic information by drug resistance class

Three of the non-B cases with TDR acquired the infection in Australia; two females carrying 01_AE and single PI mutations, both born in Australia who reported acquisition through heterosexual transmission are were diagnosed in 2006 and 2007 respectively. The other case was a MTC transmission occurring in Australia, a male with a *pol*-ISR A/AE variant that carried dual resistance to NNRTIs and NRTIs.

The remaining six cases were born overseas, four in Africa (subtype A, subtype C, 02_AG and *pol*-ISR A/AE), one in Europe (subtype C) and one in Asia (01_AE). All reported heterosexual transmission except the 02_AG case (unknown), and the *pol*-ISR case (blood contact overseas). All six reported acquiring the infection overseas. Only one case was female (subtype C), diagnosed in 2001.

Three of the four African-born people carried the Y181C mutation, the fourth carried K70E and was the only person in the entire cohort to carry that mutation.

Of note is the high resistance carried by both *pol*-ISR cases. One carried high-level resistance to NVP, EFV, RPV, ETR, 3TC, and FTC and low-level resistance to ddI and ABC, while the other carried high-level resistance to NVP, intermediate to high-level resistance to ETR and EFV, and low-level resistance to RPV.

5.6 Single or multi-drug class resistance

The following section reports the overall prevalence of single or multi-class resistance within the population of people with TDR. In the 77 TDR cases, resistance was confined to a single drug class (NNRTI, NRTI or PI) for 63 sequences (82%) and 50 of these (79%) had a single mutation. Almost half the TDR cases harbored resistance only to NNRTIs

(44%, 34), 29% (23) only to NRTIs, and 8% (6) only to PIs. Single class resistance with multiple NNRTI or NRTI mutations occurred in four and eight cases respectively. All cases with PI resistance carried single mutations.

Dual-class resistance to NRTIs/NNRTIs was found in eight (1.6%) of the 496 TN cases (10% of TDR cases) and dual-class resistance to both NNRTIs/PIs, and to NRTIs/PIs occurred in two cases respectively. Triple class resistance also occurred in two cases. Table 15 and 17 shows TDRMs observed in the 77 cases.

Table 15. Genotypic TDR mutations detected and predicted phenotypic resistance in HIV-1 cases diagnosed in South Australia 2000–2013

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
								NNRTI	NRTI	PI	NNRTI	NRTI	PI
1015	2000	M	50	U	B	Australia	N/A		D67N K219Q T69D T215L			AZT(L), d4T(L), ABC(U), TDF(U), ddI(U)	
1046	2000	M	25	MSM	B	Australia	N/A		D67N T69N K70R T215F K219Q	L90M		AZT(I/H), d4T(I/H), ABC(L), TDF(L), ddI(L)	SQV(R), FPV(R), IDV(R), LPV(R), NFV(R), ATV(R), TPV(R)
1051	2000	M	36	MSM	B	Australia	N/A	Y181C Y188L			NVP(H), ETR(I/H), RPV(I/H), EFV(H)		
1057	2000	M	25	MSM/IDU	B	Australia	N/A		M41L T215E			M41L and T215Y confer AZT(H), d4T(H), ddI(I), ABC(I), TDF(I). T215E is revertant mutant of T215F/Y.	
1072	2001	F	25	Het	C	Overseas	SSA	Y181C			NVP(H), ETR(I/H), RPV(L), EFV(L)		
1087	2001	M	38	MSM	B	Australia	Australia	K103N	D67N T215E K219Q T215L	L90M	NVP(H), EFV(H)	AZT(L), d4T(L), ABC(U), TDF(U), ddI(U)	SQV(R), FPV(R), IDV(R), LPV(R), NFV(R), ATV(R)
1100	2001	M	26	Het/IDU	01_AE	Overseas	Asia	G190E			NVP(H), RPV(H), ETR(H), EFV(I)		

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
1107	2001	M	44	MSM	B	Australia	N/A		M41L T215D			M41L and T215Y confer AZT(H), d4T(H), ddI(I), ABC(I), TDF(I). T215D is revertant mutant of T215F/Y.	
1117	2001	M	19	MSM	B	Australia	N/A			I47V			DRV(R), FPV(R), IDV(R), LPV(R), NFV(R), TPV(R)
1126	2001	M	46	MSM	B	Australia	Australia	G190S	M41L D67N V75M T215S		NVP(H), EFV(H)	d4T(I), ddI(L), AZT(L), ABC(U), TDF(U)	
1132	2001	M	52	MSM	B	Australia	N/A	K103N			NVP(H), EFV(H)		
1139	2002	M	49	MSM	B	Australia	N/A	G190E		V82A	NVP(H), RPV(H), ETR(H), EFV(I)		Reduces susc. to IDV, LPV, ATV, NFV
1159	2002	M	24	MSM	B	Overseas	N/A	V179D V106A	M184V		NVP(H), EFV(I), ETR(L), RPV(L)	3TC(H), FTC(H), ddI(L), ABC(L)	
1164	2002	M	61	MSM	B	Overseas	N/A		D67N K70R M184V	M46I I84V		3TC(H), FTC(H), ddI(L), ABC(L), AZT(L), d4T(L), TDF(L)	DV(H), NFV(H), FPV(H), SQV(H), ATV(H) LPV(I), TPV(I), DRV(L)
1166	2002	M	53	MSM	B	Australia	N/A		M41L			Unknown on own	
1169	2002	M	35	MSM	B	Australia	N/A		M41L			Unknown on own	
1175	2002	F	26	Het	B	Australia	Australia		D67N T69D K219Q T215L			AZT(L), d4T(L), ABC(U), TDF(U), ddl(U)	
1178	2002	M	34	Het/IDU	B	Australia	N/A		D67N T69D K219Q			AZT(L), d4T(L), ABC(U), TDF(U), ddl(U)	

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
									T215L				
1186	2003	M	26	MSM	B	Australia	Asia	K103N			NVP(H), EFV(H)		
1187	2003	M	39	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1188	2003	M	32	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1192	2003	M	46	MSM	B	Australia	Australia		T215S			Revertant Mutant of T215Y/F.	
1198	2003	M	45	MSM	B	Australia	N/A		M41L			Unknown on own	
1201	2003	M	31	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1203	2003	M	32	MSM	B	Australia	Australia	K103N		N88S	NVP(H), EFV(H)		NFV(H), ATV(H), IDV(L), SQV(L), FPV(Increased susc)
1207	2003	M	38	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1209	2003	M	55	MSM	B	Australia	Australia	G190A Y181C	K70R T215I T215F T215S K219Q		NVP(H), ETR(I/H), RPV(L), EFV(I)	AZT(I), d4T(L), TDF(L), ABC(L), ddI(L)	
1212	2003	M	50	MSM	B	Australia	Europe	K103N	D67N T69D	L90M	NVP(H), EFV(H)	AZT(L), d4T(L), ddl(U)	SQV(R), FPV(R), IDV(R), LPV(R), NFV(R), ATV(R)
1213	2003	M	38	MSM	B	Australia	Australia			M46L			Reduces susc. to IDV, NFV, FPV, LPV, TPV when other mutations present
1214	2003	M	32	MSM/IDU	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1215	2003	M	40	MSM	B	Australia	Australia		T69D			ddI(U), d4T(U)	

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
1216	2003	M	23	MSM	B	Australia	N/A	K103N			NVP(H), EFV(H)		
1226	2003	M	44	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1245	2004	M	27	MSM	B	Australia	Asia	K103N			NVP(H), EFV(H)		
1246	2004	M	53	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1247	2004	M	46	Het	C	Overseas	Europe		L74V			ddI(H), ABC(I), AZT(U), TDF(U)	
1249	2004	M	58	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1253	2004	M	22	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1267	2004	M	43	Het/IDU	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1270	2004	M	48	MSM	B	Overseas	Australia			V82A			Reduces susc. to IDV, LPV, ATV, NFV
1274	2004	M	39	MSM	B	Australia	Australia		M41L			Unknown on own	
1275	2004	M	57	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1276	2004	M	43	MSM	B	Australia	Australia		M41L D67N T215C K219Q			AZT(L), d4T(L), ABC(U), TDF(U), ddI(U)	
1307	2005	M	31	MSM	B	Australia	Australia		T69D			Reduces susceptibility to ddI	
1324	2005	M	52	MSM	B	Australia	Australia	K103N V179T			NVP(H), EFV(H)		
1338	2005	M	28	MSM	B	Australia	Australia	K103S			NVP(H/I), EFV(L/I)		
1374	2005	M	36	Het	B	Australia	Australia	K103N K103S K101P	M41L T69N		NVP(H), EFV(H), RPV(H), ETR(I)	Unknown	
1382	2005	M	46	Het	B	Australia	Australia	K103N			NVP(H), EFV(H)		

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
1412	2006	F	49	Het	01_AE	Australia	Australia			M46L			Reduces susc. to IDV, NFV, FPV, LPV, TPV when other mutations present
1432	2006	M	13	U	02_AG	Overseas	SSA	Y181C G190A	M184V		NVP(H), ETR(I), RPV(I), and EFV(I)	3TC(H), FTC(H), ddI(L), ABC(L)	
1446	2006	M	37	MSM	B	Australia	Australia		K219E			AZT(U), d4T(U)	
1462	2006	M	35	MSM/IDU	B	Australia	Australia		D67N K219Q T69D T215L			AZT(L), d4T(L), ABC(U), TDF(U), ddI(U)	
1466	2006	M	46	U	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1476	2007	M	33	Het/IDU	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1500	2007	F	48	Het/IDU	B	Australia	Australia	Y181C	K219E L74I		NVP(H), ETR(I/H), RPV(L), EFV(L)	AZT(U), d4T(U)	
1509	2007	M	69	MSM	B	Australia	Europe	K103N V106I			NVP(H), EFV(H)		
1522	2007	M	47	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1537	2007	F	36	Het	01_AE	Australia	Australia			L76F			L76V reduces susc. to IDV, LPV, DRV and FPV. Increases susc. to ATV, SQV and TPV.
1538	2007	M	48	Het	A	Overseas	SSA		K70E			(L/I) to TDF, ABC, DDI & possibly 3TC and FTC. Increases susc. to AZT.	
1568	2008	M	38	H/IDU	B	Australia	Australia		T69D L74V			ddI(H), ABC(I), AZT(U), TDF(U),	

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
												d4T(U)	
1569	2008	M	71	MSM	B	Australia	Europe	K103N			NVP(H), EFV(H)		
1570	2008	F	31	Het	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1577	2008	M	35	Het	B	Overseas	Australia	K103N	K219E		NVP(H), EFV(H)	AZT(U), d4T(U)	
1585	2008	M	43	MSM	B	Australia	Australia		T215S			Revertant mutation emerging from T215Y in the absence of NRTIs	
1590	2008	M	3	MTCT	A/AE	Australia	Australia	K103N M230L	M184V		NVP(H), EFV(H), RPV(I/H), ETR(I/H)	3TC(H), FTC(H), ddI(L), ABC(L)	
1591	2008	M	36	MSM/IDU	B	Australia	Australia	K103N			NVP(H), EFV(H)		Unknown
1596	2008	M	38	MSM	B	Overseas	America	K103N			NVP(H), EFV(H)		
1652	2009	M	48	MSM/IDU	B	Australia	Australia		M184V			3TC(H), FTC(H), ddI(L), ABC(L)	
1655	2009	M	56	MSM	B	Australia	Australia		T69D			ddI(U), d4T(U)	
1656	2009	M	39	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1678	2009	M	51	Blood contact overseas	A/AE	Overseas	SSA	Y181C G190A V106I			NVP(H), ETR(I/H), RPV(L), EFV(I)		
1708	2010	M	37	MSM	B	Australia	Australia			L90M			SQV(R), FPV(R), IDV(R), LPV(R), NFV(R), ATV(R)
1728	2010	M	40	MSM/IDU	B	Australia	Australia	K103N			NVP(H), EFV(H)		
1734	2010	M	40	MSM	B	Australia	Australia		D67N			AZT(L), d4T(L), ddI(U)	
1785	2011	M	37	MSM	B	Australia	Australia	K103N			NVP(H), EFV(H)		

ID	Yr Dx	Sex	Age	Risk	Subtype	Location acquired	Region born	Resistance mutations			Predicted phenotypic resistance (level of resistance is classified by mutation with highest resistance)		
1868	2012	M	39	MSM/IDU	B	Australia	Australia		M184V			3TC(H), FTC(H), ddI(L), ABC(L)	
2030	2013	M	63	Het	B	Australia	Australia	K103N			NVP(H), EFV(H)		

Key: NNRTIs (non-nucleoside reverse transcriptase inhibitors); NRTIs (nucleoside reverse transcriptase inhibitors); PIs (protease inhibitors); M (Male); F (Female); Het (Heterosexual); Het/IDU (Heterosexual sex with IDU risk); MSM (Men who have Sex with Men); MSM/IDU (MSM with IDU risk); U (Unknown risk); SSA (Sub-Saharan Africa); H (high level resistance); I (Intermediate level resistance); L (Low level resistance); R (Reduces susceptibility to inhibitors); I (Increases susceptibility to inhibitors); NVP (nevirapine); EFV (efavirenz); ETR (etravirine); RPV (Rilpivirine); 3TC (lamivudine); FTC (emtricitabine); ABC (abacavir); ddI (didanosine); TDF (tenofovir); d4T (stavudine); ZDV (zidovudine); ATV (atazanavir); DRV (darunavir); FPV (fosamprenavir); IDV (indinavir); LPV (lopinavir); NFV (nelfinavir); SQV (saquinavir); TPV (tipranavir).

Table 16. NNRTI, NRTI and PI resistance by subtype

Genotype	T/N cases N	T/N subtype with TDR N (%)	Multi-class resistance N	NNRTI resistance N	NRTI resistance N	PI resistance N
B	365	68 (19)	12	41	31	10
C	41	2 (5)		1	1	
AE	51	3 (6)		1		2
AG	16	1 (6)	1	1	1	
D	2	0 (0)				
A	4	1 (25)			1	
A/AE	5	2 (40)	1	1	1	
A/B	1					
AE/AG	1					
AE/B	1					
AG/B	4					
B/C	1					
B/D	1					
B/G	1					
G	2					
Total	**496**	**77**	**14**	**45**	**35**	**12**

5.7 Amino acid substitutions by drug class

Amino acid substitutions leading to drug resistance can differ between subtypes and CRFs. The following sections reports substitutions by each of the three drug classes, and analyses demographic links between mutations and the population.

5.7.1 NNRTI Mutations

Thirty-seven TN cases carried one surveillance mutation to NNRTIs, eight cases carried two and one case carried three (Tables 15 and 17). The most prevalent sites for NNRTI mutations were K103 (36 cases), followed Y181 (6 cases) and G190 (6 cases).

5.7.1.1 Mutations at position K103

Nearly half of all cases harboring resistance (36/77) carried mutations at position K103; 34 with the N mutation, one with the S mutation and one with both (hereafter referred to as K103N and K103S respectively). K103N reduces susceptibility to NVP (~50-fold) and EFV (~20-fold) while K103S causes intermediate to high-level resistance to NVP, and low to intermediate resistance to EFV.

The majority of K103 cases were diagnosed between 2003 and 2004 (44% or 16 of all K103 cases) and another spike occurred between 2007 and 2008 (25% or 9 of all K103 cases), Figure 18.

Thirty-five (97%) of the 36 K103 cases were of a B subtype, 35 were males (the one female had subtype B), 33 were people who acquired the infection within Australia and 28 were Australian-born. The two subtype B cases acquired overseas were males in their 30s, one American-born who acquired HIV via MSM transmission, the other Australian-born via heterosexual contact. Both cases formed part of the 2007–2008 spike (Figure 18). The one non-B case was a *pol*-ISR A/AE variant, carried by a male infant who was infected via MTCT in Australia.

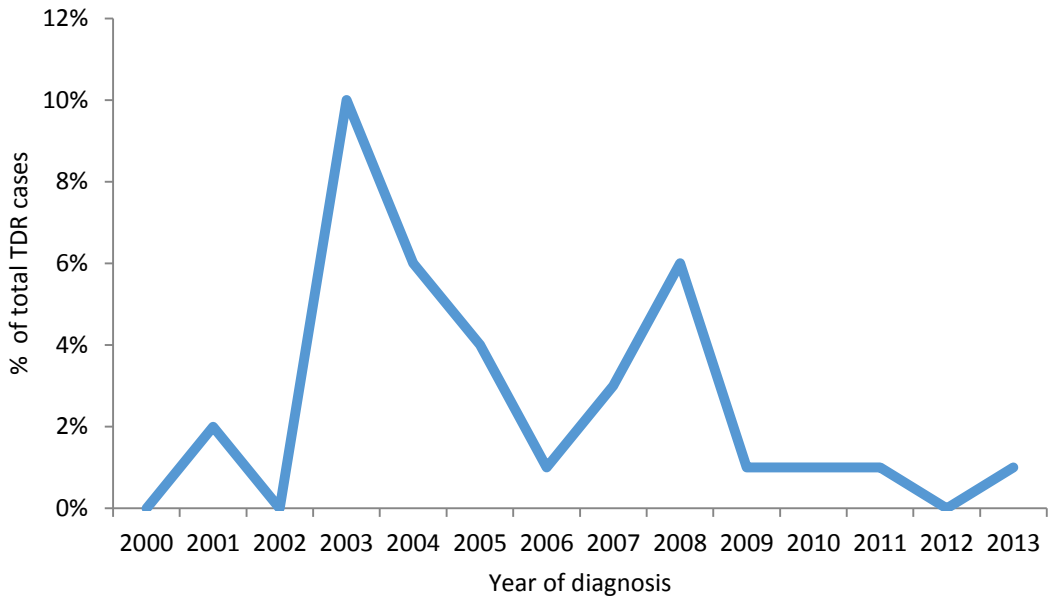


Figure 18. Proportion of cases with mutations at position K103 from total TDR cases (n=78) over the entire period.

Just over two-thirds of cases with a mutation at K103 were via MSM transmission (67%, 24/36), and all 24 cases were a B subtype. In 23 cases the person acquired the infection in Australia and in 16 cases the person was born in Australia. The overall prevalence of K103 mutations in the TN MSM cohort was 8% (25/302).

The remaining twelve K103 cases were reported as being transmitted as follows: three through MSM with IDU risk, two through heterosexual contact with IDU risk, five through heterosexual contact, one through MTCT and one unknown. All twelve cases were people born in Australia, and all but one acquired the infection in Australia.

5.7.1.2 Mutations at positions Y181, Y188 and G190

Mutations at positions Y181, Y188 and G190 were identified in nine cases. Three cases carried both the G190A and Y181 mutations, two cases carried a single Y181C mutation, three cases carried a single E or S mutation at G190, and one case carried the Y181C and Y188L mutations.

Together, mutations at positions Y181 and G190 reduce susceptibility to NVP (≥50-fold), ETR (≥10-fold) and RPV (3-fold). Y181C reduces susceptibility to EFV 2-fold, G190A reduces it 5- to 10-fold, and G190ES reduces it ≥50-fold. G190E also confers high-level resistance (≥10-fold) to RPV and ETR. The L mutation at Y188 reduces susceptibility to NVP and EFV ≥50-fold, and RPV 5-fold.

Just over half (5/9) of the cases were subtype B, all diagnosed in Australia, and three of them carried various NRTI accessory and TAM mutations (K219EQ, T215IS, K70R, D67N, M41L). One case also carried the F mutation at position T215, which usually occurs with TAM mutations. Of the four non-B cases, all were acquired overseas and were people born overseas.

Dual resistance occurred in five of the nine cases, four cases carried dual NRTI resistance and one carried dual PI resistance, Table 15 and 17. Of note is one case in which the person was born and infected in Africa and carried mutations Y181C, G190A, H221Y (not shown in data) and the NRTI mutation M184V. H221Y often occurs in combination with Y181C and is commonly selected in people receiving RPV. M184V causes high-level resistance to 3TC and FTC, and low-level resistance to ddI and ABC.

5.7.1.3 Mutations at positions K101, V106, and M230

A single P mutation at position K101 occurred in the same case that harbored mutations N and S at position K103 (ID number 1374). K101P and K103N both cause high-level resistance to NVP and EFV, and K101P also causes high-level resistance to RPV and intermediate/high resistance to ETR.

Another case that was also carrying the N mutation at K103 also harbored resistance to the L mutation at position M230 (ID number 1590). This is an uncommon mutation usually selected in people receiving EFV, NVP and RPV combination therapy. It causes intermediate to high-level resistance to all NNRTIs. This case was a child diagnosed at 3

years of age, with a *pol*-ISR A/AE variant that had been transmitted through MTCT. The child also carried the NRTI V mutation at position M184.

Only one case carried the A mutation at position V106 that causes high-level resistance to NVP and EFV. This was a male diagnosed in 2002 who acquired the infection overseas and whose country of birth was listed as unknown.

5.7.2 NRTI Mutations

NRTI resistance occurred in 35 (7%) TN cases, and 49% of these (17) had two or more mutations. Eight cases carried dual resistance to NRTIs and NNRTIs, two cases carried dual resistance to NRTIs and PIs, and two cases carried multi resistance to all the drug classes.

The most prevalent sites for NRTI surveillance mutations were D67 (11 cases), T69 (11), T215 (10), and K219 (11), followed by M41 (9), M184 (6), K70 (4), L74 (3) and V75 (1).

5.7.2.1 Mutations at positions D67, T69, M184 and K219

Six of the 35 NRTI resistant cases had mutations at positions D67 (N), T69 (D), and K219 (Q), all of which confer resistance against AZT, d4T, and ddI. All cases were a B subtype, and all but one were male cases. In total, 11/77 (14%) cases carried a K219 mutation, eight carried mutation Q and three carried mutation E, 11 (14%) cases carried a D67N mutation and 11 cases carried a T69D mutation.

Six cases carried mutation V at position M184. Three of these carried a single NRTI mutation while the other three had dual resistance with multiple NNRTI mutations. M184V causes high-level resistance to 3TC and FTC and low-level resistance to ddI and ABC. However, this mutation also increases susceptibility to AZT, TDF and d4T and is associated with clinically significant reductions in HIV-1 replication.

5.7.2.2 Mutations at positions M41 and T215

Nine cases carried mutation L at position M41 with seven carrying single class resistance including four who harbored a single mutation only. The remaining two cases had dual resistance to NRTIs and NNRTIs. Eight of the nine cases were reported as MSM transmission, all acquired in Australia, all subtype B, and diagnosed between 2000 and 2005.

Four cases that carried M41L also carried a T215 revertant, which emerges from T215Y or F in the absence of NRTI treatment. The presence of these revertant mutations suggests that a) the person may once have harbored a T215T/F mutation as a result of treatment, b) the person may once have harbored a T215T/F mutation transmitted to them during infection which then reverted in the absence of NRTI treatment, or c) the revertant mutation was transmitted to them during infection from someone who was either on treatment or had been on treatment. In the presence of treatment is it possible that the T215 revertants could mutate to the Y mutation. Together, M41L and T215Y confer high-level resistance to AZT and d4T, and intermediate-level resistance to ddI, ABC and TDF.

One case that carried M41L and a T215 revertant also carried the N mutation at position D67 and the Q mutation at position K219. All four of these mutations are TAMS that, when combined, reduce susceptibility to AZT, d4T, ABC, TDF and ddI. TAMS are selected by AZT and d4T and facilitate primer un-blocking.

Another case carried all the same mutations as the one above except at K219, with an additional mutation (M) at V75 which reduces susceptibility to d4T and ddI, and the NNRTI mutation S at position G190 which causes more than 50-fold reduced susceptibility to NVP and EFV.

Another M41L case also carried mutation T at position T69, as well as three NNRTI mutations, N and S at K103, and P at K101. This Australian-born person was the only one

of the M41 cases that reported heterosexual transmission acquired in Australia. The combined mutations confer extremely high level resistance to a number of NNRTIs and NRTIs.

5.7.3 PI mutations

In the protease gene a high frequency of various naturally occurring polymorphisms and non-surveillance mutations was recorded, but only seven different surveillance mutations, the most common of which were L90M (4 cases) and M46L (3 cases).

Twelve cases carried resistance to PIs. Four carried the L90M mutation, which confers resistance to SQV, FPV, IDV, LPV, NFV, and ATV. Three of these cases also had NRTI mutations that confer low to intermediate resistance to AZT, d4T, ABC, TDF and ddI and two of these three cases carried triple class resistance, also harboring the NNRTI N mutation at position K103, which confers high-level resistance to NVP (50-fold) and EFV (20-fold). All four L90M cases were MSM with subtype B acquired in Australia.

Two cases carried the L mutation at position M41 and one case carried the I mutation, both of which reduce susceptibility to IDV, NFV, FPV, LPV and ATV when present with other mutations. M46L also reduces susceptibility to TPV. Only one of the three cases carried other mutations, the PI mutation I84V that causes high resistance to ATV, FPV, IDV, NFV and SQV, intermediate resistance to LPV and TPV, and low resistance to DRV, and three NRTI mutations D67N, K70R and M184V, which cause high resistance to 3TC and FTC, intermediate resistance to AZT and low resistance to d4T, ddI, ABC and TDF. The first diagnosis of a case with the M46 mutation was in 2002, through MSM transmission acquired overseas. The next case was diagnosed in 2003 also through MSM transmission but acquired in Australia. The last case was diagnosed in 2006, a female who reported heterosexual acquisition in Australia.

One PI resistant case carried N88S, which causes high-level resistance to NFV and

ATV, and low-level resistance to IDV and SQV, a B subtype transmitted by MSM contact within Australia. One case carried I47V which reduces susceptibility to all PIs except SQV and ATV, another B subtype transmitted through MSM within Australia.

5.8 Demographic analysis of treatment naïve persons with TDRMs

This section reports demographic characteristics associated with the transmission of drug resistant HIV strains, in order to elucidate possible prevention strategies.

5.8.1 Transmission Risk

Of the total TN cohort, 86% (427/496) were male. MSM was reported in 71% (301/427) of male cases, including 25 MSM who also reported IDU risk.

Eighteen percent (55/301) of TN MSM carried at least one mutation, and of the MSM with IDU risk subset group, 28% carried at least one mutation. The proportion of TN MSM carrying TDRMs has decreased over time, from 36% (36/99; 2000–2004), to 11% (14/127; 2005–2009) to 7% (5/75; 2010–2013, $p \leq 0.0001$). Age at diagnosis ranged from 19 to 71 years.

Of the 187 males who reported transmission risk other than MSM or unknown, 10% (19) carried mutations. Seventeen were reported as heterosexual transmission (including six with IDU risk) and of these heterosexual transmission cases, five were acquired overseas and twelve in Australia. Overall TDR prevalence among TN heterosexual males (not including IDU risk) was 9% (11/125). TDR prevalence among the 31 heterosexual males with IDU risk was 19% (6). Age at diagnosis ranged from 26 to 72 years.

Table 17. TDR distribution in 77 cases

Cases with a TDR at position	N	Proportion of TDR cases (%) (N=77)	Proportion of TN cases (%) (N=496)
Any	77	100%	15.5%
Multiple Drug Classes	14	18%	2.8%
NNRTI			
At least 1	46	60%	9.3%
Two or more	9	12%	2.0%
K101	1	1%	0.2%
K103	36	47%	7.3%
V106	3	4%	0.6%
Y181	6	8%	1.2%
Y188L	1	1%	0.2%
G190	6	8%	1.2%
M230	1	1%	0.2%
V179	2	2%	0.4%
NRTI			
At least 1	35	45%	7.1%
Two or more	18	23%	3.6%
M41	8	10%	1.6%
D67	12	15%	2.4%
T69	11	14%	2.2%
K70	4	5%	0.8%
L74	3	4%	0.6%
V75	1	1%	0.2%
M184	6	8%	1.2%
L210	1	1%	0.2%
T215	11	14%	2.2%
K219	11	14%	2.2%
PI			
At least 1	12	15%	2.4%
Two or more	1	1%	0.2%
M46	3	4%	0.6%
I 47	1	1%	0.2%
L76	1	1%	0.2%
V82	2	3%	0.4%
I84	1	1%	0.2%
N88	1	1%	0.2%
L90	4	5%	0.8%

Key: TDR (Transmitted Drug Resistance), TN (Treatment Naïve).

Two male cases were not sexually transmitted, a MTCT acquired in Australia and a blood exposure/medical procedure overseas.

Female TDR prevalence was 9% (6/69) and all acquired their infection through heterosexual contact, including one female who reported dual IDU risk. Five cases were acquired in Australia and they were females born in Australia. One female acquired the infection overseas and was born overseas (Africa).

5.8.2 Region of birth

TDR prevalence for people who were Australian-born was 17% (51/298). All but three were subtype B cases, and of these subtype B cases all but two were acquired in Australia.

Of the three non-B cases that were Australian-born, all were acquired in Australia, and two were female 01_AE cases through heterosexual transmission. The male was a child with an A/AE variant transmitted through MTCT.

TDR prevalence for overseas-born cases was 9% (12/140). Of these 12 overseas-born cases, one was American-born (16.5% of total TN American-born people), three were Asian-born (9%), four were European-born (11%) and four were African-born (7.5%). Half (6/12) of the overseas-born cases had subtype B virus, all reported MSM transmission and all but one was acquired in Australia. One American-, two Asian- and three European-born males accounted for these cases.

Non-B cases accounted for the other 46% (6). These comprised one 01_AE, two C, one 02_AG, one A, and one *pol*-ISR (A/AE variant). These six non-B cases were all people born overseas who acquired the virus overseas, only one was female (heterosexual

exposure). Of the five non-B male cases, three reported heterosexual transmission, one a blood exposure and one unknown exposure.

Nearly one-fifth (18%, 14/77) of cases carrying mutations did not have a country of birth listed. They were males with subtype B diagnosed between 2000 and 2003, all but two were acquired in Australia and all but two were MSM transmission.

5.8.3 Location infection acquired

TDR prevalence for cases acquired in Australia was 18% (66/350), and 95% of these cases (63) had a B subtype. Two cases were females with 01_AE and were diagnosed between 2006 and 2007. The last case was a male child with a *pol*-ISR A/AE variant, MTC transmission and the only non-B TN case to have a K103N mutation.

Just over three-quarters (77%, 51/66) of cases acquired in Australia were MSM (69% born Australia, 11% overseas, 20% unknown). Twelve cases were heterosexual transmission (including five females), one was MTCT and two were unknown.

5.8.4 Sex and age

TDR prevalence in TN females was 9% (6/69) and 16.5% (71/429) in TN males. The ratio of males to females carrying mutations was 12:1.

The majority of male cases were subtype B (92%, 65), nearly all acquired in Australia (92%, 60). Four of the 72 males (6%) were non-B cases, all born and acquired overseas. Two males were *pol*-ISR cases, both A/AE variants diagnosed between 2008 and 2009 and with one born in Australia (MTCT).

Five of the six females were Australian-born, and all acquired their infection in Australia (three subtype B cases and two 01_AE cases). One carried single NNRTI resistance, two carried single PI resistance, one carried single NRTI resistance and one carried dual NNRTI/NRTI resistance. The other female was African-born and infected

overseas with a subtype C virus that carried single NNRTI resistance.

Ten percent (8/77) of TDR carriers were under 25 years of age. One carried *pol*-ISR A/AE and one carried CRF02_AG. All cases were male, and six reported MSM exposure. The other two males were aged three at diagnosis (MTCT acquired in Australia) and aged 13 (African male with unlisted risk and infection acquired overseas).

Seven of these young people carried virus with high-level resistance to at least one NNRTI or NRTI. Two young MSM also carried PI mutations that conferred resistance to all PI drugs except TPV and DRV for one person, and SQV and ATV for the other.

Seventeen percent (13/77) of TDR carriers were aged 51 and over. All were male and all but one had subtype B. One African-born male had an A/AE ISR, contracted through direct blood exposure overseas. This male had high level resistance to two NNRTIs and intermediate/low resistance to two others.

Of the 12 subtype B males, 11 reported MSM exposure and seven of these carried K103N, which confers high level resistance to two NNRTIs.

One male reported heterosexual exposure and had a subtype B virus that harbored the K103N mutation only.

5.8.5 Year of diagnosis

The majority of TN cases carrying TDRMs were diagnosed between 2002 and 2009 (Figure 19), with the annual proportion of new diagnoses carrying TDRMs decreasing over time (Figure 20).

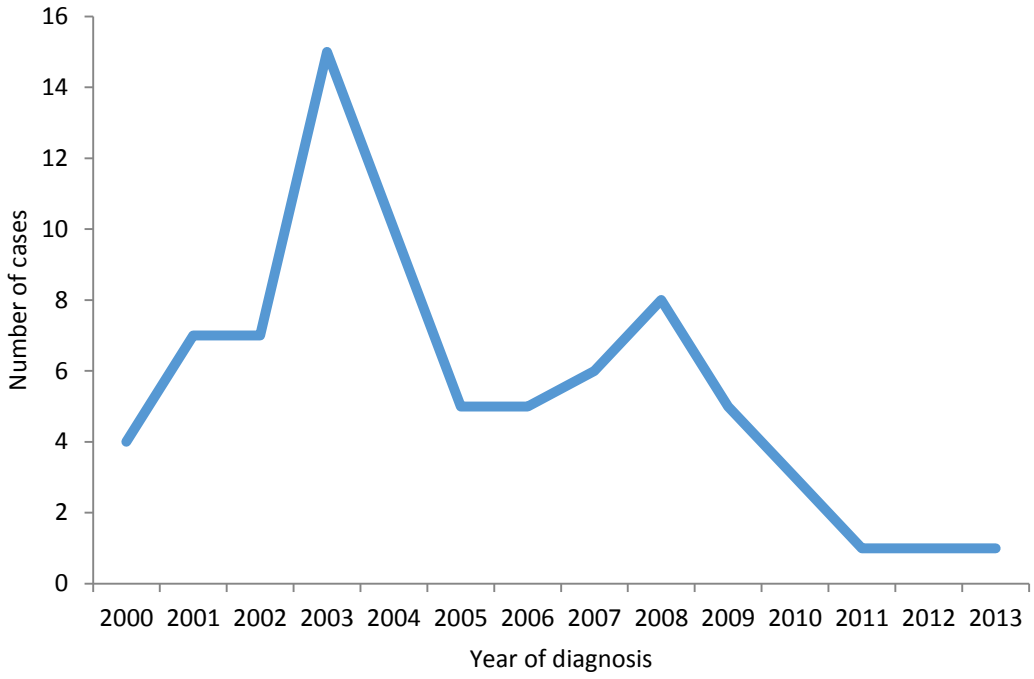


Figure 19. Number of TDR cases annually, 2000–2013.

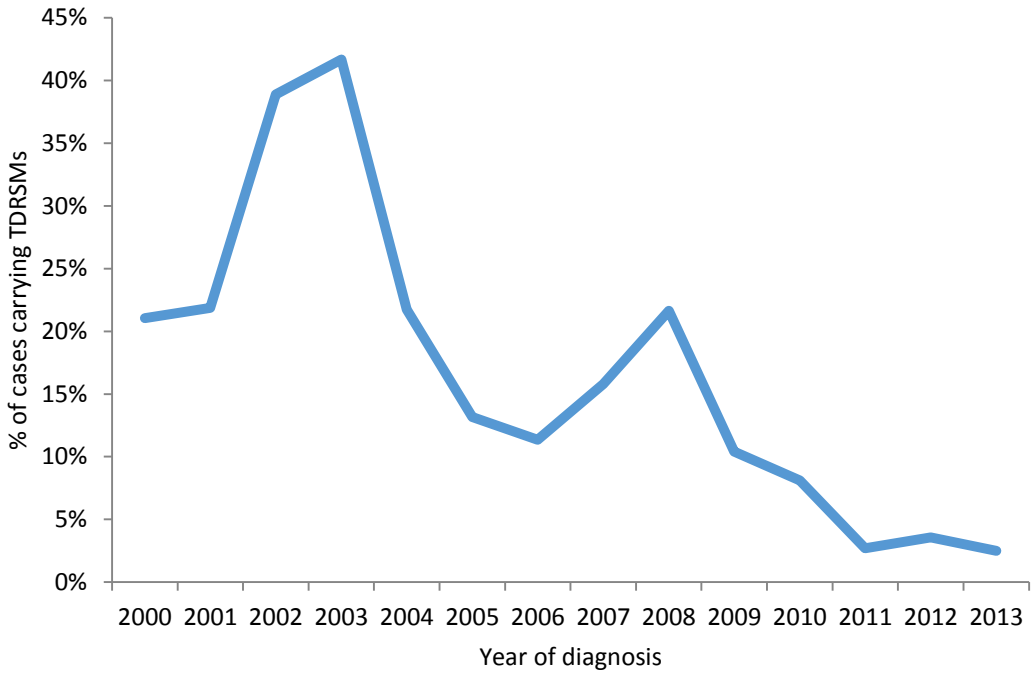


Figure 20. Proportion of annual treatment naïve cases carrying TDRMs.

5.9 Summary

The prevalence of TDR in the treatment naïve cohort was high at 15.5%, with rates significantly decreasing over time. The predominant TDR mutation was K103N, most likely because of the use of NVP as a mono or dual therapy and selection pressure during the early HAART era. TDR was highest in the MSM population and the majority of MSM cases were males born and infected within Australia.

The prevalence of TDR in non-B subtypes was moderate at 7%, and has also significantly decreased over time. The moderate levels reflect the emergence of drug resistance mutations when HAART was first introduced in LMICs and were caused by the use of drugs with a lower genetic threshold for resistance, guideline-based treatment rather than individualized care, issues with treatment adherence, the extended period between infection and diagnosis, and absence of regular viral load testing for rapid changeover from first to second line treatment in the case of treatment failure.

The decrease in TDR in both subtype B and non-B cases over time corresponds with more robust treatment regimens globally, improvements in HIV testing and early diagnosis, regular viral load checks and improved care of people diagnosed with HIV.

5.10 Discussion

Almost 30 years after the introduction of antiretroviral therapy in Australia it is timely to analyze the prevalence of TDR among treatment naïve people newly diagnosed with HIV in South Australia. Approximately 15 million people are now receiving ART worldwide, and in 2004 WHO began a global surveillance program to measure acquired and TDR as ART access was expanded to those in need. Surveillance is crucial to gauge how drug resistance may affect future treatment options, including the best commencement regimen and which treatments should be excluded, along with careful monitoring of HIV viral load and regular genotypic resistance tests to assess how a particular subtype is reacting to various

treatments.[228]

Currently, WHO defines the prevalence of TDR within a geographic area as follows: ≤5% (low), 5–15% (moderate) and ≥15% (high).[128] The present study demonstrates a high but decreasing rate of TDR in newly diagnosed individuals in South Australia between 2000 and 2013. This high TDR prevalence rate is not surprising given ART has been widely available since the late 1980s when serial mono and dual therapies were used, which often led to high levels of circulating drug resistance mutations, as did early HAART regimens.[11] The overall high prevalence rate in this study mirrors findings in Victoria[10] and other high-income countries.[18] A systematic review by Pham *et al*. found TDR prevalence was greatest in North America, followed by Western Europe and South America.[18] Studies in Latin America,[127] Germany[115], and San Francisco[229] have all reported relatively high prevalence rates similar to the present study. A Latin American study of approximately 100 people living in the Santo Domingo discovered a similarly high prevalence of the K103N mutation, although the majority of the K103N infected population were heterosexual males,[127] unlike the South Australian population in which the K103N mutation was predominantly found in viruses harbored by MSM. Recent TDR surveillance in Asia and Africa shows relatively low prevalence rates, which was supported by findings in the current study for those born in these regions.[128] However, within specific regions such as Uganda there is growing evidence of a surge in TDR, most likely because of the earlier roll-out of ART than in other regions and rapid expansion of a standardized public health first-line ART regimen using a NNRTI and dual NRTI backbone, rather than individualizing treatment for each person through genotypic resistance testing.[19,230]

The decline in TDR from 2000 to 2013 is almost certainly related to the introduction of NNRTIs and PIs to ART, whereas before 1996 treatment options were limited to the NRTI drugs ZDV, DDI or DDC.[10] This two- or three-drug class combined HAART regimen raised

the genetic threshold to resistance by increasing the antiviral potency, thereby decreasing the prevalence of acquired resistance and hence TDR as a flow-on effect. Improved treatment adherence may have also affected TDR prevalence, because of a higher level of education about resistance, a decrease in severe side effects of the antiretroviral drugs, and the change from multi-pill regimens multiple times a day to triple combination therapy contained in one tablet taken once a day.[231]

5.10.1 Type of resistance

Although current treatment regimens in South Australia use a combination of NNRTI, NRTI, PI and INSTIs,[15] only the first three were examined in this cohort, with INSTIs only introduced into treatment regimens relatively recently. Almost three-quarters of people with TDR carried high-level resistance to at least a single drug class, many with high-level resistance to multiple drugs within that class. This reduces the availability of treatment options for current first-line triple combination regimens for treatment naïve people, with a higher risk for subsequent restrictions on second and third line regimens in the event of treatment failure.

While the prevalence of TDR to all three drug classes decreased significantly over time, single NNRTI mutations were most prevalent followed by multiple NRTI mutations. This reflects the widespread use of these two drug classes, firstly NRTIs as standard serial mono and dual therapies in high income countries, then later worldwide as HAART NNRTI/NRTI combination regimens.[11,19,113,230]

The high NNRTI mutation prevalence was because of a large number of people with the K103N mutation, being 7% of all treatment naïve cases and nearly half of all TDR cases. This mutation is known to develop in people exposed to NVP or EFV as a monotherapy and leads to high level resistance to the entire drug class.[232] High rates of K103N-mutated virus have also been found circulating among untreated HIV carriers in Victoria and it is thought

to be preferentially transmitted.[10] Globally, the K103N mutation is found in approximately 37% of NNRTI treated people with a subtype B virus, but only 1.1% of the TN population.[198] In this study, K103N prevalence among TN people was over six times higher than that reported in the Stanford drug resistance database. All but one K103N cases were subtype B virus and predominantly transmitted through MSM. There were also subsets of K103N-containing viruses transmitted through MSM and heterosexual contact with IDU risk. All these subtype B K103N-containing viruses were contracted by people born in Australia and all but one acquired the infection in Australia. Most were diagnosed between 2003 and 2004 or 2007 and 2008, although there were diagnoses spread across the entire time period. The K103N mutation first rapidly spread through the MSM population, most likely because of multiple concurrent partners newly infected with the TDR strain and quite possibly undiagnosed at the time of transmission. The crossover between the MSM and heterosexual populations in South Australia (including some with IDU risk) suggests one overall sexual network, with forward transmission of K103N most likely through bisexual men and sharing of needles or other drug paraphernalia.

Only one non-B case carried K103N, a *pol*-ISR A/AE variant originating from sub-Saharan Africa transmitted through MTCT. The child was aged 3 years at time of diagnosis, and he had the rare NNRTI M230L mutation and the K103N mutation, effectively eliminating an entire drug class as a treatment option. He also had the NRTI M184V mutation, which is selected by 3TC and FTC. In Africa, the K103N mutation is often caused by the use of NVP during pregnancy, which has been widely implemented since 1999.[232] The single dose NVP protocol used during childbirth is also thought to inhibit HIV strains susceptible to treatment, which leads to a selection pressure of the background drug-resistant strains and compromises use of subsequent standard first-line NNRTI/NRTI regimen.[135,232] However, although the mother was born in Africa, she gave birth in Australia and was not

diagnosed with HIV herself until the child was aged two. The child also carried the M230L mutation which is uncommon globally, but in a recent Ghana study a prevalence rate of over 50% was found among people with predominantly 02_AG or URFs who had been on a standard first-line treatment regimen for six months, along with the K103N mutation.[129] A genotypic drug resistance profile was undertaken for the mother one year before the child was diagnosed that showed no evidence of drug resistance mutations. It is possible that the mother experienced treatment failure during first-line therapy while breastfeeding, although it is more likely she had been previously diagnosed and treated in Africa, and upon ceasing treatment to come to Australia carried a virus that had reverted back to wild-type at time the genotypic resistance testing was done.

The M230L mutation mentioned above was the only NNRTI mutation found in non-B cases only. Though not a surveillance mutation, H221Y was also noted in one African-born person, in conjunction with Y181C and G190A. These three mutations are commonly found together and are indicative of treatment failure of a regimen containing NPV.[233] Some African origin non-B cases carried Y181C and G190A together, and one also carried V106I. When these mutations are combined they confer significant resistance to etravirine, a drug normally used for second-line treatment after virologic failure while on first generation NNRTI treatment such as EFV and/or NVP, both widely used in sub-Saharan Africa.[120,234]

Only 2.4% of treatment naïve people in the South Australian cohort carried PI mutations and all but two of them were MSM with subtype B diagnosed between 2000 and 2004. This low overall prevalence of PI mutations is consistent with findings in Victoria,[10] and it is a consequence of a number of factors such as higher genetic threshold for resistance, increased antiviral potency because of combination treatment versus monotherapy and increased treatment adherence following improvements in the number of pills needed per day, size of pills, etc.[10] There may also be a smaller proportion of people using PIs compared with

NNRTIs and NRTIs because of their propensity for side effects such as gastrointestinal upset, and problematic drug interactions with other medications.[11]

PI resistance in untreated individuals is indicative of infection with a virus that had been exposed to early HAART, shown by the dual or triple class resistance seen in the majority of PI resistant cases. New PI drugs have a higher genetic threshold to drug resistance compared with those used during early HAART regimens, which also supports this decrease in PI mutation prevalence.[235] L90M was the most prevalent mutation among PIs, followed by M46I/L. These mutations are often found among those with PI resistance, and confer resistance to NFV, IDV and LPV, all of which were used in Australia during this time period ≤http://arv.ashm.org.au/arv-guidelines/appendix-b-drug-characteristics-tables/pis-characteristics≥.

People with NRTI resistance and a B subtype virus were more likely to carry multiple NRTI mutations. The majority of these were diagnosed between 2000 and 2004, while those with a non-B subtype only carried single NRTI mutations, diagnosed between 2004 and 2008. Again, this is indicative of pre-HAART serial NRTI mono and dual therapy in Australia where subtype B predominated, compared with the later implementation of HAART where non-B subtypes originate from.

TAMs were the most common NRTI resistance mutations among subtype B viruses, and this is mainly explained by the extensive use of ZDV as mono and dual therapy during the 1990s and as part of first-line HAART regimens.

Two NRTI mutations, K70E and L76F were found in non-B subtypes only. While the K70E mutation is uncommon across all subtypes overall,[236] it was harbored by a person originating from sub-Saharan Africa, where mutation prevalence among those on treatment can be as high as 25%.[236] It is found in people who are receiving regimens that include d4T, TDF and ABC. It reduces susceptibility to these drugs and can cause low-level resistance to

3TC and FTC while increasing AZT susceptibility. The L76V mutation is known to reduce susceptibility 2-6 fold to LPV, DRV, APV, and IDV, but a 7-8 fold increase in susceptibility to ATV and SQV.[237] The effect of the F mutation at position L76 is unknown.

M184V is the most common mutation among people with NRTI resistance in Australia,[10] but rates remained low with an overall prevalence rate of 1.2% in the TDR cohort. Onward transmission of M184V may have been prevented by a number of factors including low viral titre and poor replicative fitness of this mutation in the transmitter, and impaired fitness of the mutation in the treatment naïve person because of the absence of drug selection pressure.[10]

Serial NRTI monotherapy or concurrent dual treatment failure during the pre-HAART and early-HAART era contributed significantly to the emergence of high levels of acquired drug resistance which seems to have led to rapid onward transmission of these drug resistant forms of HIV through the treatment naïve population.[134] The emergence of newer, more potent drugs that are now used in first-line triple combination regimens has led to a decrease in NRTI resistance.

Between 2000 and 2003, the only dual class resistant viruses were found in MSM and were a mixture of PI, NNRTI and NRTI resistance, including two cases with triple class resistance. Between 2004 and 2013 there were only five dual class resistant cases, all to NNRTIs and NRTIs. Three were subtype B heterosexual transmissions (one with IDU risk) and two non-B cases transmitted non-sexually. This disappearance of dual and triple class TDR in the MSM population and introduction of NNRTI and NRTI dual class TDR through other modes of transmission reflect the early introduction of successful triple combination therapy regimens in Australia where MSM were the predominant HIV-infected population. The changes also reflect the increasing proportion of HIV imported from regions of high infection prevalence where the use of first-line NNRTI/NRTI HAART is relatively new and

issues such as treatment adherence, limited treatment options, no regular viral load checks and poor quality patient care all increase the risk of TDR. This finding emphasizes the importance of conducting genotypic drug resistance profiles before treatment begins both for cases acquired overseas and within Australia, ensuring people are adhering to therapy and having regular viral load and CD4+ T cell testing to monitor the effectiveness of treatment.

5.10.2 TDR and subtype

The treatment naïve cohort for this study was comprised of around 75% subtype B cases and 10% 01_AE cases, with the remaining 15% either subtype A, C, D, G and 02_AG, or *pol*-ISR cases. The latter were predominantly variants containing subtype B, 01_AE and 02_AG.

TDR prevalence was significantly higher for people with subtype B virus compared with non-B/*pol*-ISR virus because of the high number of Australian-born MSM with K103N mutations. Given that subtype is highly correlated with geographic region, either as place of birth or place of infection acquisition, it is likely that the difference between B and non-B subtypes was related more to the amount of time ART has been available in Australia where subtype B among MSM is most prevalent, compared with low and middle income countries where most of the non-B infections originated. However, overall mutations conferring resistance to all three drug classes have decreased significantly for both B and non-B subtype viruses, which supports evidence of the global introduction and increasing access of more virally potent triple combination therapies.

Of the people with dual and triple class resistance mentioned earlier, the majority had a subtype B virus, and most were Australian-born males reporting MSM contact within Australia. This supports the notion that location of HIV acquisition and location of treatment where early ART regimens are first available are better indicators of TDR prevalence than

subtype.

Although the number of non-B/*pol*-ISR cases with TDR was small and decreased over time, the overall TDR prevalence of 8% was moderate according to WHO standards,[113] and it has been well documented that HIV-1 recombination affects viral evolution and the development of drug resistance mutations.[238] Nearly half the cases were people born in sub-Saharan Africa where the rate of viral recombination is high, and carried NNRTI and/or NRTI mutations, most commonly Y181C and G190A. Both of these result in high-level resistance to multiple NNRTI treatments and are associated with use of EFV and NVP in Africa as part of first-line treatment.[135] The major drug resistance mutations found in the African-born cohort were similar to those found in a previous study of African patients, which also found TDR was predominantly single-class resistance as found in our cohort.[19] This moderate prevalence can be explained by the initial implementation of ART in Africa and other LMICs, where treatment adherence was poor because of treatment cost and availability, and treatment guidelines directed the use of dual rather than triple combination therapy.[135] This was clear in the current study. A 13-year-old African-born male carried dual resistance mutations to multiple NNRTI and NRTI treatments, having Y181C, G190A, H221Y and M184V, the last of which confers high-level resistance to 3TC and FTC and low level resistance to ABC and ddI. It is a matter of concern that someone so young is already resistant to at least four NNRTI and four NRTI treatments, given these two drug classes are the preferred first-line treatment regimens. Although triple combination therapy is now widely used in LMICs, it is crucial that accurate and extensive health histories are collected and routine HIV testing conducted on arrival in Australia, so genotypic resistance testing can be performed as early as possible to ensure timely and accurate treatment.

5.10.3 Location of infection acquisition and region of birth

Although the TDR burden has decreased significantly for people infected and/or born in

Australia, their prevalence of TDR was over double that of those infected and/or born overseas. Again, this was mainly because of the rapid transmission of K103N in the Australian MSM population and the prevalence of multiple NRTI mutations from early pre-HAART therapy.

High-income countries have a high level of ART coverage (≥75%) for those who are eligible for treatment,[239] while in 2012 approximately 34% of eligible people in LMICs had access to ART.[240] The higher TDR prevalence in Australian-born and infected people was not surprising given the disparity in ART availability between high-income and LMICs, the difference in length of time ART has been available, and the propensity for multidrug resistance to occur during early treatment regimens because of selection pressure of certain drugs.[216] Other factors that affect TDR include early diagnosis and accurate treatment, regular viral load checks to identify treatment failure early, and monitoring of treatment adherence.

The majority of first-line ART regimens have been introduced into LMICs relatively recently in comparison, and consist of two NRTIs combined with one NNRTI. This has been proven to be extremely effective at sustaining a suppressed viral load and reducing the likelihood of selection for resistant strains and thereby onward transmission of resistant strains.[11,135,154,216]

However, TDR prevalence among overseas-acquired and overseas-born cases is still considered moderate by WHO standards and almost all those with TDR carried single class resistance to either NNRTIs or NRTIs, indicative of selection pressure while on triple combination therapy. It is possible some of these TDR mutations were acquired DR mutations. Although people are carefully screened at time of diagnosis in Australia about past diagnosis and treatment history they may not always report accurate information. It is also possible that some non-B cases with mutations at drug resistant sites were actually

polymorphisms that appeared after infection.[232] Conversely, the stage of infection at diagnosis was not available for analysis, but it is highly likely some overseas-acquired cases were chronic undiagnosed infections in which TDR mutations had reverted back to wild-type or had reduced to undetectable levels (i.e: not the dominant circulating strain any longer) and so TDR prevalence may actually be underrepresented.[216]

There were no TDR cases in the overseas-born population between 2010 and 2013, which reflects the extension of effective HAART access in low- and middle-income regions such as Africa, South America, and Asia, from where nine of the 13 overseas-born people originated. There have been major achievements with initiating ART in developing countries. Reports from Africa have shown that where ART is available there have been reductions in viral load similar to those from developed nations.[135] However, challenges still remain. Acquired and TDR surveillance is currently limited, which affects current practice and policy. ART use is increasing despite this lack of surveillance in a setting where viral load testing is sporadic, genotypic resistance testing not routinely conducted before treatment initiation, and treatment failure to first-line regimens is prolonged before switching to second-line treatments.[128] To halt the spread of TDR, it is essential that surveillance efforts are increased, genotypic resistance testing is conducted before the initiation of treatment and there is a high quality of care to ensure early diagnosis, high level treatment adherence and ongoing education about prevention.

5.10.4 TDR in the young population

Ten percent of TDR cases were people under 25 years of age, and the prevalence of TDR in the under 25 TN population was 14%. All were male and all but two were MSM with subtype B virus diagnosed between 2000 and 2004. Of concern in this young cohort is that all of them harbored high level resistance to one or more of the three drug classes, including three who carried high-level resistance to NNRTIs and NRTIs, and two with a single PI

mutation that confers resistance to all but two of the PI treatments. There are many implications to consider for young people beginning HIV treatment, including further drug resistance, treatment failure, treatment switching, adherence to treatment, and long-term toxicities caused by the length of time on treatment.[126]

There were also two young people with non-B viruses diagnosed in 2006 and 2008 respectively. Both were males born in Africa, one was aged 3 at diagnosis and the other aged 13. These cases have been described previously in this chapter. They are of concern because both children had high-level resistance to two of the most commonly used drug classes in first-line treatment regimens, which limits treatment options to PIs and the newer INSTIs, puts them in a high-risk category for treatment failure with further restrictions on treatment options, and exposes them to long-term toxicity side effects because of the need for lifelong treatment.[126]

5.10.5 Type of transmission

5.10.5.1 MSM contact

MSM represent a disproportionate number of people with TDR worldwide. This is not surprising given MSM represent a disproportionate number of people with HIV in high-income countries where treatment has been available for much longer. This higher prevalence of TDR in the MSM population is mainly due to risk behaviors such as unprotected sex, multiple concurrent partners and high frequency of partner swapping, including among untreated men both diagnosed and undiagnosed, and in the early stages of infection when TDR mutations are likely to be present.[241,242]

The prevalence of TDR among the MSM population has significantly decreased over the last 15 years in Australia, although it still has the highest rate globally.[18] The overall TDR prevalence among MSM in South Australia was approximately 20%, which is 5% higher than national rates and much higher than reported in North America (14%), Western Europe

(11%) and South America (8%).[18] This could be explained by the high prevalence of K103N forward transmission as discussed earlier. The current findings are also consistent with others who have found high prevalence rates of TDR among MSM compared with heterosexual and other populations.[72] However, TDR to all three of the drug classes decreased from extremely high prevalence to low prevalence. This contrasts with the increase in resistance to NNRTIs and PIs among MSM seen globally.[18] This global increase is partly caused by selection pressure of certain drugs in triple combination regimens, which may be the result of unequal adherence to each medication and drugs in the combination that have a lower genetic threshold for resistance. The decrease seen in the current study indicates that first-line combination treatment regimens within South Australia are effective at reducing viral load, have a high genetic threshold to resistance, and generate high medication compliance by combining multiple drugs into one pill to be taken once daily and reducing severe side effects. However, given TDR forms of HIV are still circulating in the MSM population, including the K103N mutation, there is still an urgent need for targeted prevention of transmission, and for mandatory genotypic resistance testing at time of diagnosis.

5.10.5.2 Heterosexual contact

In regions like North America, Australia, Western Europe and South America, HIV transmission through heterosexual sex or IDU is less common than through MSM but rates of TDR are quite high, while in places such as Eastern Europe, Central Asia, and some South-East Asian regions infections are predominantly transmitted through intravenous drug use yet the prevalence of TDR is quite low.[18] Where heterosexual transmission is most prevalent, such as sub-Saharan Africa, TDR is also quite low. The current study mirrors these findings. TDR prevalence among heterosexual people born and/or infected in Australia was moderate and has fallen significantly over time. Conversely, heterosexual Asian- and

African-born people had low TDR prevalence rates. A very high TDR prevalence was found among injecting drug users in the current study, in both the MSM and heterosexual populations born in Australia and overseas. People who inject drugs may be more at risk of TDR because of sporadic adherence to treatment, treatment interruption, and a higher risk of transmitting TDR before virus detection and diagnosis. The IDU population is more at risk of HIV in general because of the direct access to the bloodstream of an uninfected person, while people engaging in anal sex have also been found to be at higher risk because of the high number of CD4+ T cells found in the gastrointestinal tract, including the rectal mucosa which can be easily disrupted during intercourse and provides HIV access straight to the bloodstream.[243,244] The high TDR prevalence rates among IDUs in Australia compared with the relatively low prevalence recorded in LMICs can be explained by the longer time period of accessible ART and resulting resistance circulating in the population. The prevalence was high across the first two time periods, then dropped to a moderate level in the most recent time period, which correlates with the decrease in TDR noted in the MSM and heterosexual populations.

5.10.6 Strengths and Limitations

This study included all TN people newly diagnosed with HIV infection in South Australia between 2000 and 2013. This allowed an accurate calculation of the prevalence rate of TDR in the South Australian HIV-infected community over a long period of time and a through a number of changes in treatment practices, but caution is warranted when extrapolating results to other Australian states, or countries of origin for the cohort born or infected overseas. South Australia implemented routine genotype/drug resistance profiling within 12 months of diagnosis in 2000 and uses a standardized protocol for collection of notification data, which allowed analysis of transmitted surveillance mutations within the treatment naïve cohort.

Demographic and genetic data were available and these provided the unique opportunity to make comparisons between genetically diverse subtypes, regions of birth, regions of infection acquisition and risk behavior. This has assisted in understanding the circulating drug-resistant HIV-1 strains at the population level. To date, such data have not been published.

A limitation of the study was the lack of precise information about where the infection was acquired, which would have allowed for more detailed analysis of the origin of TDR strains. The time between infection acquisition and diagnosis was not available for analysis, so it is possible the proportion of individual mutations was biased because of reversion to wild-type or the resistant virus no longer being the dominant circulating strain.

5.10.7 Recommendations

Genotype surveillance that incorporates communication between virologists, practitioners, clinicians and people infected with HIV is crucial to enhancing our epidemiological knowledge about TDR and treatment response for treatment naïve people with diverse strains of HIV.[115] Sequence diversity and resistance in the global population needs to be monitored to aid treatment optimization and ensure the best treatment outcomes, especially for treatment naïve people. HIV-1 evolves very rapidly, and although subtype variation may not predict how a person will respond to treatment, emerging evidence supports the belief that natural polymorphisms or drug associated mutations that are common to certain subtypes and CRFs can affect treatment outcomes.[134]

Ideally, HIV-1 genotype and resistance testing should occur at time of diagnosis before any treatment has occurred. This not only aids population surveillance and development of guidelines but also allows clinicians to access genotype and mutation information to guide decision making about individual treatment plans. This has been progressively introduced in South Australia over the last decade, with the majority of newly diagnosed cases being

genotyped within 12 months of diagnosis according to South Australian guidelines, usually at time of diagnosis and again before treatment commencement.

However, this is not viable for many countries, particularly those with limited resources. Instead, the majority of resource-poor countries follow WHO HIV drug resistance guidelines, which include routine monitoring of factors associated with emerging drug resistance, surveys to assessing prevalence of drug resistance and to monitor the emergence of drug resistance, and program evaluation of populations receiving ART. Data from these surveys are then used to make decisions about treatment regimens. As findings from sub-Saharan Africa demonstrate, TDR is rising.[135] Given the expanding access to ART in Africa, the growing African community in Australia and the increase in new diagnoses that may or may not be chronic infections, we need financially viable ways to ensure early diagnosis and genotypic testing. Resistance data in pretreatment populations provide important information about the probable effectiveness of available regimens for each region, and lessen the probability that the virus has reverted to wild-type or that mutated strains have declined to levels undetectable by population-based genotyping. Such information also ensures that appropriate first-line treatment is given in a timely manner. Targeted prevention and support strategies among migrant and refugee communities, should also be a priority.

Triple combination therapies are much more effective in suppressing viral replication. This, combined with improved treatment strategies to ensure adherence, more tolerable side effects and regular viral load checks for early management of treatment failure, have most likely led to the significant decrease in TDR.[115] The low prevalence of TDR viruses still circulating may be the result of poor engagement with medical care and low rates of ART adherence among those already on treatment. Given that the time between becoming infected and being diagnosed can be quite significant if there are no symptoms, and that there may 20% more people infected with HIV than are diagnosed, it is quite possible that resistant

viruses continue to be transmitted between undiagnosed people. This would explain the large number of K103N viruses diagnosed between 2003 and 2004, and again between 2008 and 2009. People in the early stages of HIV infection who are unaware of their infection or who may know they are infected but have not yet started treatment may be responsible for a large number of forward transmissions of HIV, because this is the time period where HIV is replicating most rapidly and RNA levels are highest.[229]

Further studies should include surveillance data on stage of infection at diagnosis, virologic failure, first and subsequent drug regimens used, treatment adherence levels, CD4+ cell count and viral load. It would also be beneficial to investigate transmission patterns and networks by phy. This could be used to validate personal reports at time of diagnosis and assist with contact tracing, and to identify transmission patterns and networks. Findings could then be used to develop targeted prevention and education strategies.[161] However, caution must be taken with this approach because attempting to link viruses through transmission and possible contact tracing can invoke a number of ethical and legal issues around patient confidentiality and disclosure of HIV status.[245]

5.10.8 Conclusion

In conclusion, the findings show that there is a high but decreasing level of TDR in the South Australian HIV-infected population, largely stemming from forward transmission of the K103N mutated virus in the subtype B MSM cohort before 2010. There is also moderate but decreasing levels of TDR in people born or acquiring their infection overseas, predominantly NNRTI and NRTI resistance which is also seen in host countries where ART has been introduced relatively recently. Although decreasing, the moderate level of TDR is still of concern. In a world where ART is rapidly expanding, early diagnosis, pre-treatment drug resistance testing, effective treatment regimens and ongoing surveillance are of the utmost importance. Despite decreasing rates, the number of TDR strains circulating in the

MSM population in Australia remain a concern, which highlights the need for continued surveillance, education and early diagnosis. There is also a steady influx of people migrating from LMICs to Australia, including those with HIV infection and TDR. It is essential that HIV testing is encouraged and supported, to identify infection early and test for resistance before treatment. Resistance patterns in non-B persons infected overseas may influence treatment choice and viral suppression beyond our current understanding of the historical subtype B infection circulating in Australia, and these cases should be monitored closely.[232]

CHAPTER 6: PHYLOGENETIC ANALYSIS

6.1 Overview

A comprehensive understanding of transmission patterns, subtype distribution, and risk factors is crucial when designing targeted prevention strategies for different population groups and networks to halt onward transmission of HIV.

The introduction of routine drug resistance testing of the HIV-1 *pol* gene in South Australia has provided a wealth of generated sequence data. This allows the assessment of subtype distribution in the population and to conduct phylogenetic analyses to investigate the virus in ways that were not possible using non-genetic surveillance data alone.[194] It also provides insight into whether routine information collected at the time of diagnosis is accurate, such as reports of transmission risk, location of infection acquisition etc.

The sequencing of multiple genes increases the robustness of the analysis of genetic diversity. By combining epidemiological and genetic data, HIV strains can be closely examined for evidence of recombination and mutation and this information can then be compared with demographic information. The results can be used to identify population transmission patterns, subtype distribution, strains that are currently circulating and by what exposure risk, and which strains are being imported or transmitted locally.

In Chapter Four, results of analysis of *pol* gene sequence data which represented all newly diagnosed people in South Australia between 2000 and 2013 showed a subtype distribution changing significantly from predominantly subtype B infections circulating among the Australian MSM population to an extensive range of B and non-B subtypes and CRFs, including *pol*-ISRs, circulating in MSM and heterosexual populations and in children and adults through direct blood contact. It was also found that many non-B subtypes and *pol*-ISRs were being imported and there was evidence of local ongoing transmission.

Despite the wealth of routinely collected sequence data, genetic diversity and transmission patterns within the South Australian HIV population remain largely

unexplored. In the present study this diversity is examined in more depth using phylogenetic reconstruction of HIV-1 *pol* and *env* gene sequences.

Phy can be used alongside more traditional epidemiological methods such as contact tracing to identify evolutionary relationships among patient sequences and understand HIV transmission dynamics globally. It can often confirm or refute personal reports and this leads to more accurate network mapping, identification of high risk populations and identification of broader clusters of people between whom the virus has been transmitted over time.[163] This information can then be used for prevention and intervention strategies. The comparison of genetic sequences can also shed light on intersubtype diversity between people who share similar strains of the same subtype or CRF.

Phylogenetic analyses were performed using *pol* and *env* sequences of 221 newly diagnosed cases in South Australia between 2000 and 2012, to ascertain subtypes and to explore subtype and demographic characteristics of people with sequences in highly reliable clusters (≥70% bootstrap value from common ancestral node) compared with people who did not. Demographic characteristics for those that formed part of transmission clusters were also examined.

Query sequences of interest were subtyped using phy, according to which reference sequences they most closely clustered with. Sequences that clustered together with an ancestral node bootstrap value of ≥70% were called 'high reliability clusters' and sub-clusters were defined as sequences with an ancestral node bootstrap value of ≥70% located within the larger ≥70% clusters.

Transmission cluster membership was defined as two or more sequences with a bootstrap value of 98% or higher, and each sequence with a genetic distance of ≤1.5% from at least one other sequence in the cluster.[163] These sequences may be directly related (Persons A and B; Person A transmitted the virus to Person B), related by a shared transmission (Persons A,

B and C; Person B transmitted the virus to Person A and Person C in the cluster), or related by intermediary transmission (Person A infects Person B, then Person B infects Person C and so forth). Within one cluster there may be a person who has transmitted the virus to multiple people within that cluster, and there could also be people in the cluster who have been infected by someone who has not yet been diagnosed and is therefore not in the cluster, while other people in the cluster have also been infected by the undiagnosed person.

6.2 Cohort demographics

The cohort consisted of 174 males and 47 females. Ninety people were born in Australia, 85 were born overseas and 46 did not have a listed region of birth. The mean age at diagnosis was 37 years (SD=12, range 3–72). There were 114 cases acquired within Australia, 100 acquired overseas and seven with an unlisted location. Route of transmission was heterosexual (including IDU risk) in 100 cases, bisexual in five, MSM contact (including IDU risk) in 92, medical procedure or MTCT in 15 and nine cases did not have a listed transmission risk.

Between 2000 and 2006 there were 92 cases newly diagnosed, with 129 between 2007 and 2012. These year groupings were chosen for statistical power purposes, so univariate analysis could be conducted. Plasma samples for cases diagnosed in 2013 were not available for *env* sequencing at the time of the study, due to clinical reasons.

6.3 PCR amplification and sequencing of *env*-gp41 fragments

In total, multiple PCR amplification attempts for *env*-gp41 were conducted on 332 samples (from 296 unique cases). There were 42 cases that had one or more plasma samples available for amplifying/sequencing *env*-gp41, but after multiple attempts this was unsuccessful. These 42 cases were diagnosed between 2000 and 2006, and the plasma samples were taken between these dates. Twenty-six (62%) had Stanford CPR determined *pol* subtype B, 11 (26%) 01_AE, three (7%) subtype C, one (2%) 01_AG, and one (2%) had

pol-ISR B/C. These 42 cases were excluded from the study. A further 33 cases were excluded because one or more of the inclusion criteria were not met, for example being diagnosed interstate or before 2000.

Of the available *pol* sequences for the 513 newly diagnosed cases in South Australia between 2000 and 2012, 45% (233) were utilized, that had been previously created and stored on the secure SA Pathology server. Each represented a single case except for 12 pairs of multiple sequences used as controls. A total of 233 samples from the same 221 unique cases were used to amplify and sequence the *env*-gp41 region. Examples of PCR results (visualized on agarose gel) are illustrated in Figure 21.

Nearly all (91%, 202) cases had *pol* and *env* sequences taken either from the same plasma sample or from plasma samples collected less than 12 months apart. Study information by date of diagnosis is shown in Table 18. Overall, 41% of the total number of newly diagnosed and subtyped cases between 2000 and 2012 were included in this study, with larger percentages from 2009, 2010 and 2011. This was governed by the availability of plasma samples for extraction.

Table 18. Cases used for *pol* and *env* phy

Year dx	Total number originally subtyped N	*pol*/*env* sequenced cases N	%	*pol/env* seq from same sample or samples ≤12 months apart	%
2000	23	7	30%	5	71%
2001	41	16	39%	10	63%
2002	21	6	29%	6	100%
2003	39	12	31%	9	75%
2004	47	20	43%	19	95%
2005	41	12	29%	11	92%
2006	46	19	41%	17	89%
2007	38	14	37%	12	86%
2008	42	10	24%	10	100%
2009	49	28	57%	26	93%
2010	38	34	89%	34	100%
2011	56	37	66%	37	100%
2012	32	6	19%	6	100%
TOTAL	**513**	**221**	**41%**	**202**	**90%**

Key: Year dx (year of diagnosis), Total number originally subtyped (number of cases newly diagnosed between 2000 and 2012 that had a stored *pol* subtype on file), *pol*/*env* sequenced cases (number and proportion of cases with a *pol* and *env* sequence that were used for this study). All *pol* subtyping in this table was based on the Stanford CPR online subtyping tool as per subtyping protocol at SA Pathology ≤http://sierra2.Stanford.edu/sierra/servlet/JSierra?action=sequenceInput≥.

Twelve pairs of sequential sequences from 12 different cases (sequence pairs known to be from the same person) were included in each tree as quality controls to ensure accurate calculation of the MEGA phylogenetic tree. The extra sequence was derived from a plasma sample taken in a later year. Each pair is labelled the phylogenetic trees (Figure 22 and Figure 23). Each of the 12 cases had both the *pol* or *env* sequences clustering together on the respective trees, Table 19. There was a large variation in bootstrap values with an average of 78% (range, 23–100%) and the average genetic distance between the two sequences for each case was 0.9% (range, 0–3.3%). Seven of the 12 *pol* pairs had bootstrap values of ≥95%

and genetic distances of ≤2.1%, 10 *env* pairs had bootstrap values of ≥95% and genetic distances of 3.3%. Lower bootstrap values and higher genetic distance may have been caused by a variety of factors including number of years between the two samples, natural genetic drift, treatment interruption, high mutation rate of a particular subtype/CRF, or virus reversion to wild-type. Cases 1 and 11 had sequences from samples taken a number of years apart, with a much higher viral load in the first sample compared with the second, it was likely they experienced issues with treatment failure as evidenced by frequent viral load changes between the dates. Later sequences from cases 1, 5, and 11 also carried a number of nucleotide changes which may have been due to treatment failure, or natural evolution during treatment interruption. Cases 4, 7 and 9 carried drug resistance mutations and likely experienced treatment failure at least once between the times the two sequences were taken.

The overall results show that the *pol* and *env* ML trees were robust. The most recent plasma sample sequence for each of the 12 cases was then excluded for the remainder of the analyses.

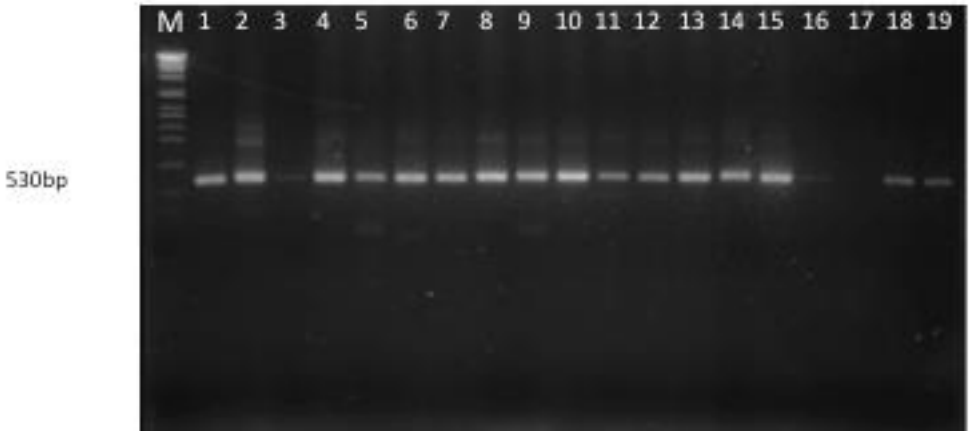


Figure 21. Agarose gel electrophoresis of amplified *env*-gp41 HIV-1 PCR products. Lanes: Lane M – 100-bp marker, Lanes 1–16 – 530-bp *env*-gp41 PCR products (negative outcome for samples in lanes 3 and 16) Lane 17 – negative control, Lanes 18–19 – positive controls.

Table 19. Cases with sequential *pol*/*env* sequences for quality assurance of ML trees

		pol					***env***				
Case No	**Year Dx**	**Sample 1/2**	**Same cluster**	**B/S (%)**	**Genetic distance (%)**	**phy**	**Sample 1/2**	**Same cluster**	**B/S (%)**	**Genetic distance (%)**	**phy**
1	2003	2003/2008	Y	97	0	01_AE	2003/2008	Y	95	0.3	01_AE
2	2004	2008/2011	Y	99	0	D	2008/2011	Y	96	0	D
3	2005	2005/2011	Y	100	0.3	C	2005/2011	Y	99	0.9	C
4	2005	2005/2011	Y	57	1.1	B	2005/2011	Y	92	0.6	B
5	2005	2005/2007	Y	23	1.7	B	2005/2007	Y	96	0	B
6	2006	2009/2011	Y	100	0.1	02_AG	2009/2011	Y	99	0	02_AG
7	2006	2006/2010	Y	55	1	B	2006/2010	Y	39	3.1	B
8	2009	2010/2011	Y	99	0.7	C	2010/2011	Y	99	3.3	C
9	2000	2004/2010	Y	99	2.1	B	2004/2010	Y	99	1.9	B
10	2001	2001/2010	Y	94	1.8	01_AE	2003/2010	Y	99	1.6	01_AE
11	2003	2003/2011	Y	37	1.7	B	2003/2011	Y	42	2.6	B
128	2009	2009/2012	Y	99	0.2	01_AE	2011/2012	Y	99	0.3	52_01B

Key: Year Dx (year of diagnosis), Sample 1 and 2 refer to the years that each of the two samples used for sequencing was taken, B/S (bootstrap).

6.4 Subtyping and cluster analysis of the partial *pol* and *env* regions

The following sections report phylogenetic tree analysis using both the *pol* and *env* genes, to determine subtype and transmission clusters within the HIV cohort diagnosed between 2000 and 2012.

Initial subtyping of the *pol* and *env* sequences was determined using maximum likelihood analysis to construct a tree phylogenetically (phylogenetic tree). Sequences were subtyped where the genetic distance unambiguously resolved the sequence to a particular cluster. Cluster analyses were then conducted for the trees, assessing query sequences that were part of highly reliable clusters and those that formed part of transmission clusters. Demographic analyses were also conducted on both types of clusters.

Online virus subtyping tools (Stanford CPR, jpHMM, REGA, SCUEAL, COMET and LANL BLAST) were used to assess the accuracy of the phylogenetic tree, and to uncover any potential unique ISRs. These will be discussed in Chapter Seven. A complete table of case subtype results by phy and online tools is presented in Appendix One.

6.5 Subtype distribution

The phylogenetic trees derived from the selected *pol* and *env* sequences are presented in Figures 22 and 23. Subtype distribution was similar between the two trees, with 45% (100/221) of the total *pol* sequences clustered with pure subtype B compared with 52% (104/221) of the total *env* sequences, Figures 22 and 23. The remaining 121 *pol* sequences clustered with three pure non-B subtypes and ten CRFs, and the remaining 117 *env* sequences clustered with three pure non-B subtypes and 12 different CRFs, Figures 22 and 23. Both trees showed a similar overall topology, with all cases sitting within the same broad clusters in both trees. Bootstrap values of 70% or higher are indicated on the branches. The number of sequences within high reliability clusters differed slightly between the two trees, with 53% (118/221) of *pol* sequences being part of high reliability clusters and 56% (124)

of *env* sequences being so. The remaining sequences either clustered with other sequences with a bootstrap value less than 70% (n=93 *pol*, and n=90 *env*), or were outliers, sitting outside of clusters (n=10 *pol*, n=7 *env*).

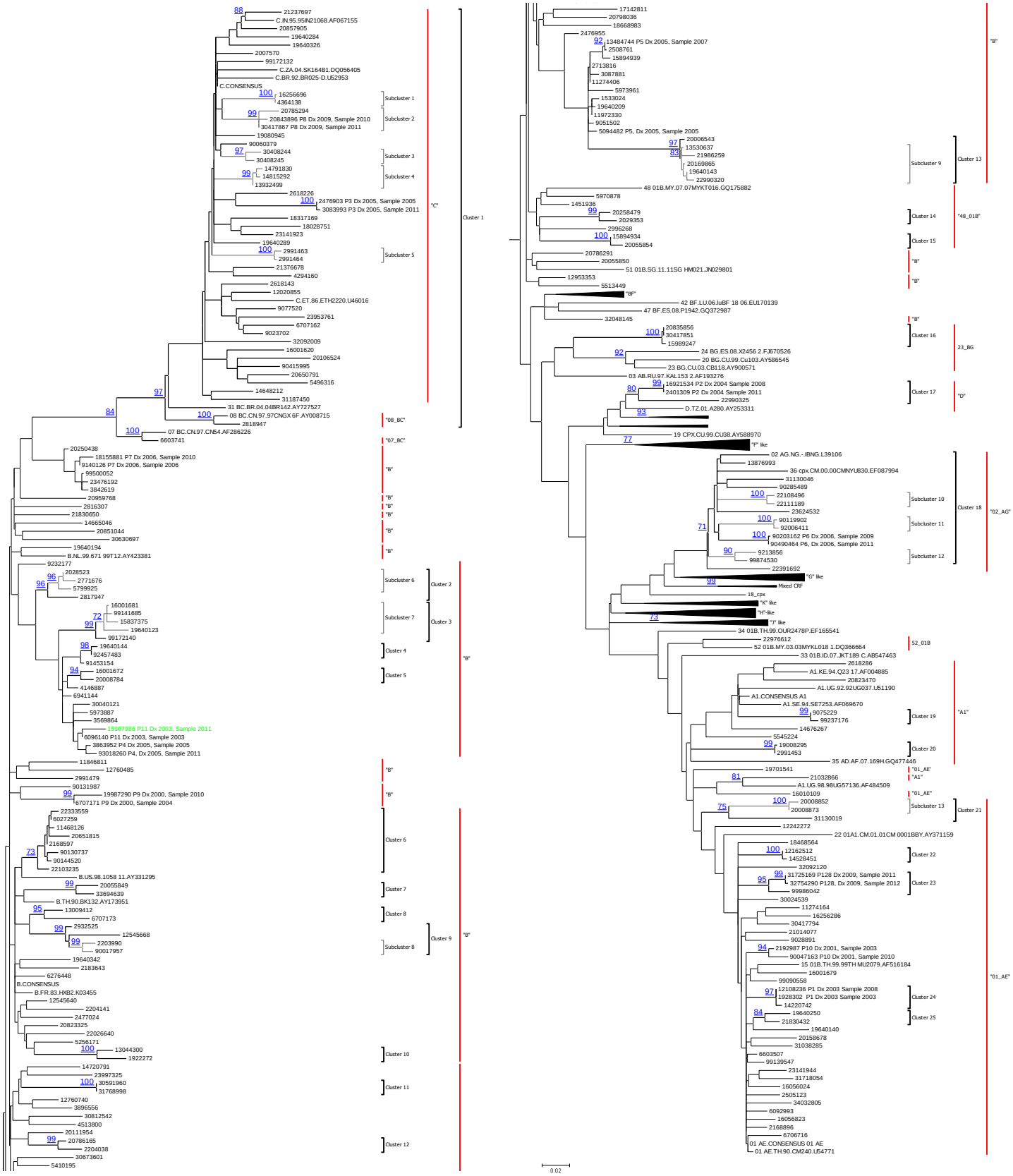


Figure 22. Maximum likelihood tree representing the phylogenetic relationships between HIV-1 *pol* sequences. The tree was constructed using the GTR+I+G model of evolution and query sequences were aligned against references sequences taken from the Los Alamos HIV database. Bootstrap values ≥70% are indicated on the branches. Red lines denote reference subtypes and CRFs that query sequences clustered or paired with; Cluster (sequences that paired or clustered together with a bootstrap value of ≥70%); Sub-Cluster (sequences that paired or clustered together within a larger clusters and sub-cluster had a bootstrap value ≥70%); P (Person); Dx (Year of diagnosis); S (Sample and year it was taken).

Figure 23. Maximum likelihood tree representing the phylogenetic relationships between HIV-1 *env* sequences. The tree was constructed using the GTR+I+G model of evolution and query sequences were aligned against references sequences taken from the Los Alamos HIV database. Bootstrap values ≥70% are indicated on the branches. Red lines denote reference subtypes and CRFs that query sequences clustered or paired with; Cluster (sequences that paired or clustered together with a bootstrap value of ≥70%); Sub-Cluster (sequences that paired or clustered together within a larger clusters and sub-cluster had a bootstrap value ≥70%); P (Person); Dx (Year of diagnosis); S (Sample and year it was taken).

6.6 High reliability cluster analysis

All cases that clustered together in the *pol* tree also clustered together in the *env* tree but with varying bootstrap values. In total, there were 170 (77%) cases that met the criteria of being either a *pol* or *env* high reliability cluster, or both. The *pol* tree identified clusters that had bootstrap values of ≥70% where the same clusters were found in the *env* tree with bootstraps ≤70% and *vice versa*; of the 170 cases in a high reliability cluster, 27% (46) were part of a *pol* ≥70% cluster only, 31% (52) an *env* ≥70% cluster only and 42% (72) were part of a *pol* and *env* ≥70% cluster.

There were 103 *pol* sequences that did not form part of a high reliability cluster. These were most closely related to pure subtypes or CRFs in the phylogenetic tree, as seen in Table 20.

Table 20. *pol* sequences that were not part of high reliability clusters.

phy subtype	N
A1	6
B	59
01_AE	25
02_AG	1
07_BC	1
15_01B	2
47_BF	1
48_BF	3
51_01B	4
52_01B	1

There were 97 *env* sequences that did not form part of a high reliability cluster. These were most closely related to pure subtypes or CRFs in the phylogenetic tree as seen in Table 21.

Table 21. *env* sequences that were not part of high reliability clusters.

phy subtype	N
A1	6
B	45
C	29
02_AG	5
03_AB	1
07_BC	1
08_BC	1
14_BG	1
15_01B	2
28_BF	1
36_cpx	1
45_cpx	1
47_BF	3

6.6.1 pol tree

A total of 25 *pol* high reliability clusters were identified from the tree topology (Figure 22), comprising 53% (118) of cases, Table 22. Most (64%, 16/25) were pairs. Just over half the pairs/clusters were cases with *pol* sequences clustering with non-B (including 'B-like') reference sequences, Table 23. Table 22 shows the proportion of cases that were part of high reliability clusters by subtype/CRF. All subtype C sequences were part of high reliability clusters, compared with 31% and 41% of 01_AE and subtype B sequences respectively. Sequences that clustered with 07_BC, 15_01B, 47_BF, 51_01B or 52_01B (n=9) did not form part of any high reliability *pol* cluster.

The largest *pol* cluster was comprised of 42 sequences, 41 were assigned subtype C and one was sub-clustered with 08_BC. This large cluster had a bootstrap value of 97%, indicating it was highly reliable, and the sequences were genetically similar (<10% genetic distance). The *env* sequences for these 42 cases similarly formed one big cluster but with a bootstrap value of 67% and a larger range of genetic distance. Only 11 of the sequences were part of high reliability clusters, Table 23.

Table 22. Proportion of subtypes/CRFs that formed part of a *pol* high reliability cluster

pol	**Total sequences (N=221)**	**Sequences in high reliability clusters (N)**	**Proportion in high reliability clusters**
01_AE	36	11	31%
02_AG	12	11	92%
08_BC	1	1	100%
23_BG	3	3	100%
48_01B	7	4	57%
A1	10	4	40%
B	100	41	41%
C	41	41	100%
D	2	2	100%

The second largest *pol* cluster had 11 02_AG sequences, with a bootstrap value of 71%. All the matching *env* sequences also clustered together but with very low bootstrap value however six of the *env* sequences sub-clustered together in three pairs with high bootstrap values.

There were 12 subtype B *pol* clusters, ranging from pairs (n=7) to a cluster of eight (n=1), and 40 subtype sequences in total, Table 23. All subtype B pairs/clusters had bootstrap values of 94–100%, except the cluster of eight, which had a value of 73%. The 40 matching *env* sequences all clustered together in one group with a very low bootstrap value, but 31 were part of smaller high reliability sub-clusters.

There were five 01_AE high reliability *pol* clusters, four pairs and one cluster of three. The cluster of three had a bootstrap value of 75% and two of the three sub-clustered together with a bootstrap ≥98%. The same two also had *env* sequences clustered together with a bootstrap of 99%. The four *pol* pairs had bootstrap values of 84%, 95%, 97% and 100% respectively and all eight of these cases had *env* sequences clustered together with a bootstrap value of 90%.

The two A1 pairs, 48_01B pairs, and cluster of three 23_BG sequences all had bootstrap values ≥98%, and the cluster patterns were the same for the *env* sequences, all with bootstrap values ≥90%. The subtype D pair clustered with a bootstrap value of 80% and had a matching *env* pair with a bootstrap value of 94%.

Table 23. High reliability *pol* clusters

***pol* gene**	**A1**	**B**	**C**	**D**	**01_AE**	**02_AG**	**23_BG**	**48_01B**	**Total**
Pairs	2	7		1	4			2	**16**
3					1		1		**2**
4		2							**2**
5		1							**1**
6		1							**1**
8		1							**1**
11						1			**1**
42			1*						**1**
Total	**2**	**12**	**1**	**1**	**5**	**1**	**1**	**2**	**25**

Key: * (one sequence in the subtype C cluster of 42 was sub-clustered with 08_BC)

Thirty-three of the 118 (28%) clustered *pol* sequences also formed part of 13 sub-clusters, Table 24. Again, these were predominantly pairings (69%, 9/13). Sub-clusters were found in the subtype B, C, and CRF01_AE and 02_AG larger clusters, Table 24.

Five sub-clusters were found in the large cluster of 42 subtype C cases, consisting of four pairs and one cluster of three.

Four sub-clusters were found in four larger subtype B clusters. A sub-cluster of five was found in the subtype B cluster of six, a sub-cluster of four was found in a cluster of five, a sub-cluster of three was found in a cluster of four, and a sub-cluster pair was found in a cluster of four.

One 01_AE sub-cluster pair was found in a larger cluster of three 01_AE sequences, and three 01_AG sub-cluster pairs were found within the 02_AG cluster of 11 cases.

Table 24. High reliability *pol* sub-clusters

***pol* gene**	**B**	**C**	**01_AE**	**02_AG**	**Total**
Pairs	1	4	1	3	**9**
3	1	1			**2**
4	1				**1**
5	1				**1**
Total	**4**	**5**	**1**	**3**	**13**

6.6.2 env tree

Over half (56%, 124) of the 221 *env* sequences formed part of 24 high reliability pairs/clusters, Table 25. As with the *pol* sequences, the majority of sequences (67%, 16/24) were clustered in pairs, Table 26. The majority of pairs and clusters clustered with non-B subtypes (including 'B-like') on the *env* tree. Table 25 below shows the proportion of sequences that were part of high reliability clusters, by subtype/CRF.

In contrast to the *pol* sequences, only 28% of subtype C sequences were part of high reliability clusters. All *env* 01_AE sequences were part of high reliability clusters as were 57% of subtype B sequences.

Cases with phy-assigned 03_AB, 07_BC, 08_BC, 14_BG, 15_01B, 28_BF, 36_cpx, 45_cpx and 47_BF (n=12), did not form part of any *env* high reliability clusters.

The largest high reliability cluster was comprised of all 36 01_AE sequences from the cohort, with a bootstrap value of 90%. The *pol* sequences for these 36 cases all broadly clustered together but in smaller groups rather than one big cluster, and only 10 of the 36 *pol* sequences were part of ≥70% clusters (5 pairs). One of the five pairs was assigned subtype A1 on the *pol* phylogenetic tree, the other four 01_AE.

There were 10 subtype B clusters in total, ranging from pairs (n=6) to a cluster of 26 (n=1), with bootstrap values ranging from 86% to 99%. All other subtype/CRF clusters had

bootstrap values ranging from 75% to 100%.

The second largest cluster on the *env* tree was the group of 26 subtype B sequences with a bootstrap value of 91%. The *pol* sequences for these 26 cases also clustered together (in smaller groups) but only 13 sequences were in ≥70% clusters (2 pairs, one cluster of four and another cluster of five).

Table 25. Proportion of subtypes/CRFs that formed part of an *env* high reliability cluster

env	**Total sequences (N=221)**	**Sequences in high reliability clusters (N)**	**Proportion in high reliability clusters**
01_AE	36	36	100%
02_AG	12	7	58%
29_BF	2	2	100%
A1	12	6	50%
B	104	59	57%
C	40	11	28%
D	3	3	100%

Table 26. High reliability *env* clusters

***env* gene**	**A1**	**B**	**C**	**D**	**01_AE**	**02_AG**	**29_BF**	**Total**
Pairs	3	6	4			2	1	16
3			1	1		1		3
4		1						1
6		1						1
11		1						1
26		1						1
36					1			1
Total	**3**	**10**	**5**	**1**	**1**	**3**	**1**	**24**

Just over one-quarter (26%, 32) of clustered *env* sequences also formed part of 15 sub-clusters (Table 27) and all but one of the 15 were pairings, predominantly assigned subtype B (9 pairs), 01_AE (4 pairs), and 02_AG (1 pair). There was also one sub-cluster of 4 cases within the cluster of 26 subtype B cases.

Table 27. High reliability *env* sub-clusters

***env* gene**	**B**	**01_AE**	**02_AG**	**Total**
Pairs	9	4	1	14
4	1			1
Total	**10**	**4**	**1**	**15**

6.7 Transmission cluster analysis

As stated earlier, transmission clusters were defined as sequences that paired/clustered together with bootstrap values ≥98% and all sequences within the cluster having a genetic distance of ≤1.5% from at least one neighbour. Figures 24 and 25 show the transmission clusters on the phy *pol* and *env* trees respectively.

6.7.1 pol tree

Fifteen percent (33) of all *pol* sequences were identified as being part of 12 transmission pairs and three transmission clusters, Table 28.

Four *pol* A1 and three *pol* 23_BG sequences were part of transmission clusters, Table 29. Of the 11 02_AG sequences in high reliability clusters, four were part of transmission clusters. Only 9% of subtype B sequences were found in transmission clusters compared with 41% found in high reliability clusters, and only 17% of the subtype C sequences compared with the 100% of subtype C sequences found in high reliability clusters, Table 29. None of the 08_BC or D sequences that formed part of high reliability clusters were part of transmission clusters.

Table 28. *pol* transmission clusters

***pol* gene**	**A1**	**B**	**C**	**01_AE**	**02_AG**	**23_BG**	**48_01B**	**Total**
Pairs	2	3	2	2	2		1	12
3		1	1			1		3
Total	**2**	**4**	**3**	**2**	**2**	**2**	**1**	**15**

Table 29. Proportion of subtypes/CRFs that formed part of a *pol* transmission cluster

pol	**Total sequences (N=221)**	**Sequences in transmission cluster (N)**	**Proportion in transmission cluster**
01_AE	36	4	11%
02_AG	12	4	33%
23_BG	3	3	100%
48_01B	7	2	29%
A1	10	4	40%
B	100	9	9%
C	41	7	17%

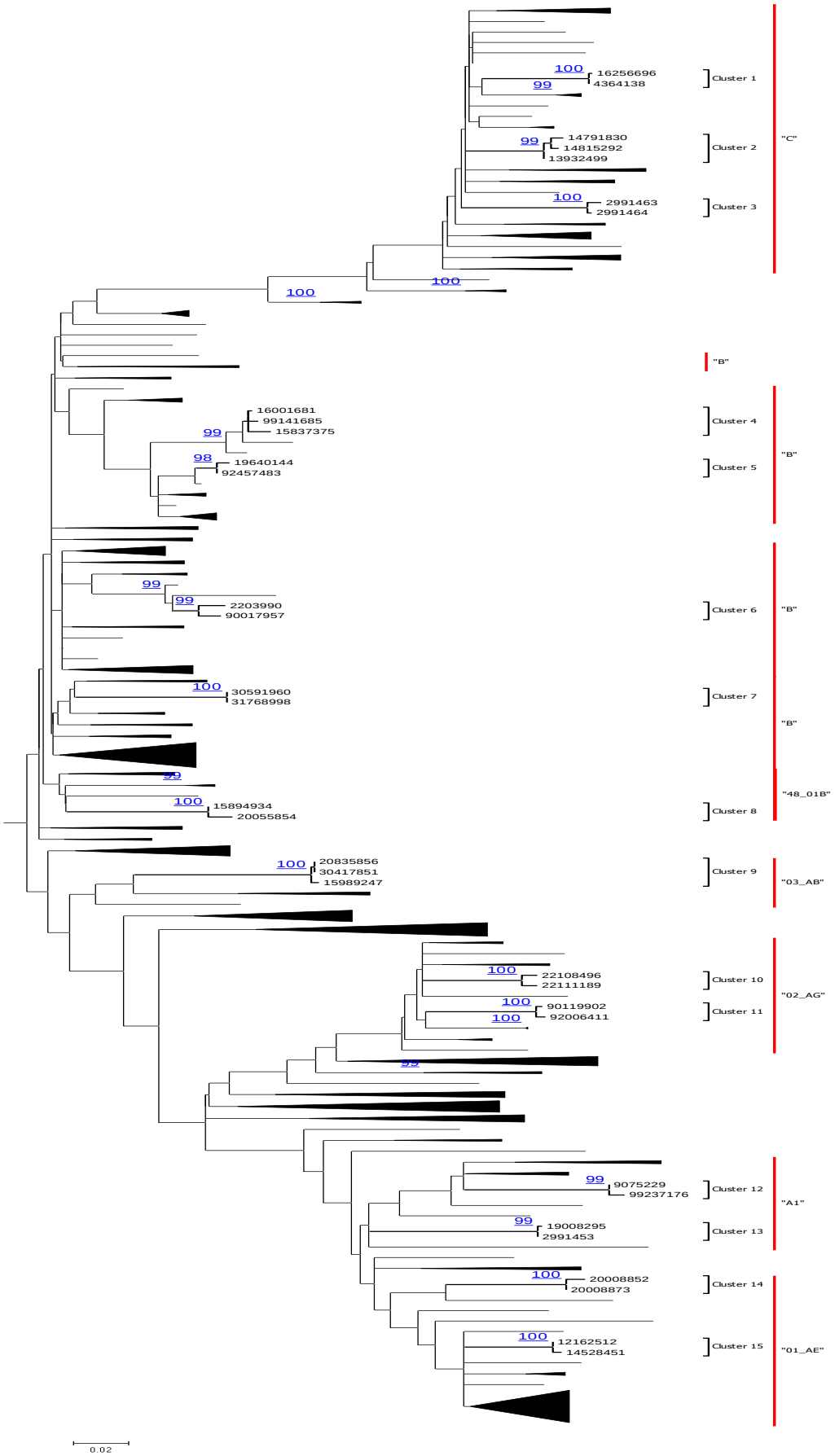


Figure 24. Transmission clusters in the *pol* phylogenetic tree.

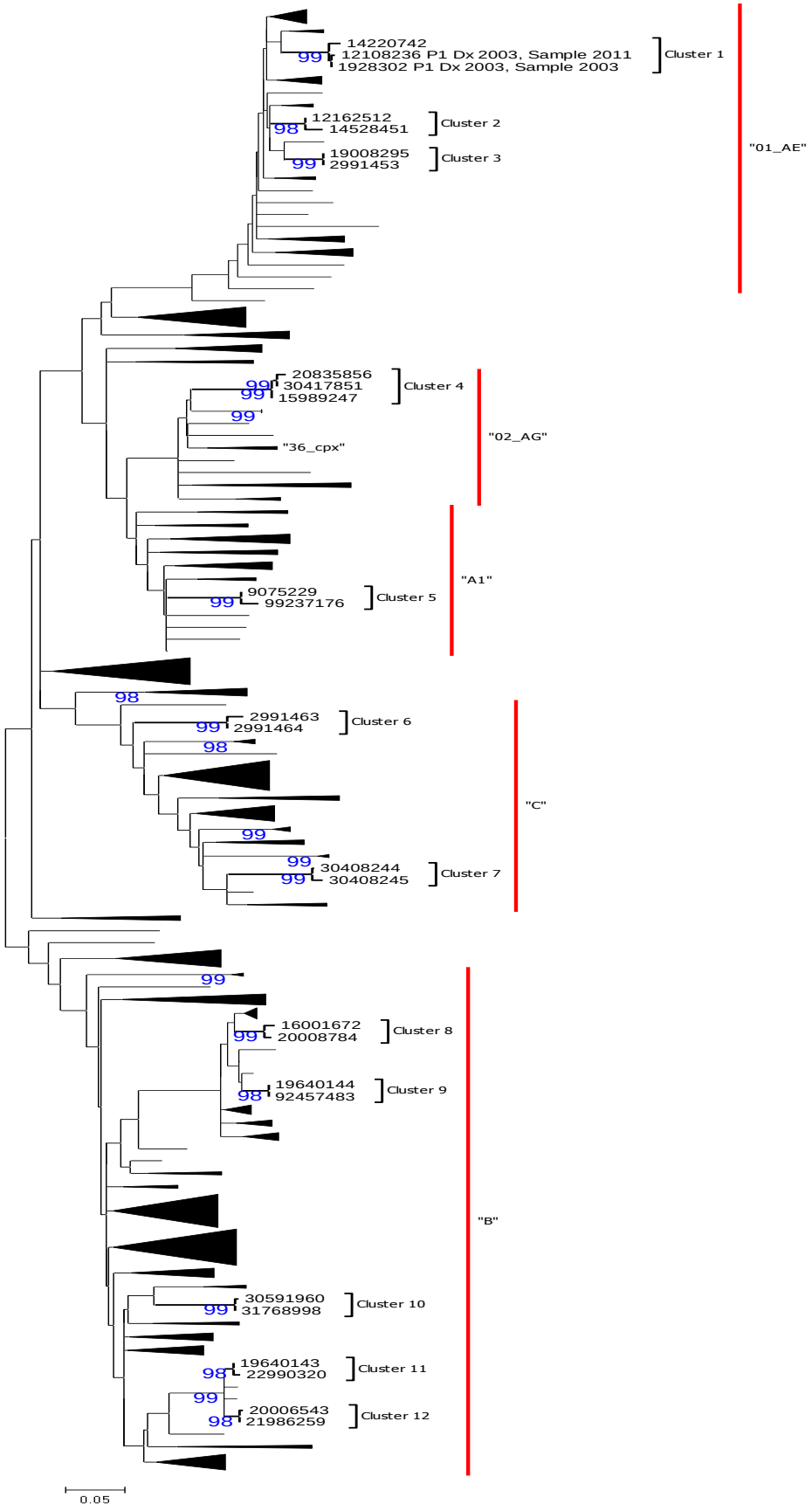


Figure 25. Transmission clusters in the *env* phylogenetic tree.

6.7.2 env tree

Twelve main transmission clusters were found within the *env* tree, comprised of 11% (25) of all *env* sequences, Table 30. All but one (92%, 11) were transmission pairs. The other was a transmission cluster of three people with CRF02_AG, Table 30.

All 01_AE sequences in the *env* tree formed part of high reliability clusters, but only 17% formed part of transmission clusters. There were 25% of all *env* 02_AG sequences that formed part of transmission clusters, 17% of subtype A1, and 10% each of B and C sequences, Table 31.

Table 30. *env* transmission clusters

env	**A1**	**B**	**C**	**01_AE**	**02_AG**	**Total**
Pairs	1	5	2	3		11
3					1	1
Total	**1**	**5**	**2**	**3**	**1**	**12**

Table 31. Proportion of subtypes/CRFs that formed part of an *env* transmission cluster

env	**Total sequences (N=221)**	**Sequences in transmission cluster (N)**	**Proportion in transmission cluster**
01_AE	36	6	17%
02_AG	12	3	25%
A1	12	2	17%
B	104	10	10%
C	40	4	10%

6.8 Univariate analysis of cases in clusters vs cases not in clusters

Univariate analyses were conducted to see if there were any significant demographic differences between people with sequences that clustered and those that did not. Characteristics for *pol* clusters are shown in Table 32 and *env* clusters in Table 33. Data for high reliability and transmission clusters were both reported.

6.8.1 Cluster membership – high reliability

There were significant associations between *pol* clustering and subtype, year of diagnosis, sex, and transmission risk ($p \leq 0.05$), Table 32. Only location of infection acquisition and region of birth were associated with *env* high reliability clusters, Table 33.

A higher than expected proportion of non-B cases was found in *pol* high reliability clusters ($p \leq 0.05$), and significantly more cases diagnosed in 2007–2012 were in clusters than those diagnosed in 2000–2006 ($p \leq 0.05$).

People infected in Australia were significantly more likely to have an *env* sequence in a high reliability cluster than those infected overseas ($p \leq 0.05$) but there was no association between location of infection acquisition and *pol* clusters, Table 32 and Table 33.

The observed frequency of Australian-born people in an *env* cluster was greater than expected ($p \leq 0.05$) but there were no differences between Australian- or overseas-born people being in *pol* clusters.

Females were significantly more likely to be part of *pol* high reliability clusters than males ($p \leq 0.05$) but there was no significant association between female *env* sequences and high reliability clustering.

A significantly higher than expected proportion of direct blood contact transmission cases (including MTCT) and heterosexual (including IDU risk) transmissions were part of ≥70% *pol* clusters ($p \leq 0.05$). There was no significant association between transmission risk and high reliability *env* clustering.

6.8.2 Cluster membership – transmission

There was a significant association between *pol* transmission clusters and subtype as defined by phy ($p \leq 0.05$), but not as defined by LANL BLAST ($p=0.31$). Year of diagnosis, sex and transmission risk were also significantly associated with *pol* transmission clusters (*p*

≤0.05), Table 32. Only year of diagnosis was associated with *env* transmission clusters, Table 33.

As with the *pol* high reliability clusters, a higher than expected proportion of phy defined non-B cases was found within *pol* transmission clusters, and significantly more cases were found in *pol* clusters when diagnosis occurred between 2007 and 2012 ($p \leq 0.05$), Table 32. Though there were no significant difference between being in an *env* high reliability cluster by year of diagnosis, there was a significant association between being in an *env* transmission cluster ($p \leq 0.05$) Table 33, with a higher than expected number of people in an *env* transmission cluster diagnosed between 2007-2012.

Females were significantly more likely to be part of *pol* but not *env* transmission clusters, ($p \leq 0.05$), Tables 32 and 33.

A significantly higher than expected proportion of cases with reported transmission through direct blood contact or heterosexual activity (including IDU risk) were found within *pol* but not *env* transmission clusters ($p \leq 0.05$), Tables 32-33.

Table 32. Factors associated with *pol* cluster membership (high reliability clusters and transmission clusters)

Characteristic (*pol* sequences)	Total	High reliability clusters (N)	%	*p* value	Transmission clusters (N)	%	*p* value
Cases	221	118	53%		33	15%	
Subtype B (phy)	100	41	41%	$p \leq 0.05$	9	9%	$p \leq 0.05$
Non-B (phy)	121	77	64%		24	20%	
Subtype B (LANL BLAST)	115	48	42%	$p \leq 0.05$	14	12%	$p=0.31$
Non-B (LANL BLAST)	106	70	66%		19	18%	
Year Dx	221	118			33		
2000–2006	92	39	42%	$p \leq 0.05$	5	5%	$p \leq 0.05$
2007–2012	129	79	61%		28	22%	
Age	214	117	55%		33	15%	
<25	25	17	68%	$p=0.10$	6	24%	$p=0.24$
25 and over	189	100	53%		27	14%	
Location acquired	214	117	55%		33	15%	
Australia	114	57	50%	$p=0.10$	16	14%	$p=0.57$
Overseas	100	60	60%		17	17%	
Region born	175	107	61%		32	18%	
Australia	90	49	54%	$p=0.09$	17	19%	$p=0.84$
Overseas	85	58	68%		15	18%	
Sex	221	118	53%		33	15%	
Male	174	85	49%	$p \leq 0.05$	21	12%	$p \leq 0.05$
Female	47	33	70%		12	26%	
Transmission Risk	212	115	54%		33	16%	
Direct blood contact/MTCT	15	13	87%	$p \leq 0.05$	5	33%	$p \leq 0.05$
Heterosexual (inc IDU risk)	100	64	64%		22	22%	
MSM (inc IDU risk)	97	38	39%		6	6%	

Key: Data represents number and proportion (%) of cases within each category. MSM (men who have sex with men), IDU (intravenous drug use), MTCT (Mother to child transmission). *p* values calculated using Pearson chi-squared or Fisher exact tests.

Table 33. Factors associated with *env* cluster membership (high reliability clusters and transmission clusters)

Characteristic (*env* sequences)	Total	High reliability clusters	%	*p* value	Transmission clusters	%	*p* value
Cases	221	124	56%		25	11%	
Subtype B (phy)	104	59	57%	*p*=0.86	10	10%	*p*=0.45
Non-B (phy)	117	65	56%		15	13%	
Subtype B (LANL BLAST)	103	54	52%	*p*=0.15	13	13%	*p*=0.71
Non-B (LANL BLAST)	108	70	65%		12	11%	
Unknown	10	7	70%		0	0%	
Year Dx	221	124	56%		25	11%	
2000–2006	92	49	53%	*p*=0.56	2	2%	*p*≤0.05
2007–2012	129	75	58%		23	18%	
Age	214	123	57%		25	12%	
<25	25	13	52%	*p*=0.71	4	16%	*p*=0.51
25 and over	189	110	58%		21	11%	
Location acquired	214	123	57%		25	12%	
Australia	114	76	67%	*p*≤0.05	15	13%	*p*=0.53
Overseas	100	47	47%		10	10%	
Region born	175	103	59%		25	14%	
Australia	90	65	72%	*p*≤0.05	14	16%	*p*=0.67
Overseas	85	38	45%		11	13%	
Sex	221	124	56%		25	11%	
Male	174	96	55%	*p*=0.62	17	10%	*p*=0.19
Female	47	28	60%		8	17%	
Transmission Risk	212	123	58%		25	12%	
Direct blood contact/MTCT	15	7	47%	*p*=0.66	1	7%	*p*=0.45
Heterosexual (inc IDU risk)	100	59	59%		15	15%	
MSM (inc IDU risk)	97	57	59%		9	9%	

Key: Data represents number and proportion (%) of cases within each category. MSM (men who have sex with men), IDU (intravenous drug use), MTCT (Mother to child transmission). *p* values calculated using Pearson chi-squared or Fisher exact tests.

6.9 Transmission clusters – case demography

Transmission clusters were also assessed by demographic information, to gain an understanding of patterns of infection in the population, such as exposure route and where

the infection was acquired. Both genes were assessed to see whether a greater number of transmission events were identified when using two genes compared to one. Only 7% (15) of cases met criteria for possible transmission events for both *pol* and *env* regions. Another 8% (18) met criteria for the *pol* region only, and 4% (10) for the *env* region only. In total, 43 cases were part of 20 possible transmission events, including 15% of all *pol* sequences and 11% of all *env* sequences. Combined, the *pol* transmission cases comprised 12 pairs and three clusters of three, while the *env* transmission cases comprised 11 pairs and one cluster of three.

All but one of the *pol*-only transmission pairs/clusters had *env* sequences that paired with bootstrap values of 90% or higher, and genetic diversity ranging from 1.9% to 4.9%. All *env*-only transmission pairs had *pol* sequences that paired or clustered (two *env* pairs had *pol* sequences that formed a cluster of four) with bootstrap values of 94% or higher, and genetic distances of less than 2%. This information indicated that the transmission events were real.

6.10 Transmission events

In total, there were 20 possible transmission events, seven of which were identified by both *pol* and *env* sequences, five by *env*-only and eight by *pol*-only. Male/female transmission pairs accounted for 55% of these events, followed by male/male pairs or clusters. One pair of child infections fitted transmission criteria, together with what appeared to be a mother/child pair, and a father, mother and child cluster. These will each be discussed in more detail below.

6.10.1 Male/female pairs

There were three male/female pairs in which the *pol* and *env* regions clustered near different CRFs or subtypes. Two of these pairs were typed as *pol* CRF01_AE and *env* A1, and all four people were diagnosed in 2009. One of the pairs was born in Africa, with the male infected overseas through blood contact and the female by heterosexual contact, while

the other pair was born in Asia and reported heterosexual transmission although only the male was infected overseas.

The other *pol/env* subtype discordant pair had *pol* 48_01B and *env* 29_BF, both were born in Australia and reported heterosexual contact. The male reported being infected overseas and probably transmitted to the female partner in Australia, both were then diagnosed in the same year (2011). The *env* sequences for this pair clustered together with a high bootstrap value and a genetic distance of 3.1%.

There were two 01_AE male/female pairs; one pair was born in Asia and infected overseas through heterosexual transmission, with the female diagnosed in 2003 and the male in 2007. The other pair was diagnosed in 2009, both Australian-born people who reported heterosexual contact with the male acquiring the infection overseas.

There were also two subtype C male/female pairs; one diagnosed in 2009 and the other in 2011. All four cases were people born in Africa and infected overseas, all reported heterosexual contact and all were aged in their 30s.

There was one male/female 02_AG pair, both in their 30s at age of diagnosis. Interestingly, the Australian-born female was diagnosed in 2004 and reported overseas acquisition while the African-born male was diagnosed in 2010 and reported being infected in Australia.

Two of the subtype B male/female pairs were born and infected in Australia, and all reported heterosexual or bisexual contact. One pair were both diagnosed in 2011, the other pair were diagnosed seven years apart. This latter pair had a *pol* genetic distance of 1.8% but when demographic data was examined it was considered to be a possible transmission link. Both people identified transmission through heterosexual contact with additional IDU risk and both were born and infected in South Australia. Furthermore, the *env* sequences for this pair also clustered with a bootstrap value of 99% and a genetic distance of 3.3%.

The final subtype B male/female pair met transmission criteria for the *env* sequence only, but the *pol* sequences for this pair clustered with *pol* sequences from a male/male subtype B pair that was also only identified by *env*. Diagnosis for all four people occurred between 2010 and 2011 and all were born and infected in Australia. The male/male pair reported MSM with IDU risk and the male/female pair reported bisexual and heterosexual contact respectively. The four *pol* sequences clustered with a 97% bootstrap value and maximum genetic distance of 1.6%.

6.10.2 Male/Male pairs or clusters

There were four male/male transmission pairs identified and two male-only clusters. Three pairs phylogenetically clustered with reference subtype B and the other with subtype C. The subtype C pair were both aged in their 50s, one reported heterosexual contact and the other MSM. The male was diagnosed in 2004 and did not have a region of birth listed but he reported an overseas-acquired infection, while the male diagnosed in 2006 was Australian-born and infected in Australia.

The three subtype B pairs were all MSM born and infected in Australia, one of the pairs has already been reported above as being linked to a male/female pair. Of the remaining two subtype B pairs, one pair was diagnosed in 2009 and in the other pair males were diagnosed in 2011 and 2012 respectively. All four males were aged between 25 and 40 years.

One male-only cluster was comprised of three males with *pol* sequences that clustered near the 23_BG reference sequence and *env* sequences that clustered near 02_AG. These males were diagnosed in consecutive years (2009, 2010, 2011). The first male diagnosed was born in Africa and infected overseas, reporting heterosexual transmission. The second male diagnosed was Australian-born and reported MSM transmission in Australia. The final male was African-born and reported infection through heterosexual transmission in Australia.

The other male-only cluster was comprised of three Australian-born males with subtype B, diagnosed in consecutive years (2007, 2008, 2009). They were aged in their 30s at time of diagnosis. The first male diagnosed reported heterosexual contact with IDU risk in Australia, the second reported heterosexual contact overseas, and the third male reported MSM contact in Australia.

6.10.3 Vertical transmission

There was one possible transmission pair between a mother and child, with a subtype A virus acquired heterosexually overseas by the mother and diagnosed in 2007 when she was aged 30. The child was reported as being born in Australia and acquiring the infection via MTCT. The child was diagnosed in 2008 at age 3 years.

There was also a possible transmission cluster between a father, mother and child with a subtype C virus. The adult male was aged 40 years, the adult female 28 years, and the child 7 years. They were all diagnosed in 2007 and all African-born and infected. Both adults reported heterosexual contact as the mode of transmission, with MTCT for the child. The *pol* sequences fit the transmission criteria, and all three *env* sequences clustered together with a bootstrap of 98% but the child's *env* sequence had a genetic distance of 4.9% from the mother, and 5.9% from the father.

6.10.4 Direct blood exposure

One pair consisting of two children aged 10 at diagnosis had *pol* sequences fitting transmission criteria. Their *pol* sequences clustered with reference sequence 02_AG and their *env* sequences clustered with A1. The children were diagnosed in 2009 and 2010 respectively, and both were born in the same region of West Central Asia, contracting HIV by direct blood contact. The *env* sequences clustered together with a bootstrap of 99% and a genetic distance of 2%.

6.11 Summary

Though the *pol* and *env* tree topologies were highly similar, with the same cases sitting with each other on both trees, the bootstrap values and genetic distances varied. This resulted in 46 cases being part of high reliability pairs or clusters only identified by *pol*, 52 only identified by *env* and the remaining 72 cases identified by both. Over half of all the cases comprised 25 separate high reliability *pol* pairs or clusters, including 15 transmission *pol* pairs/clusters, supported by high bootstrap values (≥70 and ≥98% respectively), and the genetic data was congruent with epidemiological data. Fifty-six percent of all cases comprised 24 high reliability *env* clusters, including 12 *env* transmission clusters.

The majority of high reliability groups were pairs for both the *pol* and *env* region, just under half were subtype B infections, and overall, more than half of all subtype B infections were not part of high reliability clusters. The largest high reliability *pol* cluster comprised all 41 subtype C cases and one 08_BC case, but the *env* regions of these cases clustered together with a bootstrap value of just under 70%. Conversely, almost 70% of 01_AE *pol* cases were not part of high reliability clusters but all of the matching 01_AE *env* regions were.

In total there were 20 identified transmission events, only seven were identified by both *pol* and *env*, eight were identified by *pol* only, and five by *env* only. The majority were male/female heterosexual pairs carrying non-B virus, followed by male/male pairs carrying subtype B, in which MSM, heterosexual contact and IDU were reported as the route of transmission. There was also a cluster of non-B cases which was a likely family transmission between mother, father and child, and a non-B child pair who likely acquired HIV via iatrogenic transmission in Central Asia.

Phy linked with demographic and epidemiological information highlighted the majority of subtype B infections are circulating within the local population, with some

evidence of overseas acquisition. The majority of non-B viruses are separate introductions via pairs and clusters of people being infected overseas and importing HIV infection into Australia via migration or travel overseas.

6.12 Discussion

Monitoring genetic diversity is not only important for vaccine development, diagnostic tools, blood screening, and clinical decisions around treatment regimen, but it also gives a clear picture of how different viruses are related to each other, and the degree of similarity or dissimilarity between viruses within a group.

Analysis of molecular sequencing via phylogeny is very useful for estimating evolutionary relationships of HIV infection, and identifying transmission dynamics and epidemic linkages. Linking phylogeny with epidemiological information such as route of transmission and geographic region of acquisition or birth, provides a robust way to closely monitor epidemics and identify priority areas and groups for prevention and intervention strategies. It has been shown genetic diversity continues to be generated through mutation and recombination and spread through population mobility worldwide.[75] Compared with current epidemiological methods such as contact tracing, phylogeny can provide scientific information to assess the epidemic structure. Contact tracing has proven successful at identifying undiagnosed infection, which obviously cannot be determined using phylogeny, but difficulty lies in often not knowing what stage of infection a person is at the time of diagnosis, how many people they may have infected, the rate of transmission per sexual act, number and contact details of sexual partners, and whether they can identify their origin of infection.[246]

Phy is therefore often used to confirm or refute patient reported linkages,[163,247] and linking these two types of data is important both for validation of sequence relatedness of

the tree, and of patient reports, which may not always be accurate.[162] Phylogeny has proved a very useful tool for revealing information, including identifying groups of infected people which share a common virus, whether this be direct transmission or part of a larger group,[248–254] identifying high-risk subpopulations, and imported infections that are genetically diverse. However, the use of phylogeny also comes with risk to patient confidentiality and steps must be taken to ensure privacy is maintained for individuals. Fear of social stigmatization often prevents people revealing information such as number of sexual partners, sexual practices, ethnic preference, use of condoms, location of infection and travel history.

Historically, phylogeny has also been used to identify transmissions between individuals and regardless of an infected person's intent, some cases have been prosecuted, including in Canada, the US, and Australia.[255–258] Despite claims it is possible to prove direction of HIV infection due to the genetic bottleneck that occurs during transmission and resulting paraphyly of source viruses within the recipient,[257] other researchers state it is only possible to determine whether viruses are likely to be linked by transmission, and direction cannot be determined with certainty.[258] Compounding this is the rapid pace of global HIV genetic diversity which is proving a challenge for phy, with a growing number of sequences unable to be classified with certainty.[194] Legislation which allows identification and prosecution of people infecting others with HIV can impact greatly on prevention efforts because HIV infected individuals may be less likely to access health services or to divulge their HIV status to others.[259] In the current study, careful consideration was given to the type and level of epidemiological data reported, for instance, regions were reported rather than individual countries (except Australia which was deemed to be large enough to ensure no possibility of identification). It is also acknowledged that there are multifaceted factors underlying the epidemiological data chosen to analyze. There are known associations between HIV, region of birth and socioeconomic status for example,[260] with clustering patterns likely to be

impacted by financial, social, environmental and political forces in particular countries rather than the actual birth region itself. HIV risk is also known to be higher among people who suffer from more structural barriers such as lower income, education, access to treatment and services, language barriers, higher rates of incarceration, and unemployment etc.[260] This study sought to limit or indeed prevent stigma of any particular high-risk or at-risk group. With that in mind, it was highly beneficial to link genetic and epidemiological data. There was a high congruency between both types of data, such as known subtypes and CRFs matching geographic association of infection and route of infection, cases clustering together on the trees with supporting epidemiological information, and strong evidence to suggest accurate self-reporting in this cohort.[194]

In the beginning of the Australian epidemic, the prevalence of HIV was highest among MSM, with subtype B the major circulating virus. Over the last decade the prevalence of non-B variants has increased including complex variants. These non-B viruses are predominantly imported from high prevalence countries but there is evidence of increasing transmission locally.[66,69] There is now evidence of sub-epidemics in South Australia; the predominant subtype B infection in the MSM population, non-B infections in IDU populations, heterosexual overseas-born, and travelling Australian-born people being infected with B and non-B infections overseas, and the emerging complex variants being imported from overseas. Despite this evidence, only one large-scale phylogenetic analysis has been reported in Australian literature; a surveillance study published in November 2015 that assessed the changing Western Australian HIV epidemic between 2000 and 2014, [250] including transmission dynamics within the population.

The current study sought to explore genetic relatedness of subtypes and CRFs, and genetic diversity as it relates to epidemiological surveillance, rather than trying to ascertain evolutionary relationships per se. Phy was conducted using maximum likelihood to construct

an unrooted tree and univariate analysis was used to assess demographic and epidemiologic factors associated with clustering.[163] This data demonstrates the benefit of linking phylogenetic and surveillance data in order to best inform HIV prevention efforts.[163]

6.12.1 Phylogenetic profile of the South Australian cohort

Phylogenetic analysis performed on the *pol* and *env* regions of this cohort revealed a geographical and genetic viral complexity in South Australia. The tree topologies for both regions were highly similar, with all cases that clustered together in one tree clustering together in the other tree. However as found by other researchers, branch lengths and bootstrap values differed between the trees;[162] 77% of all cases were part of either a *pol* or *env* high reliability pair or cluster but only 42% were part of both.

Over half the *pol* sequences formed part of high reliability groups, and 56% of *env* sequences. Just under half the *pol* groups were subtype B pairs and clusters, the largest comprised eight query sequences. The remainder were non-B pairs and clusters, the largest comprised 42 query sequences that also clustered with the subtype C, CRF07_BC and 08_BC reference sequences.

When broken down by subtype, all subtype C, D and 23_BG cases were part of a high reliability *pol* pairs or clusters, but while the small cluster of three 23_BG cases was due to a transmission event, both subtype D cases and the majority of *pol* C cases were not. Three subtype D sequences were also part of the same high reliability *env* cluster but again, not linked by transmission, which was supported by the epidemiological data. These findings suggest that the rate of nucleotide variation has been minimal within the subtype D lineage.

The most notable high reliability cluster was the *pol* subtype C group, all subtype C sequences clustered together with very high bootstrap value. One of the sequences sub-clustered directly with reference sequence 08_BC, while another sequence sat just outside of the 97% bootstrap subtype C cluster with 07_BC. Neither person had a region of birth

listed but both listed direct blood contact as a transmission risk. Given both sequences paired with the BC recombinants with bootstrap values of 100%, and these BC recombinants are almost exclusively found circulating in China[33,52] notably in the IDU population,[261] it is likely these viruses originated from being in direct contact with contaminated blood in China. It has been reported that subtype C originating from India is one of the parental strains of both CRF07_BC and 08_BC,[168] in the current *pol* tree the subtype C query and reference sequences originating from India did cluster with each other, but did not sit with the two BC recombinants on the tree. The majority of subtype C sequences in this study were carried by people originating from Africa, this was supported in the tree with the query sequences clustering with reference sequences originating from different countries in Africa. While the matching *env* sequences also clustered together with reference sequence C, the common ancestral node bootstrap value was 67% and only 11 of these 42 sequences were part of smaller high reliability subgroup pairs and one cluster. The genetic distance from the common node for the *pol* cluster ranged between 1.5% and 10%, while for the *env* cluster it genetic distance started around 6% and was over 25% for some sequences. For this particular cohort of subtype C infected individuals, there was greater variation within the *env* gene, while the *pol* region remained conserved over time. Differences between the two genes will be discussed in the next section.

In contrast, the biggest high reliability group identified by *env* was all the 01_AE cases, they clustered with a high bootstrap value (90%) to reference sequence 01_AE and had genetic distances from the ancestral node ranging from 7% to 13%. The matching *pol* sequences did cluster together but had a common bootstrap value less than 70%, and genetic distances from the common node ranging from 5% to 13%. Though all the 01_AE *env* sequences clustered together, only 17% were due to transmission events. CRF01_AE was discovered in the 1990's, when branches of what was then known as subtype E virus would

inconsistently cluster in phylogenetic trees depending on which genomic fragment was sequenced. The *env* region clustered as a distinct "E" group, but the gag and *pol* regions clustered with subtype A sequences. This led to the conclusion that subtype E was most likely an A/E recombinant. However it was also suggested that the discordant branching patterns between gene regions may be due to unequal evolutionary rates, where one gene region has evolved faster in and thus branch as a distinct group. Recent research by Ng and colleagues highlights this,[262] they conducted a genealogical analysis of CRF51_01B and discovered that the time to most recent common ancestor was around 2002 (CI=1999-2005) for the subtype B PR region, 2004 (CI=2002-2006) for the subtype 01_AE/B gp120 region, but 1996 (CI=1992-2001) for the 01_AE/B gp41 region. The authors postulated that the significantly older gp41 region may reflect at least two recombination events that involve CRF01_AE and multiple subtype B lineages with distinct evolutionary histories, or an as yet unidentified intermediate recombinant form for CRF51_01B that may have become extinct.[262] In the current study, the average genetic distance between sequences clustering with 01_AE was similar between the two trees, but the clustering pattern was different, including markedly different bootstrap values. The *pol* region of the CRF01_AE is subtype A, which is known to be highly variable.[263,264] The 01_AE cluster did sit underneath the subtype A reference sequences in the *pol* tree, but did not cluster directly with them. These findings suggest that the *pol* region may have a distinct evolutionary lineage compared to the *env* region.

The profile for subtype B cases differed to subtype C and 01_AE. Forty-one percent of subtype B sequences were part of high reliability pairs and clusters, the largest cluster comprised eight cases while just under two thirds were part of high reliability *env* groups, the largest was a cluster of 26 cases. Approximately 20% in both *pol* and *env* trees were due to transmission events. The *pol* sequences for this large cluster did not cluster together in

one large group, but rather in smaller groups, and only 13 of the sequences were in high similarity pairs or clusters. Subtype B linkages were most certainly impacted by missing sequences and it is highly likely some transmission links were not captured; just over one third of all new diagnoses were included from the first time period when the proportion of subtype B was highest, but nearly half the new diagnoses were included from the second time period when the proportion of non-B infections in the population was much higher. The difference between the *pol* and *env* regions for subtype B may be due to the length of time ART has been available in Australia where the majority of these subtype B cases were acquired. Australian treatment has been almost exclusively focused on PR and RT inhibitors; drug resistance mutations or natural evolution during times of increased viral load, such as after treatment failure or treatment interruption may impact on *pol* and leave *env* relatively conserved.

Almost all of the 02_AG *pol* sequences were part of one large high reliability cluster but only roughly one third were due to transmission events, while just over half of the 02_AG *env* sequences were part of a high reliability cluster, and 43% were due to transmission events. It is likely that the clustering patterns of 02_AG infections are impacted by the introduction of infections from different regions of Africa, with evidence of infection within the African-born and Australian-born population. The inconsistency between these particular sequences is likely due to a high diversity, with many recombinant events occurring among these strains which have resulted in the merging of several lineages into a single diverse group.[193] A study of Ghana sequences found the majority of 02_AG sequences were actually recombinant sequences made up of several genomic fragments of different 02_AG strains.[265] It has also been shown with phy of *env*-gp41 and *pol*-PR/RT regions that these strains do not have a strong phylogenetic relationship.[265] It is therefore possible that *pol* and *env* genes have evolved from different ancestral sequences and are undergoing

complex recombination processes which create an array of strains that complicate its use for subtyping. The current data support this view and other studies have found the PR region of 02_AG is unsuitable for phylogenetic analysis alone.

6.12.2 Transmission dynamics

Only 9% of the subtype B cases in high reliability *pol* clusters were due to transmission events, compared to significantly more non-B cases (20%). However the proportion due to transmission events in the *env* tree was very similar between subtype B (10%) and non-B (13%) cases. The reason for this discrepancy was due to the transmission criteria, in total, 43 cases were part of 20 possible transmission events but of these, eight events were identified only by *pol*, and five only by *env*. However the gene region which did not meet transmission criteria did cluster together with the same sequences, with bootstrap values 90% or higher, and genetic distances less than 5%.

Only counting cases identified by both *pol* and *env* using the transmission criteria would have led to non-identification of 13 transmission events, including a child pair, a family cluster, and subtype B sequences that crossed over MSM and heterosexual populations. Likewise, if only *pol* or only *env* had been used for this analysis, five and eight extra transmission events would not have been identified respectively. Transmission parameters will be discussed later.

The majority of transmission events were male/female pairs, followed by male pairs or clusters, a MTCT pair, a family cluster, and a child pair. Most of the male/female pairs carried non-B variants including intergenomic recombination, primarily transmitted via heterosexual contact overseas. There was evidence of non-B importation and forward transmission within Australia, including transmission within racial groups and between different racial groups, notably Australian-born people acquiring the virus overseas, and transmitting to Australian and non-Australian born people. One male/female pair had

sequences clustering with *pol* 48_01B and *env* 29_BF. Both were Australian-born but the infection was acquired overseas by one of the pair. The CRF48_01B was first described in Malaysia by Liu et al in 2010,[266] and has been typed as 01_AE/B at the PR/RT region of the *pol* gene. The *env* CRF29_BF sequences did not meet transmission criteria, with a genetic distance of 3.1%. CRF29_BF was first described in Brazil around 2006,[267] and carries subtype B in the *env* region. It is possible that phylogenetic analysis incorrectly classified the *env* region for this pair. The *pol*-PR/RT region for CRF48_01B is comprised of 01_AE and B, it is likely this pair carried a unique 01_AE/B variant most likely originating from Malaysia, with different subtype B lineages in the *pol* and *env* regions respectively.[92]

Only one of the male only pairs carried a non-B variant; subtype C. One male of unknown birth reported heterosexual contact overseas, and an Australian-born male reported MSM contact. The other male only non-B transmission occurred in the cluster of three, all the males were typed as *pol* 23_BG *pol* and *env* 02_AG. Phylogenetic analysis showed the sequences were highly similar, and demographic information showed diagnosis occurred one year apart. Two of the men reported heterosexual contact, both born in Africa, but only one was infected overseas and one male reported MSM contact locally.

The remaining male-only transmission events were subtype B cases, the majority were MSM and infected locally though as previously mentioned, one of the pairs was linked to a male/female subtype B pair. The subtype B cluster of three was similar to the subtype C cluster, all three were diagnosed one year apart from each other, two reported heterosexual contact (one overseas, one locally) while the other reported MSM contact in Australia. This discrepancy between the patient-reported transmission risk, and the transmission events identified by genetic analysis is indicative of broader issues surrounding fear of stigma, cultural barriers, and other factors that make it difficult for people to report information accurately.

There were two likely vertical transmissions, one involving a mother and child, the other a father, mother and child. The mother and child had subtype A virus, originating from Africa, though the child was reported as being infected in Australia. The family cluster carried subtype C, all originating from Africa and infected overseas. Though they were all diagnosed within one year of each other, the child's *env* sequence was 4.9% genetically distant to the female and 5.9% to the male. It is unknown why the *env* region was so variable across the three cases, but one explanation may be the genetic bottleneck of transmission theory, where only a subset of the viral population was passed from mother to child.[268]

Of special interest was a child transmission pair diagnosed one year apart and identified by *pol* only. Both regions were typed as 02_AG, with the children born in the same country in Central Asia. Further LANL BLAST analysis found the *pol* region was most similar to a 02_AG/A1 recombinant while *env* was most similar to 63_02A1, a CRF known to be circulating around Central Asia at the time the children were diagnosed. This will be discussed further in Chapter Seven. An analysis of these cases by Goldwater[269] in 2013 found the infections were likely to be iatrogenic. The children were not familial and it was likely they were infected with contaminated blood possibly from the same hospital. The infections were reported to be a result of the well-known issues of infection control in Central Asia, due to the political collapse and civil war that left devastating consequences on health systems and clean blood supplies for hospitals.[269] Given that both *pol* and *env* sequences clustered together with very high bootstrap values, and both had a genetic distance of 2% or less, these cases pose the question of whether they were infected with blood from the same infected person. An alternative proposition could be if the children do carry the relatively new CRF63_02A1 strain, it may still be quite conserved as it has so far only been found circulating near the border between Russia and Central Asia.

6.12.3 Nucleotide variability

Viral diversity differs between subtypes, but also between sites along the genome. Some sites are more functionally constrained while changes in other positions may be favored and this differs by genome region.[270] The *env* region provides rich information on circulating subtypes and CRFs within a geographic region, and products within the *env* gene are targets for vaccine development and fusion inhibitors.[122,201,271] The *env* region can be as virally diverse as 35% between subtypes, though most of the genetic variation occurs in gp120.[23] The C2-V3 fragment has historically been used for HIV-1 subtyping to identify ancestral forms and newer variants, however there is a high level of nucleotide changes within this region, with considerable heterogeneity within across the group M subtypes. Pieniazek and colleagues found there was enough nucleotide divergence within the gp41 region to use it for phylogenetic cluster analysis and subtype assignment, but enough conservative sites which to design a set PCR primers.[272] This has also been supported by other research in which 99% of HIV infected people were found to produce antibodies against the gp41 immunodominant regions in which there were very low levels of mutations,[200] and the region has been found to remain conserved even during long-term treatment with a fusion inhibitor.[201]

In this study, there was a high nucleotide variability among the query sequences (excluding reference sequences) for both the *pol*-PR/RT and *env*-gp41 regions; 49% (540/1098) of the *pol* sites were variable and 59% (324/545) of the *env* sites. This high rate of variability was unsurprising given the large variety of B and non-B subtypes and CRFs in the dataset, including potential unique recombinant forms. Further studies assessing the nucleotide variability within subtype clusters would be beneficial to see if there are significant subtype-specific differences between the two genes.

The *env* region in general has the highest level of diversity along the genome, and this is thought to impact on the *env* structure between subtypes. Gao et al found subtype-specific

differences impacted on genetic, phylogenetic, and biological properties and subtype B strains differed from non-B in terms of changes in the number and distribution of cysteine residues, substantial length differences in hypervariable regions, and premature truncations in the gp41 domain.[271] The V3 region of gp120 has been found to differ between subtype B and subtype C virus, with subtype-specific epitopes that impact antibody cross-reactivity.[273] Contrasting evidence exists for the *env* gp41 region, minimal variation was found in the HR-1 region of gp41 between subtypes in a study by Xu et al,[274] but a study by Carmona et al[275] found a significantly higher number of naturally occurring *pol*ymorphisms to enfuvirtide in non-B subtypes compared with subtype B.

During HIV infection, it is believed that much of the *env* evolution is due to targeted cell-mediated and humoral immune responses.[53,276]. A study by Choisy et al[277] found subtypes A, B, C and D experienced positive selection pressure at similar positions in *env*, but the magnitude of selection was statistically different between subtype B and the other subtypes, reflecting subtype-specific adaptive evolution.[190] Another study by Travers et al[55] found between the different group M subtypes there was a heterogeneity of selective pressure; for subtypes C, F1 and G there were specific observed sites that had undergone positive selection while the same sites in other subtypes had undergone purifying selection.[278] Subtypes A and K both had sites that had undergone purifying selection while at the same sites positive selection had occurred for other subtypes. These sites were located throughout the gp160 region, including gp120 and gp41. The findings from the current study support previous research that indicate the difference in evolutionary mechanisms between subtypes may have an impact on viral fitness for each subtype and adaptive pressures shape each lineage differently.[190] In addition to specific host responses, *env* differences between subtypes could also be influenced by geographic region the infection is circulating in, route of infection and epidemic patterns.[190]

It is very important for future vaccine and treatment design that a comprehensive understanding of why and how there are subtype-specific differences at the *env* region is obtained.[64,190]

6.12.4 Univariate analysis

There has been a decreasing use of *p* values by statisticians over the last 30 years, with an increasing understanding that exploratory data analysis should retain as much raw data as possible.[179] Analysis of DNA sequences is complex and multidimensional, to add to that by performing multivariate analysis which introduces demographic information may lead to the loss of valuable meaning of the original data. For this reason only simple univariate analysis was performed.

There was a significantly higher than expected number of non-B *pol* sequences that were part of high reliability clusters and transmission clusters, and more cases diagnosed between 2007-2012. These findings are most likely related; there is more heterogeneity among subtype B strains in Australia due to longer circulation and more frequent infection rates, therefore the propensity to mutate becomes greater, amplified by long-term exposure to treatment and subsequent treatment failures. A large proportion of subtype B sequences were not included in the subset examined which impacted the representativeness of the subtype B cohort, 35% of all newly diagnosed people were included from the first time period (where subtype B predominated), but 48% were included from the second time period (where the proportion of non-B infections in the population was much higher).

Females were significantly more likely to be part of *pol* high reliability and transmission clusters than males but there was no significant association for *env*. Again, this is most likely due to the higher proportion of non-B sequences forming part *pol* clusters, among which is a greater proportion of females, compared to subtype B infections which are predominantly male. The observed frequency of both Australian-acquired infections, and Australian-born

people in an *env* high reliability cluster was greater than expected. It is unclear whether this finding is meaningful, though there were no significant associations for the *pol* region. This is most likely related to the fact that there were no significant differences between B and non-B viruses in *env* clusters, due to the 40+ cases with subtype C virus not being part of high reliability *env* clusters.

In a very recent study, Lubelchek and colleagues[163] used a representative dataset of one third of all new diagnoses between 2008 and 2011, and found the subset mirrored the general population generally, though they did concede a higher proportion of racial and ethnic minorities which is also true in the present study. Ultimately, sampling of the entire diagnosed and subtyped cohort in South Australia may have biased the *pol* analysis toward identification of people infected overseas in high prevalence countries,[279,280] or within Australia for the *env* analysis. However the overrepresentation of the former if any, may be justified given this subpopulation is at a disadvantage in Australia in regards to prevention, treatment and other HIV-related disparities.[163]

6.12.5 Strengths and Limitations

As sequence analysis is increasingly being used in clinical practice and to assess HIV epidemic structures and transmission dynamics, combined with a growing genetic diversity, great care must be taken when choosing a phylogenetic method.[270] Branch lengths (genetic distance) and bootstrap (reliability) values can both be impacted by the choice of model which can lead to unreliable trees and unrealistic evolutionary relationships. If sequence data are biased, such as when sequences share an unusually high GC content, this will impact bootstrap estimates, the sequences may artificially cluster together and will be supported by a high bootstrap value. Sequences with an increased evolutionary rate may also artificially group together. If a substitution model is chosen that is too simple for the complex dataset, distant sequences may cluster together or be drawn toward the root of the tree. The bootstrap

trees are then inferred on the basis of this incorrect evolutionary model, which leads to 'long branch attraction' (the artificial clustering of long branches together with a high bootstrap value). For this study, the GTR method with *gamma* distributed and invariant sites was assessed as the best model to ensure reliable phylogenetic inference from the dataset. This model is the most robust when accounting for bias in nucleotide composition and preferred substitutions and high rates of mutation, especially for *env* fragments.[270] In order to ensure the relative stability of pairs and clusters within a phylogenetic tree, it is recommended that between 200 and 2000 bootstrap iterations of the dataset are conducted.[281] In the present study 1000 bootstrap iterations were chosen based upon previous studies of high quality.[112,262,282–284] A bootstrap value of 70% or over at the common node was chosen as branches or groups supported by these values are considered to be reliably placed on the tree.[281] For both the *pol* and *env* tree there were no artificial artefacts or long branch attraction noted, and the epidemiologic data supported correct clustering.[281]

Phylogeny is the most attractive and robust approach for assessing HIV-1 evolution and transmission, not only for surveillance purposes but also for clinical and legal reasons. Whole genome sequencing is the gold standard, however there are still relatively few full-length sequences available for reference comparison, and limitations such as financial cost, access to new sequencing techniques, and computational power, are still common barriers, especially in LMICs. As has been mentioned, the region spanning PR and RT of the *pol* gene is routinely used for genotypic drug resistance testing and is therefore the most commonly used for phy to complement epidemiological information including contact tracing, and to assess genetic similarity or diversity within and between subtypes.[162] The *pol* gene is also the most conserved of all the genes with the lowest rate of nucleotide substitution, while the *env* gene is highly variable and known to undergo distinctive evolutionary dynamics.[64,162,285] There has been conjecture about whether *pol* is too genetically conserved and therefore

suboptimal by some for transmission analysis,[162] however Hué and colleagues reported both the *pol* and *env* regions are robust for phy,[162] finding identical clustering patterns during analysis of both regions, with similar bootstrap values. They did observe differences in branch length, and clustering patterns of unrelated sequences and concluded the *pol* gene carries adequate genetic variability for phy of HIV infection, though characterisation of phylogenetic relationships may also be confirmed with other more variable gene regions such as *env.* In the current study, phy of both these regions was used to assess the relatedness of viruses.[162] Both regions were found to have high proportions of variable sites, but the *pol* and *env* tree topologies were consistent, with the same groups identified. As the other authors found, bootstrap values and branch lengths for some groups differed between the genes.[162] Basing inferred linkage on a single genetic region may be undermined by recombination events, mutations and genetic variability of the region. Using two genes in this study increased identification of high reliability and transmission pairs/clusters compared to using one gene region alone.

The analysis included sequential sequences to asses within individual evolution and conduct quality assurance of the trees. The majority of the cases met the criteria for transmission by *pol* or *env*, and all cases had sequential sequences sitting within the same cluster. Of those where there was a genetic distance greater than 1.5% and/or a bootstrap value less than 98%, the majority carried subtype B acquired in Australia. It may be that the diversity between sequential sequences was due to treatment failure and resulting drug resistance mutations, though other studies have demonstrated that even when known drug resistant sites are excluded from analysis the phylogenetic tree is not altered.[162,250] It is more likely that the genetic distance is due to the natural evolution and error prone nature of HIV during replication and transcription during periods of time when a person is experiencing treatment failure, switching treatment, or having a treatment interruption, which is usually

when the viral load increases dramatically.[44]

This then raises the question of whether the criteria for transmission clusters was too conservative. If phy has limitations associated with viral divergence then what are the optimum parameters to define transmission clusters, by both subtype, and gene region? Bootstrap values of 98% or higher and genetic distance of 1.5% or less are often used for transmission identification,[44,241,250,283] to ensure unrelated sequences are not incorrectly identified as being related by transmission. This value has been chosen as the estimated distance between unrelated sequences is ≥5%.[44,241,250,283] However, the findings from the 12 cases with sequential sequences, and the transmission pairs and clusters identified by only *pol*, or only *env*, suggest that these parameters are too conservative. In addition, though a strength of this study was that it spanned a long period of time (2000-2013) which allowed a comprehensive overview of the HIV-1 genomic diversity and transmission dynamics, only one sequence was included for each case, from a sample within 12 months of diagnosis. This combined with the very conservative transmission criteria is likely to have led to only direct infections or infections occurring within a short space of time being identified.[262] It is highly likely some transmission links were not captured, for example, person A diagnosed in 2000 may have infected person B in 2010 after failing treatment, and person B was then diagnosed in 2011. Person A's 2000 sequence and person B's 2011 sequence may not meet transmission criteria due to natural evolution of HIV over time, super-infection and recombination. Other transmission linkages may not have been identified due to an intermediate HIV infected person that was infected by person A and then infected person B, but who did not have sequence included in the dataset. This is especially likely for subtype B sequences in which half of the known diagnoses were not included.

A strength of this study was the access to baseline HIV-1 *pol* sequencing data and original samples in which to sequence the *env* region. Though a number of the older baseline

samples were no longer in archive storage or were too degraded, nearly half of all new diagnoses between 2000 and 2012 were included in the study including a relatively representative sample of B and non-B infections. The 2013 samples could not be accessed for research due to ongoing use for clinical reasons at the time the laboratory work was conducted. The relatively large size of the dataset and the diverse nature of the population included allowed the data to remain meaningful. The majority of *env* sequences were extracted from plasma samples used to sequence the *pol* region, or a sample taken within 12 months of it, this reduced the likelihood of sequencing a different strain from the within host viral population. A small number of *env* sequences may have been taken from samples after a person had begun treatment and drug resistance mutations may have been present. However previous studies have shown that drug resistance mutations do not impact the phylogenetic signal as to impede subtype assignment and reconstruction of transmission history.[162,194,286]

In order to create the dataset for ML, a very limited amount of data processing was conducted. An initial NJ tree analysis identified reference sequences that did not cluster with or near any query sequences and these were excluded from the dataset in order to minimise unnecessary nucleotide variability. No query sequences were excluded, including those that were not genetically close with other query sequences, and did not have a reliable bootstrap value (≥70%).[162] Limiting the pre-processing of data has been shown to prevent unnecessary impact on the results such as bias toward favouring the likelihood of clusters within trees. There is a possibility that having complex recombinant sequences in the dataset may have interfered with the overall tree structure, where sequences may have clustered together artificially (long branch attraction). As referred to earlier in this section, the GTR method with *gamma* distributed and invariant sites was chosen to counteract any possible bias including highly mutated fragments.The epidemiological data was also found to be

congruent with the genetic data, indicating a robust tree, However caution should be taken when interpreting both trees due to the complex recombinant inclusions, and further studies that remove these sequences and reassess the tree are warranted.[162]

Another strength of the study was the almost complete demographic and epidemiological dataset, including route of transmission, region of birth, location of infection acquisition, year of diagnosis and sex. This enabled univariate cluster analyses to be conducted, though for some of sub-analyses such as transmission clusters by region of birth, it is noted that there were very small numbers. Linking genetic and epidemiological data enabled detection of information not available via normal surveillance, including epidemiological shifts of potential significance. Pairs and clusters were identified that would not have been possible using either genetic or epidemiological data alone. An example of this was a MSM pair and heterosexual pair with subtype B which were likely carrying a related virus. There was also evidence of non-B variants being transmitted from Australian-born people infected overseas, to overseas-born people within Australia, and new complex non-B variants being transmitted within Australia.

6.12.6 Recommendations

Recommendations from these findings include specific HIV prevention strategies for HIV subpopulations within the South Australia:

1. Overseas-born and Australian-born MSM, including increased education and support for men who may not openly report having sex with other men. This is especially important for overseas-born men who are at increased risk of suffering from multiple stigmas, and may have cultural reasons for not disclosing their sexual orientation.
2. People who inject drugs and/or engage in high-risk sexual behaviour in both the MSM and heterosexual population.

3. Heterosexual people infected overseas, or infected within Australia, with a non-B variant.

Some researchers do not include pairs in phylogenetic analysis. Pairs were included in this study to assess the proportion of infections that may be partners entering care together (being diagnosed together) or whether pairs were due to risky behaviour. Though it is less likely pairs represent sources of continuous ongoing transmission it is possible that pairs are part of larger existing clusters and these infections may continue to be transmitted in the wider population.[163,165,251] Non-B transmission cases were almost all male/female pairs and most pairs were diagnosed within a short space of time from each other. This suggests a reduced risk of forward transmission into the wider community than subtype B infections; some of these pairings are likely long-term partners. In contrast, Subtype B was more reflective of a broader epidemic occurring in South Australia. There were subtype B pairs and clusters closely linked to other subtype B pairs and clusters, spanning female and male heterosexual, MSM, and IDU transmissions.[194] Though these separate pairs and clusters did not meet transmission criteria with one another, it is highly likely that the viruses within these pairs and clusters are related, suggesting the need for careful surveillance and targeted educated and prevention strategies.

Despite research that supports a very restricted criteria for identification of transmission events,[162,163,165,248,253,287–290] findings from the present study strongly suggest the parameters are too conservative, and more research on parameters by subtype and gene region should be conducted. While the criteria most certainly excluded unrelated infections it also likely excluded related cases from clusters. It is a difficult balance to ascertain correct parameters, the rationale behind the use of phylogeny is to inform public health strategies that target subpopulations with active HIV transmission to try and prevent new infections. This in itself

justifies a tight definition. However criteria that is too tight may also overlook transmission dynamics that could greatly impact on public health and clinician resources. For instance, a cluster of people diagnosed across a large span of time may have all been infected by the same person who has not yet been diagnosed. Identifying this cluster by broadening the transmission criteria could lead to identification, diagnosis and treatment of undiagnosed and highly infectious people in the community. It could also lead to a targeted public health message about the acute and chronic stages of HIV, treatment failure, and the huge risk of infectiousness from all of these stages. Careful and considered reporting of combined phylogenetic and epidemiological information is crucial, given the reality of stigma and isolation many HIV positive people face. Focus should be solely on the epidemiological characteristics of the virus rather than the individual themselves. This will increase the likelihood of testing, disclosing status, accessing treatment and reporting potential exposures.

Further analysis using a complete dataset of all HIV infected individuals in South Australia, including multiple sequences for each person may identify more, and larger clusters. The *pol* region is routinely sequenced and is a good place to initiate this breadth of analysis. However full genome sequencing is highly recommended in order to accurately characterize HIV subtypes and CRFs, identify transmission dynamics and gain a robust understanding of the epidemic structure.

Finally, sequences analysed from cases of overseas acquired infection should be freely accessible globally, to be used as reference sequences. This is especially important for the regions in which the virus originated from, and also other countries where the same variants may have recently begun circulating.

These findings warrant ongoing research and attention from clinicians, scientists and policy makers not only in South Australia, but Australia wide, and from the countries of

origin. The data show that HIV transmission is not bound by geography and while there were a high number of imported infections, there were also a number of local non-B virus transmissions originating from overseas, and subtype B infection originating from Australia transmitted into the overseas-born population. It is currently not required to obtain patient consent to conduct phy using routinely generated patient sequences in South Australia, though this project underwent an extremely stringent ethics process spanning six months. As sequencing techniques continue to be refined, and the use of phy increases, it is strongly advised that a genotypic resistance profile be conducted at time of diagnosis, and informed patient consent is obtained about the use of sequencing data and potential linkage to epidemiological data. This will require education of clinicians, laboratory scientists, and a review of current policy.[162] Given that sequencing technologies are constantly improving, and full genome reference sequences are increasing all the time, it is recommended that full genome phy be conducted in the future for a more complete characterization of the epidemiological associations and transmission patterns.[194]

6.12.7 Conclusion

This study is the first in Australia to link genetic and routine surveillance data including sex, route of infection, location of acquisition, region of birth and age, to provide an in-depth characterisation of the changing HIV epidemic in South Australia. The findings demonstrate that analysis of relatively large datasets of *pol* and *env* sequences using modern phylogenetic methods is quick and robust,[164] and provides valuable information about the HIV epidemic structure in the region of interest. These results are an important resource for targeted public health initiatives,[164] and should be incorporated into current surveillance strategies in Australia.

Over time the traditional geographical and risk-group segregation of HIV subtypes and CRFs are becoming less distinct, with evidence of crossover between MSM, IDU and

heterosexual populations. The impact of a growing number of non-B infections into a) a historically subtype B restricted region and b) into the heterosexual community, on clinical outcome or vaccine efforts is unknown. This dearth of knowledge in Australia combined with the increasing genetic diversity including complex and unique variants, highlights the need for careful monitoring of new infections, and the immediate need for non-B subtype research in Australia.

These findings are similar to other reports globally.[164,291,292] Cluster analysis combined with routine surveillance information identified subpopulations which is crucial in order to create targeted subgroup validated evidence-based interventions and preventions that promote a safe and stigma free environment, and to identify infections early for immediate treatment.[163] Prevention and intervention strategies need to be designed and implemented at multiple levels – individual, partner, group, community and structural, in order to successfully halt the increase of new diagnoses in the population, and ensure adequate and sustained clinical care is offered to every HIV infected person.[109,293,294]

CHAPTER 7: SUBTYPE AND RECOMBINATION ANALYSIS USING ONLINE TOOLS

7.1 Overview

Phylogenetic analyses are not widely used to determine subtype in clinical settings because of time constraints and the complexity of determining recombinant strains.[182,192] Subtyping is most often interpreted by the *pol* sequences obtained following routine drug resistance testing, using rapid online subtyping tools. In South Australia *pol* sequences are submitted to the Stanford CPR tool for resistance testing and subtyping.

This study evaluated the reliability of six rapid online subtyping tools in identifying pure subtypes, CRFs, ISRs, InSRs and URFs from a subset of newly diagnosed infections between 2000 and 2012. Both *pol*-PR/RT and *env*-gp41 sequences were used to determine concordance between online tools when assessing different regions of the genome and to ascertain the prevalence of genomic recombination. Sequences were subtyped using phy and six online tools (jpHMM, REGA, Stanford CPR, SCUEAL, LANL BLAST and COMET). Five of the six online tools provided percentage support values for the assigned subtype: model averaged support (SCUEAL), % similarity to reference sequence (LANL BLAST and Stanford CPR), and bootstrap support (REGA and jpHMM). Subtypes with a support value of 70% or over were defined as reliably assigned. The results from each online tool were compared with phy to assess the degree of similarity.

The online tool parameters as reported in Chapter Three were used to identify **inferred subtypes** and evaluate the proportion of pure subtypes, CRFs and possible unique intersubtype recombinants (ISRs). In brief, all six online tools were used to assess *pol* sequences and if five out of six tools assigned the same subtype then the *pol* sequence was assigned as that subtype, referred to as the **inferred subtype**.

Four online tools were used to assess *env* sequences (jpHMM, REGA, LANL BLAST and COMET) and if three out of four tools assigned the same subtype then the *env* sequence was assigned as that subtype **(inferred subtype**).

SCUEAL and Stanford CPR do not have the capability to subtype the *env* gene. REGA cannot perform recombination analyses on short fragments (≤800bp), so these were only assigned pure subtypes by this tool and while jpHMM is capable of pure subtype and recombinant analysis for short fragments but may incorrectly classify short fragments located near the 3' end.

7.2 Subtype distribution – similarity between phy and online tools

The first part of the study was to report the degree of similarity between MEGA phy and online subtyping tools in order to assess how accurate phylogeny is in comparison with the more often used rapid subtyping tools. HIV-1 subtypes defined by phy were compared with those provided by the online subtyping tools (six tools for *pol* and four tools for *env*). The number and proportion of subtypes/CRF by each tool or phy is shown in Tables 34 (*pol*) and 35 (*env*). Full details of each case can be found in Appendix One.

There were five CRFs on the *pol* phylogenetic tree that sequences clustered with but no online tools identified (23_BG, 47_BF, 48_01B, 51_01B and 52_01B), and seven CRFs on the *env* phylogenetic tree (03_AB, 14_BG, 28_BF, 29_BF, 36_cpx, 45_cpx, and 47_cpx).

Conversely, the online tools identified some sequences as A2, G and 19_cpx but these sequences did not cluster with those reference types on the *pol* phylogenetic tree. Online tools also identified some *env* sequences as A2, A3, G, J and 63_02A1 but these sequences did not cluster with those reference types on the *env* phylogenetic tree.

Online tools also identified a number of sequences as *pol* or *env* ISRs but MEGA phy did not have the capability to identify recombinants that were not similar to known CRF reference sequences included in the tree.

Figure 26 shows the proportion of subtypes, CRFs and ISRs for *pol* phy and each of the online tools. The proportions classified as subtypes A1, B, C, D, and CRF01_AE were quite

similar between all the online tools and phy. The highest proportion of identified ISRs was by jpHMM and SCUEAL.

Table 34. Number and proportion of *pol* sequences by subtype.

	MEGA phy N (%)	jpHMM N (%)	REGA N (%)	Stanford CPR N (%)	SCUEAL N (%)	COMET N (%)	LANL BLAST N (%)
01_AE	36 (16)	31 (14)	35 (16)	38 (17)	28 (13)	33 (15)	38 (17)
02_AG	12 (5)		11 (5)	12 (5)		12 (5)	9 (4)
07_BC	1 (0.5)		1 (0.5)			1 (0.5)	1 (0.5)
08_BC	1 (0.5)		1 (0.5)			1 (0.5)	1 (0.5)
15_01B	2 (1)				6 (3)	2 (1)	2 (1)
19_cpx					1 (0.5)		
23_BG	3 (1)						
47_BF	1 (0.5)						
48_01B	7 (3)						
51_01B	4 (2)						
52_01B	1 (0.5)						
A1	10 (5)	16 (7)	10 (5)	3 (1)	9 (4)	12 (5)	8 (4)
A1-like			2 (1)				
A2					1 (0.5)		
B	100 (45)	117 (53)	108 (49)	111 (50)	108 (49)	108 (49)	115 (52)
B-like			3 (1)				
C	41 (19)	38 (17)	41 (19)	42 (19)	41 (18)	41 (19)	40 (18)
D	2 (1)	1 (0.5)	2 (1)	2 (1)	1 (0.5)	2 (1)	2 (1)
G		1 (0.5)	1 (0.5)		2 (1)		
ISR		17 (8)	6 (3)	13 (6)	24 (11)	9 (4)	5 (2)
Total	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**

Table 35. Number and proportion of *env* sequences by subtype.

	MEGA phy N (%)	**jpHMM N (%)**	**REGA N (%)**	**COMET N (%)**	**LANL BLAST N (%)**
01_AE	36 (16)	33 (15)		36 (16)	36 (16)
02_AG	12 (5)			2 (1)	11 (5)
03_AB	1 (0.5)				
07_BC	1 (0.5)			1 (0.5)	1 (0.5)
08_BC	1 (0.5)			1 (0.5)	1 (0.5)
14_BG	1 (0.5)				
15_01B	2 (1)			3 (1)	
28_BF	1 (0.5)				
29_BF	2 (1)				
36_cpx	1 (0.5)				
45_cpx	1 (0.5)				
47_BF	3 (1.4)				
63_01A1					2 (1)
A1	12 (5)	24 (11)	39 (18)	9 (4)	7 (3)
A2					2 (1)
A3					2 (1)
B	104 (47)	107 (48)	114 (52)	111 (50)	103 (47)
C	40 (18)	38 (17)	53 (24)	39 (18)	38 (17)
D	3 (1)	2 (1)	3 (1)	3 (1)	2 (1)
G			11 (5)		
J			1 (0.5)		
ISR		3 (1)		16 (7)	6 (3)
N/A		14 (6)			10 (5)
Total	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**	**221 (100)**

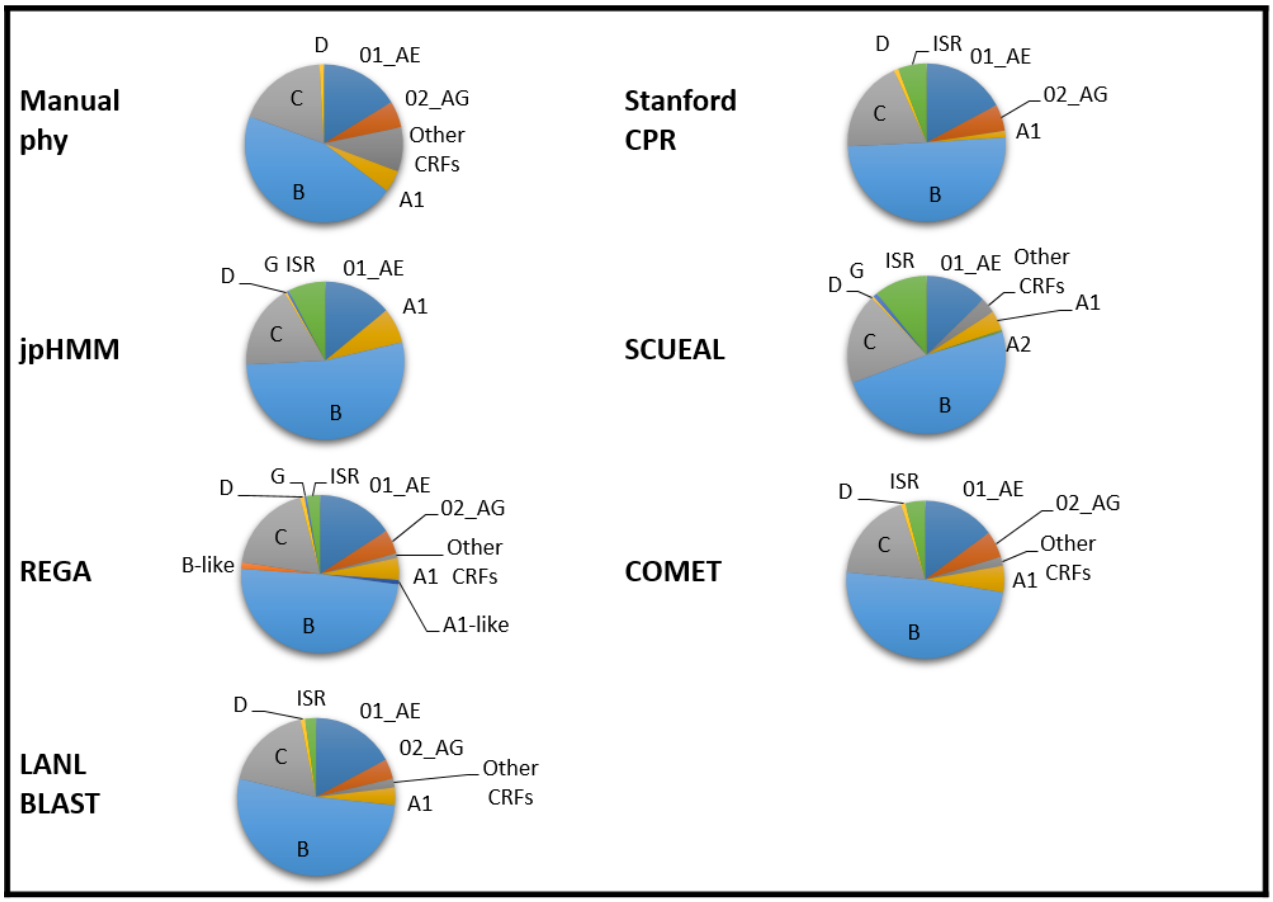


Figure 26. Proportion of pure subtypes/CRFs and ISRs assigned for the *pol* gene by each online tool and MEGA phy.

Figure 27 below shows the proportion of subtypes, CRFs and ISRs for *env* phy and each of the online tools. The proportions classified as subtypes B, C and D were quite similar between all the online tools and MEGA phy. COMET identified the highest proportion of ISRs, REGA was unable to identify any CRFs so the proportion of pure subtypes is larger compared with the other online tools and phy.

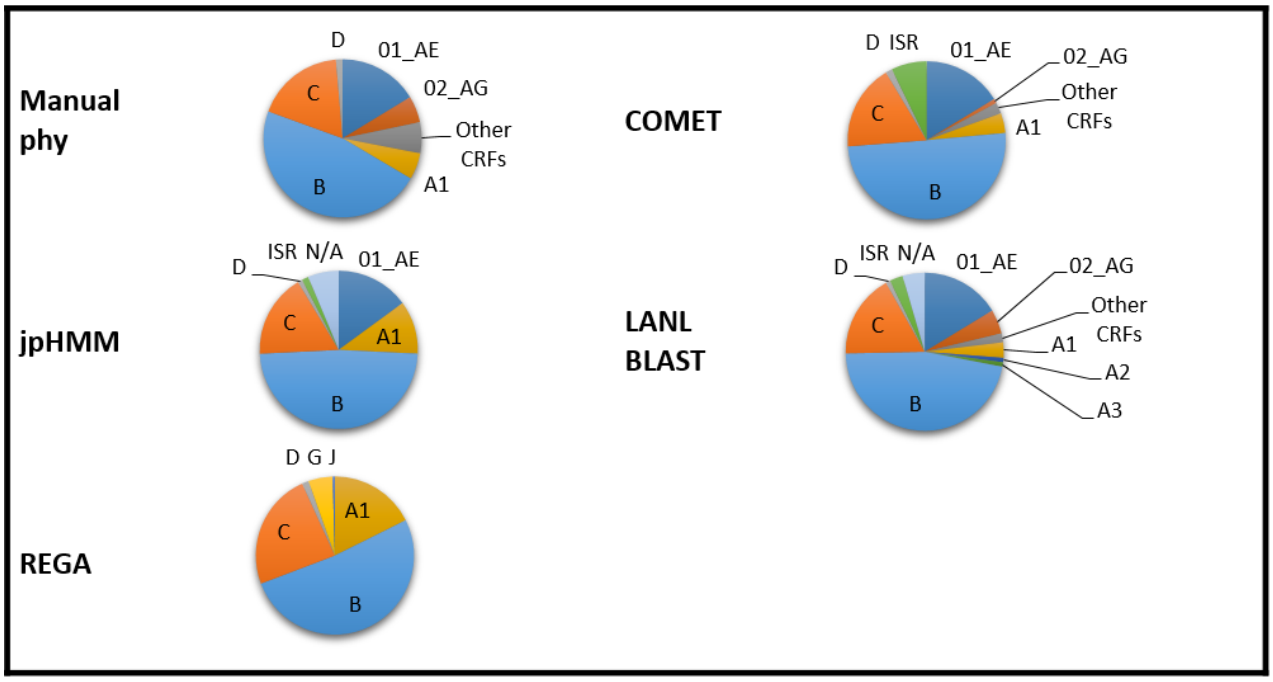


Figure 27. Proportion of pure subtypes/CRFs and ISRs assigned for the *env* gene by each online tool and phy.

Tables 36 and 37 below show the level of concordance between the phylogenetic tree and each of the online subtyping tools for the *pol* and *env* sequences. The online tools with the highest concordance to the *pol* phylogenetic tree were COMET and REGA (both 54%) followed by BLAST (52%), Stanford CPR (37%), SCUEAL (30%) and jpHMM (28%). The last two also assigned the highest number of unique ISRs during analysis, which partly accounts for the lower concordance.

The online tools with the highest concordance to the *env* phylogenetic tree were COMET and BLAST (both 42%) followed by jpHMM (28%), and REGA (25%). It must be noted that only jpHMM and COMET had the capability to recognize CRFs or ISRs for the *env* region and jpHMM could not classify 14 of the 221 sequences.

Although the online tools did not always identify the exact CRF that an individual sequence clustered with on the phylogenetic tree (which is what the percentage is based upon in Tables 36 and 37), they did often identify the same subtype or recombination variant for

that particular section of the sequence (see Key under each Table). For example, for one sequence clustered with 07_BC on the *pol* phylogenetic tree, jpHMM did not identify 07_BC but it did identify a B/C variant.

Another example is the four sequences that clustered with 51_01B on the *pol* phylogenetic tree; none of the online tools identified 51_01B, but they all identified subtype B, which is the correct subtype for that section of *pol* 51_01B.

The concordance between phy and the online tools for all of the pure subtypes, and for CRFs 01_AE and 02_AG was quite high, Tables 36 and 37.

Table 36. Concordance between *pol* phy and online subtyping tools.

pol **subtype defined by phy**	**N**	**Concordance with phy (%)**					
		jpHMM	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**
01_AE	36	75%	100%	89%	75%	81%	92%
02_AG	12	0%	92%	100%	0%	100%	75%
07_BC	1	0%	100%	0%	0%	100%	100%
08_BC	1	0%	100%	0%	0%	100%	100%
15_01B	2	0%	0%	0%	50%	0%	0%
23_BG	3	0%	0%	0%	0%	0%	0%
47_BF	1	0%	0%	0%	0%	0%	0%
48_01B	7	0%	0%	0%	0%	0%	0%
51_01B	4	0%	0%	0%	0%	0%	0%
52_01B	1	0%	0%	0%	0%	0%	0%
A1	10	80%	70%	30%	50%	80%	70%
B	100	100%	98%	99%	97%	96%	100%
C	41	93%	100%	100%	100%	100%	98%
D	2	50%	100%	100%	50%	100%	100%
Total	**221**	**28%**	**54%**	**37%**	**30%**	**54%**	**52%**

Key: ***phy 02_AG*** (SCUEAL assigned two sequences as subtype G, and the remaining 10 sequences as ISRs, REGA assigned one sequence as subtype G. jpHMM assigned six sequences as A1/G variants, one sequence each as subtype A, B or G, one as a B/G variant and one as a A1/B variant. LANL BLAST assigned one sequence as 02_AG/A1 and one as 02_AG/U), ***phy 07_BC and 08_BC*** (the online tools that did not assign 07_BC/08_BC did assign a B/C variant), ***phy 15_01B*** (SCUEAL assigned one sequence 15_01B and the other sequence an ISR. All other tools assigned subtype B. This section of the 15_01B reference sequence is 01_AE), ***phy 23_BG*** (all tools assigned a B/G or B/02_AG variant except LANL BLAST which assigned subtype B, this section of 23_BG is a mixture of B and G), ***phy 47_BF*** (all tools assigned subtype B except SCUEAL which assigned an B/29_BF variant, this section of 47_BF is a mixture of B and F), ***phy 48_01B*** (all tools assigned subtype B except REGA which assigned one sequence as B-like and one as B/F1, this section of 48_01B is a combination of 01_AE and B), ***phy 51_01B*** (all tools assigned subtype B. This region of 51_01B is entirely subtype B), ***phy 52_01B*** (all tools assigned A1/B or 01_AE/B variants. This region of 52_01B is a combination of 01_AE and B).

Table 37. Concordance between *env* phy and online subtyping tools.

***env* subtype defined by phy**	**N**	**Concordance with phy (%)**			
		jpHMM	**REGA**	**COMET**	**BLAST**
01_AE	36	92%	0%	100%	100%
02_AG	12	0%	0%	17%	83%
03_AB	1	0%	0%	0%	0%
07_BC	1	0%	0%	100%	100%
08_BC	1	0%	0%	100%	100%
14_BG	1	0%	0%	0%	0%
15_01B	2	0%	0%	0%	0%
28_BF	1	0%	0%	0%	0%
29_BF	2	0%	0%	0%	0%
36_cpx	1	0%	0%	0%	0%
45_cpx	1	0%	0%	0%	0%
47_cpx	3	0%	0%	0%	0%
A1	12	92%	100%	75%	42%
B	104	100%	100%	88%	90%
C	40	90%	100%	98%	95%
D	3	67%	100%	100%	67%
Total	**221**	**28%**	**25%**	**42%**	**42%**

Key: ***phy 02_AG*** (COMET assigned two sequences as 02_AG and remaining 10 sequences as 'check for 02_AG, jpHMM assigned all but one sequence as A1, and one as A1/B. LANL Blast assigned two as A1), ***phy 03_AB*** (all tools assigned the sequence as subtype B. This region of 03_AB is entirely subtype B), ***phy 07_BC and 08_BC*** (COMET and LANL BLAST both assigned the same CRF as MEGA phy), ***phy 14_BG*** (all tools assigned subtype B except LANL BLAST; unassigned. This region of 14_BG is predominantly subtype B with G toward the 3' end of the fragment), ***phy 15_01B*** (all tools assigned subtype B. This region of 15_01B is entirely subtype B except a small fragment of 01_AE at the 3' end), ***phy 28_BF*** (all tools assigned subtype B, the entire *env* region is subtype B for 28_BF), ***phy 29_BF*** (all tools assigned subtype B except COMET, assigned one sequence as 15_01B. The entire *env* region is subtype B for 29_BF), **phy 36_cpx** (jpHMM and REGA assigned A1, COMET assigned 'A1/check for AG' and LANL BLAST assigned 02_AG. This region of 36_cpx is a combination of A and 02_AG), ***phy 45_cpx*** (jpHMM and REGA assigned A1, COMET assigned 'A1/check for AG' and LANL BLAST assigned 02_AG/A1. This region of 45_cpx is a combination of A and 02_AG), ***phy 47_BF*** (all tools assigned subtype B. This region of 47_BF is entirely subtype B).

7.3 Online tool parameters – inferred subtype

The next part of the study was to report an inferred subtype based on the chosen parameters, in order to compare the difference in results with the current protocol for subtyping in South Australia (Stanford CPR tool).

7.3.1 pol – concordance between all six online tools

There was concordance between all six online tools for 161 (73%) *pol* sequences with a subtype distribution as follows: 100 subtype B sequences, 37 subtype C sequences, 21 01_AE sequences, 2 A1 sequences, and 1 subtype D sequence.

7.3.2 pol inferred subtype – concordance using online tool parameters (five of the six tools were concordant)

When assessing subtype by the set online tool parameters (five of the six tools), 90% (198) of cases met this criterion with a subtype distribution as follows: 111 subtype B, 41 subtype C, 35 01_AE, 9 subtype A1, and 2 subtype D. The next section details the number of inferred cases for each subtype or CRF.

7.3.2.1 Subtype A1

For the nine inferred subtype A1 *pol* sequences, Stanford CPR identified five as A/01_AE ISRs, and another one as an unknown PR/01_AE RT sequence. SCUEAL identified a sequence as an A1/A-ancestral/A3 unique ISR, and LANL BLAST identified three sequences as 01_AE (MEGA phy also identified the same three sequences as 01_AE.

7.3.2.2 Subtype B

Of the 111 inferred subtype B sequences, all clustered with or around subtype B on the phylogenetic tree, but 12 sub-clustered with B-like CRFs (47_BF, 48_01B or 51_01B).

REGA assigned one of the phy 48_01B sequences as B/F1, and one phy 48_01B

sequence as ‘B-like’. SCUEAL assigned the phy 47_BF sequence as a unique ISR; U/B/29_BF.

7.3.2.3 Subtype C

Of the 41 cases assigned *pol* subtype C by at least five of the online tools, all clustered with subtype C on the phylogenetic tree. jpHMM identified two sequences as subtype B and one as a B/C ISR, and LANL BLAST identified a fourth sequence as being most similar to a A1/C ISR.

7.3.2.4 Subtype D

Both subtype D sequences were identified as subtype D by all the online tools except SCUEAL, which identified one sequence as 19_cpx (the PR/RT *pol* fragment of 19_cpx is subtype D). These sequences also clustered with subtype D on the phylogenetic tree.

7.3.2.5 CRF01_AE

At least five online tools assigned 35 cases *pol* CRF01_AE, all but four of these clustered with the CRF01_AE reference sequence on the phylogenetic tree. Two sequences clustered with 15_01B on the phylogenetic tree, both sequences were also identified as either 15_01B or a complex ISR by SCUEAL. Two sequences clustered with A1 on the phylogenetic tree, SCUEAL identified one as 15_01B and the other as 01_AE. All other online tools assigned these four sequences as 01_AE. There were another seven sequences inferred as CRF01_AE that one online tool classified as either an A1 or 15_01B sequence.

7.3.3 pol – discordance between online tools

In addition to identifying inferred subtypes, the study aim was also to report where there was discordance between the online subtyping tools and to examine these cases in more detail. There was discordance between at least three of the six online tools for the remaining 23 *pol* sequences. Each sequence was assigned an ISR by at least one online subtyping tool

and these will be discussed later in the chapter.

7.4 Online tool parameters – *env* subtype

7.4.1 env – concordance between all four online tools

There was concordance between all four online tools for 62% (136) of *pol* sequences with a subtype distribution as follows: 95 subtype B sequences, 34 subtype C sequences, 5 A1 sequences, and 2 subtype D sequences. The remaining 85 sequences were broken down into sequences that were assigned the same subtype/CRF by at least three of the four tools (n=62) or sequences that could not be assigned because of differences between the online tools (n=23).

7.4.2 env inferred subtype – concordance using online tool parameters (three of the four tools were concordant)

When assessing subtype by concordance between three of the four tools, 90% (198) of cases met this criterion with a subtype distribution as follows: 114 subtype B, 39 subtype C, 34 01_AE, 9 subtype A1, and 2 subtype D.

7.4.2.1 Subtype A1

All nine A1 sequences had concordant subtypes between jpHMM, REGA and COMET (MEGA phy was also concordant). LANL BLAST assigned two of the sequences as A2, one as A3 and one as an A1/D ISR.

7.4.2.2 Subtype B

Of the 114 cases assigned *pol* subtype B by at least three of the online tools, all clustered with or around subtype B on the phylogenetic tree, but 10 clustered more closely with B-like CRFs (14_BG, 15_01B, 03_AB, 28_BF, 29_BF, or 47_BF). COMET assigned one of the phy 29_BF sequences as 15_01B, and LANL BLAST was unable to match the one 14_BG phy sequence to a reference sequence. REGA was unable to assign seven of the 114 subtype

B assigned sequences, COMET identified three sequences as 15_01B, LANL BLAST identified one sequence as a 01_AE/B ISR, and could not assign a subtype for nine sequences; seven of which were identified as a Simian-Human Immunodeficiency Virus (SHIV).

7.4.2.3 Subtype C

Concordance between all online tools and phy occurred for 34 of the 39 subtype C sequences, four could not be classified by jpHMM, and LANL BLAST identified one as being most similar to a B/C ISR.

7.4.2.4 Subtype D

Concordance between all online tools and phy occurred for the two subtype D sequences.

7.4.2.5 CRF01_AE

All the 34 01_AE sequences were assigned 01_AE by COMET, LANL BLAST and phy, and all but one were assigned 01_AE by jpHMM (one was unclassified). REGA is only capable of identifying pure subtypes when sequences are less than 800bp and it identified 12 as A1, 10 as C, 11 as G and 1 as J. None of the REGA sequences had a bootstrap value ≥70%.

7.4.3 env – discordance between online tools

There was discordance between at least three online tools for 23 sequences (case 68 and 133 had an even split between the four online tools). The 23 *env* sequences were assigned various subtypes/CRFs as seen in Table 38. One sequence was identified as 07_BC (case 133) and one as 08_BC (case 68) by COMET and LANL BLAST (and MEGA phy) while jpHMM and REGA assigned these sequences as subtype C. Fifteen sequences (case 6,65-67,108-111,114-117,124,139,145) were identified as 'A1/check for 02_AG' by COMET, LANL BLAST assigned nine of these as 02_AG, two as 63_02A1, two as subtype A1, one

as subtype A3 and one as an 02_AG/A1 recombinant. The same 15 sequences were assigned subtype A1 by REGA and all but two by jpHMM (two were unassigned). There was one sequence (case 2) assigned as subtype D by COMET and REGA (MEGA phy also assigned D), while LANL BLAST assigned the same sequence as an A/D variant and jpHMM could not assign a subtype. Two sequences (case 112 and 113) were assigned as 02_AG by COMET and LANL BLAST (and MEGA phy), but were typed as A1 by REGA, and A1 or A1/B by jpHMM. The final two sequences (case 157 and 162) were assigned 01_AE by COMET and LANL BLAST (and MEGA phy) but both were assigned as 01_AE/B variants by jpHMM and A1 or C by REGA.

Table 38. Discordance between three or more online tools – *env* sequences

Case	jpHMM	REGA	COMET	LANL BLAST	MEGA phy
2	-	D^	D	A/D^	D
6	A1	A1^	A1/check for AG	02_AG^	02_AG
65	A1	A1^	A1/check for AG	02_AG^	02_AG
66	A1	A1^	A1/check for AG	02_AG^	02_AG
67	A1	A1^	A1/check for AG	02_AG^	02_AG
68	C	C^	08_BC	08_BC^	08_BC
108	A1	A1^	A1/check for AG	02_AG^	02_AG
109	A1	A1	A1/check for AG	02_AG^	02_AG
110	A1	A1^	A1/check for AG	63_01A1^	A1
111	-	A1^	A1/check for AG	63_02A1^	A1
112	A1/B	A1	02_AG	02_AG^	02_AG
113	A1	A1^	02_AG	02_AG^	02_AG
114	A1	A1^	A1/check for AG	02_AG^	02_AG
115	A1	A1^^	A1/check for AG	A1^	02_AG
116	A1	A1^	A1/check for AG	02_AG^	36_cpx
117	A1	A1	A1/check for AG	02_AG^	02_AG
124	A1	A1^	A1/check for AG	02_AG/A1^	45_cpx
128	01_AE	A1	01_AE	01_AE^	01_AE
133	C	C^	07_BC	07_BC^	07_BC
139	A1	A1^	A1/check for AG	A3^	A1
145	A1	A1	A1/check for AG	A1^	02_AG
157	01_AE/B	A1	01_AE	01_AE^	01_AE
162	01_AE/B	C	01_AE	01_AE^	01_AE

Key: ^ (denotes a bootstrap or similarity percentage of ≥70%).

7.5 Genomic diversity

The majority (86%, 191) of cases had concordant *pol* and *env* subtypes using the same online tool criteria mentioned earlier in the chapter (5/6 online tools for *pol*, 3/4 for *env*, MEGA phy not included), with a subtype distribution as follows: 7 subtype A1, 111 subtype B, 39 subtype C, 1 subtype D and 33 01_AE. As stated above, the remaining 15% (30) of cases were assessed further to determine if they were genomic recombinants and will be discussed in the next section.

7.6 Recombination identification

Recombinant cases were identified as intergene recombinant (different subtype for *pol* and *env*) or intragene recombinant (different subtype at PR and RT within the *pol* region) SCUEAL also identified intrasubtype (sub-subtypes within a subtype, i.e: subtype B1, B2, B3) recombination using *pol* sequences only. SCUEAL was considered to be the most reliable online tool to detect recombination events and correctly identify subtype, it utilizes up-to-date reference sequences taken from the LANL BLAST reference database and is capable of detecting recombination breakpoints.

The 221 sequences were split into five groups, as seen in Table 39.

Table 39. Subtype distribution and recombination events using online tool criteria

Group	***N (%)***
1. pol and *env* concordant	191 (86)
2. pol and *env* discordant	3 (1)
3. pol unassigned/*env* subtyped	4 (2)
4. pol subtyped/*env* unassigned	4 (2)
5. pol and *env* unassigned	19 (9)

Groups 1 and 2 comprised 88% (194) of all cases. These groups were considered successfully assigned subtypes using the online tool parameters (5/6 tool concordance for

pol, or 3/4 for *env)*. Group 1 had a subset of 20 cases which had ISR identified by one online tool, but because the other 5 tools (*pol*) and 3 tools (*env*) were all concordant, these cases were not considered to carry ISR sequences for the purpose of this study.

Group 2 cases carried intergene recombinant viruses, *pol* and *env* were successfully assigned discordant subtypes.

Group 3 consisted of four cases with successfully subtyped *env* sequences but concordance between the online tools was not reached for *pol*, leaving the sequences unassigned. The same occurred in reverse for group 4. These cases were further analyzed.

The 19 cases in group 5 were not assigned an inferred subtype for either *pol* or *env* and were also further analyzed.

Groups 2–5 will be discussed in the next section. All cases in these combined groups (n=30) had at least one online tool that identified ISR. Table 40 below shows the different variants that each online tool identified for the 30 cases.

Table 40. ISR variants identified within 30 cases by online subtyping tools

ISR variant identified	***pol* (N)**	***env* (N)**	**ISR variant identified**	***pol* (N)**	***env* (N)**
jpHMM			**SCUEAL**		
Complex	3		Complex	4	-
A1/B	2	1	A-ancestral/G	1	-
AE/B		1	A/G	3	-
C/B	1		A1/U	1	-
G/B	3		A4/G	1	-
G/A1	6		AE/B	2	-
LANL BLAST			AE/G	2	
A/D		2	B/C	2	-
AE/B	1		B/D	1	-
AG/A1	1	1	B/G	2	-
AG/22_01A1	1		G	1	-
AG/U	1		**REGA**		
B/C		1	A1/B	1	-
Stanford CPR			A1/K	1	-
A1/AE	1	-	A1-like	2	-
AE/B	1	-	B/G	3	-

AE/AG	1	-	B-like	1	-
AE/K	1	-	**COMET**		
AG/B	3	-	Complex	4	
B/C	1	-	B/AE	1	

Key: - (online tool did not subtype this region); Complex (online tool identified the sequence as a complex recombinant form consisting of two or more breakpoint locations).

7.6.1 Group 1. Concordant pol/env sequences

These cases were not intergene recombinant viruses and are not discussed further in this chapter.

7.6.2 Group 2. Intergene recombination (discordant pol/env sequences)

Of the cases that were successfully assigned a subtype, three had discordant *pol*/*env* subtypes as shown in Table 41. These cases will be discussed in detail later in the chapter.

Table 41. Assigned genomic recombinant cases

Case	*pol subtype*	*env subtype*
136	A1	B
135	A1	D
91	C	A1

7.6.3 Group 3. Intragene recombination within pol

Four cases had a successfully assigned pure subtype or CRF *env* sequence, and an unassigned, possible ISR *pol* sequence, Table 42. SCUEAL identified all four as *pol*-ISRs, jpHMM and COMET identified two, and Stanford CPR and LANL BLAST both identified one. REGA identified one sequence as an ISR and another two as 'like' a pure subtype, Table 42. All six online tools identified case 153 with a *pol*-ISR and three online tools identified case 200 with a *pol*-ISR. The likely *pol* subtype for each case is shown in the last column in Table 42, with the overall likely *pol/env* subtype for these four case as follows:

	pol	*env*
Case 38	ISR	B
Case 125	ISR	A1
Case 153	ISR	01_AE
Case 200	ISR	B

Table 42. Cases with an assigned pure subtype or CRF *env* region, and unassigned (and possible ISR) *pol* region

Case	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	Likely *pol* subtype
38	B	B-like^	B^	B/D^	B	B^	B/D
125	A1	A1^	01_AE^	A1/ U^	A1	A1^	A1/U
153	A1/B	A1/B	01_AE^/B^	01_AE/B^	31_BC/A1/B	01_AE/B^	01_AE/B
200	A1/B	A1-like^	01_AE^	01_AE/B ^	B/01_AE	01_AE^	01_AE/B

Key: U (Unassigned an inferred subtype), ^ (denotes a bootstrap or similarity percentage of ≥70%). The likely *pol* subtype was identified by SCUEAL and COMET, with SCUEAL reporting a similarity percentage ≥70%.

7.6.4 Group 4 – Intragene recombination within env

Four cases had a successfully assigned pure subtype or CRF for the *pol* sequence and an unassigned and possible ISR for the *env* sequence, Table 43.

Table 43. Cases with a pure subtype/CRF *pol* region and possible *env* ISR

Case	jpHMM	REGA	COMET	LANL BLAST	Likely *env* subtype
2	U	D^	D	A/D^	A/D
99	C	C^	C/check for 07	B/C^	07_BC
157	01_AE/B	A1	01_AE	01_AE^	AE
162	01_AE/B	C	01_AE	01_AE^	AE

Key: U (Unassigned an inferred subtype), ^ (denotes a bootstrap or similarity percentage of ≥70%). The likely *env* subtype was identified by COMET and LANL BLAST, with LANL BLAST reporting a similarity percentage ≥70%.

Cases 157 and 162 were classified by LANL BLAST as 01_AE but jpHMM classified them as 01_AE/B ISRs. jpHMM may incorrectly classify small sequences near the 3' end of the genome and it is likely these sequences were 01_AE.

LANL BLAST classified case 2 as an A/D ISR and case 99 a B/C ISR. Case 99 was most likely 07_BC as evidenced by the COMET classification, Table 43.

The likely *env* subtype for each case is shown in the last column in Table 43, with the likely *pol/env* subtype for these four cases as follows:

	pol	*env*
Case 2	D	ISR
Case 99	C	07_BC
Case 157	01_AE	01_AE
Case 162	01_AE	01_AE

7.6.5 Group 5. Possible unique recombinants – pol-ISR and env-ISR

The final group comprised of cases that were possible unique recombinants, due to discordance between online tools, and one or more tools assigning both the *pol* and *env* regions as complex recombinant forms. The 19 cases with unclassified *pol* and *env* sequences were not assigned an inferred subtype because they did not meet the criteria using the online tool parameters. These cases are shown in Table 44.

Further analysis of the 19 cases identified 18 as possible unique intergenomic recombinant virus and one case was assessed as a *pol* 02_AG/*env* 02_AG virus.

In total, 27 cases combined from groups 2–5 had viruses classified as intergenomic recombinants, that is, complex recombinant viruses. These will be discussed in the next section.

Table 44. Cases from group 5, with *pol* and *env* sequences not assigned an **inferred subtype**

Case	*pol gene*						*env gene*				Possible unique recombinant *pol*/*env*
	jpHMM	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**	**jpHMM**	**REGA**	**COMET**	**LANL BLAST**	
124	A1	A1 K	K^ 01_AE^	A1^	A1	A1^	A1	A1^	A1 check AG	02_AG A1^	ISR/ISR
6	A1 B A1	02_AG^	02_AG^ 02_AG^	CISR^ A1 G	02_AG	02_AG^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
65	G B	B G	02_AG^ B^	CISR^ B G	CISR Complex	B^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
66	G B	B G	02_AG^ B^	CISR^ B G	CISR Complex	B^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
67	G B	B G	02_AG^ B^	CISR^ Complex	CISR Complex	B^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
108	G A1	02_AG^	02_AG^ 02_AG^	CISR^ A1 G	02_AG	02_AG^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
109	A1	02_AG^	02_AG^ 02_AG^	CISR^ Complex	02_AG	02_AG^	A1	A1	A1 check AG	02_AG^	ISR/02_AG
113	G A1	02_AG^	02_AG^ 02_AG^	G^	02_AG	02_AG^	A1	A1^	02_AG	02_AG^	02_AG/02_AG
112	B	G^	02_AG^ 02_AG^	CISR^ G	02_AG	02_AG^	A1 B	A1	02_AG	02_AG^	ISR/02_AG
114	G	02_AG^	02_AG^ 02_AG^	CISR^ 01_AE G	02_AG	02_AG U^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG

Case	*pol gene*						*env gene*				
	jpHMM	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**	**jpHMM**	**REGA**	**COMET**	**LANL BLAST**	**Possible unique recombinant *pol*/*env***
117	G A1	02_AG^	02_AG^ 02_AG^	CISR^ 01_AE G	02_AG	02_AG^	A1	A1	A1 check AG	02_AG^	ISR/02_AG
115	G A1	02_AG^	02_AG^ 02_AG^	G^	02_AG	02_AG^	A1	A1^^	A1 check AG	A1^	02_AG/A1
145	B	02_AG^	02_AG^ 02_AG^	CISR^ A4 G	02_AG	02_AG^	A1	A1	A1 check AG	A1^	ISR/A1
133	B C B	07_BC^	B^ C^	CISR^ B C	07_BC	07_BC^	C	C^	07_BC	07_BC^	ISR/07_BC
68	C B	08_BC^	C^ C^	CISR^ B C	08_BC	08_BC^	C	C^	08_BC	08_BC^	ISR/08_BC
116	G A1	02_AG^	02_AG^ 02_AG^	CISR^ A-ancestral G	02_AG	02_AG^	A1	A1^	A1 check AG	02_AG^	ISR/02_AG
110	G A1	02_AG^	02_AG^ 02_AG^	CISR^ A G	02_AG	02_AG^	A1	A1^	A1 check AG	63_02A1^	ISR/63_02A1
111	G B G	02_AG^	02_AG^ 02_AG^	CISR^ Complex	02_AG	02_AG A1^	U	A1^	A1 check AG	63_02A1^	ISR/63_02A1
139	A1	A1-like^	02_AG^ 01_AE^	A2	A1	02_AG 22_01A1^	A1	A1^	A1 check AG	A3^	A-like/A-like

Key: U (Unassigned an inferred subtype), ^ (denotes a bootstrap or similarity percentage of ≥70%), ^^ (denotes a bootstrap value ⩾70% but REGA could not assign a pure subtype because the query sequence did not sub-cluster with a pure subtype reference as per REGA subtyping algorithm).

7.7 Intergene recombination – case characteristics

The 27 cases fell into the following 15 *pol/env* intergene recombinant categories, Table 45. Table 46 shows the demographic information for each case.

Table 45. Possible unique recombinant cases

N	*pol*	*env*
10	U ISR	02_AG
2	U ISR	A1
2	U ISR	63_02A1
2	U ISR	B
1	U ISR	01_AE
1	U ISR	07_BC
1	U ISR	08_BC
1	U ISR	U ISR
1	A-like	A-like
1	A1	B
1	A1	D
1	AG	A
1	C	A1
1	C	07_BC
1	D	U ISR

Key: U ISR (possible unique ISR)

Table 46. Characteristics of cases with intergene recombination (*pol/env*)

Case	Likely *pol/env* subtype	Location acquired	Region born	Sex	Risk	Year dx
38	U ISR/B	Aus	N/A	M	MSM	2001
68	U ISR/08_BC	O	N/A	M	Direct blood contact	2004
133	U ISR/07_BC	O	N/A	M	Het/IDU	2001
114	U ISR/AG	O	Africa	F	Heterosexual	2007
115	AG/A1	O	Africa	F	Heterosexual	2011
116	U ISR/AG	O	Africa	F	MTCT	2011
124	U ISR/U ISR	O	Africa	F	Heterosexual	2011
135	A1/D	O	Africa	F	Heterosexual	2001
136	A1/B	O	Africa	F	Heterosexual	2004
2	D/U ISR	O	Africa	M	Heterosexual	2004
6	U ISR/AG	O	Africa	M	N/A	2006
65	U ISR/AG	O	Africa	M	Heterosexual	2009
67	U ISR/AG	Aus	Africa	M	Heterosexual	2011
91	C/A1	O	Africa	M	Heterosexual	2008

109	U ISR/AG	Aus	Africa	M	Heterosexual	2010
117	U ISR/AG	O	Africa	M	Heterosexual	2011
125	U ISR/A1	O	Africa	M	Heterosexual	2010
139	A like/A like	O	Africa	M	Heterosexual	2009
145	U ISR/A1	O	Africa	M	Heterosexual	2008
99	C/07_BC	O	Asia	F	Het/IDU	2010
153	U ISR/AE	O	Europe	M	Direct blood contact	2011
200	U ISR/B	Aus	Europe	M	Het/IDU	2011
108	U ISR/AG	O	Australia	F	Heterosexual	2004
66	U ISR/AG	Aus	Australia	M	MSM	2010
112	U ISR/AG	Aus	Australia	M	MSM	2009
110	U ISR/63_02A1	O	CA/ME	F	Direct blood contact	2009
111	U ISR/63_02A1	O	CA/ME	M	Direct blood contact	2010

Key: Het/IDU (Heterosexual transmission with IDU risk); MSM (Men who have sex with other men); O (overseas acquired infection); CA/ME (Central Asia/Middle East); Year dx (Year of diagnosis), U ISR (possible unique ISR)

When assessing these complex recombinant cases by demographic characteristics, the majority were male (67% 18). Most males and females acquired the infection overseas (78%, 21/27) and most were born overseas (78%, 21) with African-born people predominating (76%, 16/21).

There was a notable pattern of complex recombination type by region of birth. Of the sixteen African-born cases, 10 were male and six were female. All acquired the infection overseas, and all reported heterosexual transmission except for one MTCT case and one male who did not list transmission risk. African-born people carried seven of the 10 U ISR/AG variants, and all the variants containing subtype A (n=7). The remaining two African-born people carried a subtype D variant and a complex variant possibly comprised of subtype A1, K, 01_AE and 02_AG.

Three cases were Australian-born people; two males and one female. All three had a likely complex *pol*-ISR, and *env*-02_AG virus, the two males reported MSM risk while the

female reported heterosexual contact. One male and the female reported overseas acquisition, the other male reported acquisition in Australia.

Two cases were children born in Central Asia (1 male, 1 female). Both transmissions were reported as originating from medical procedures, and both had a likely *pol*-ISR and *env*-63_02A1 virus.

Two cases were European-born males, one reported an overseas medical procedure as the route of transmission and had a possible *pol*-ISR and *env*-01_AE, the other reported heterosexual sex with IDU risk in Australia, and had a *pol*-ISR and *env*-B.

One case was born in Asia, a female who reported overseas acquisition through heterosexual sex with IDU risk, and had *pol*-C and *env*-07_BC.

7.8 *pol* intersubtype recombination – breakpoint locations

The breakpoint locations were identified for 20 of the 27 complex cases through SCUEAL. The remaining seven were not identified as complex viruses by SCUEAL and therefore did not have breakpoint locations available through the chosen tool. The *pol* sequences for the 20 SCUEAL identified intergene recombinant cases were classified as unique ISRs, including a predicted subtype, model-averaged support for that subtype, and the percent likelihood that ISR was present. The model-averaged support cutoff for evidence of ISR is ≥50%. Breakpoint locations were also identified for each *pol*-ISR sequence and information for these cases is presented in Table 47, and Figures 28, 29a and 29b.

Demographic analysis identified only four of these cases as people born in Australia. Case 66 and 112 had *pol* subtype G ISR variants acquired in Australia, case 108 had a 37_cpx variant acquired overseas and case 161 had a 15_01B variant acquired overseas. Three of these people were male (all MSM transmission) and one was a female who reported heterosexual transmission. Diagnosis occurred between 2001 and 2010.

Three cases did not have a listed region of birth. Cases 68 and 133 had B/C variants acquired overseas, and case 38 had a B/D variant acquired in Australia. All were male, with one each reporting MSM (diagnosed 2001), heterosexual sex with IDU risk (2001), and direct blood contact (2004).

The remaining 14 cases were people born overseas and comprised 10 males and 4 females, all diagnosed between 2006 and 2011. There were eight different ISR variants among this group, which consisted of ten African-born people, two people born in Central Asia, and two born in Europe. Three overseas born people acquired the virus in Australia, namely cases 67 and 109 who were both African-born males and had viruses identified as complex ISRs, and case 200 who was a European-born male who acquired a complex subtype B/01_AE virus.

The two European-born cases (153 and 200) were the only ones to have the 01_AE/B variant, both diagnosed in 2011. One reported direct blood contact overseas and one reported heterosexual sex with IDU risk within Australia.

Table 47. SCUEAL-classified unique *pol*-ISR sequences

Case	Full subtyped assigned	% support subtype	% support ISR	Breakpoint 1 (CI)	Breakpoint 2 (CI)	Breakpoint 3 (CI)	Breakpoint 4 (CI)
6	02_AG/37_cpx/G/A1	68	100	183 (169–197)	282 (183–381)	959 (958–960)	
38	B/D/B	41	94	702 (699–705)	911 (875–947)		
65	G/37_cpx/B	66	100	107 (106–108	360 (355–365)		
66	G/37_cpx/B	49	100	107 (105–109)	360 (351–369)		
67	G/A-ancestral/B/B/B	75	100	110 (109–111)	419 (418–420)	712 (695–729)	818 (738–898)
68	C/08_BC/08_BC/C	47	100	108 (100–116)	616 (616–616)	926 (925–927)	
108	37_cpx/02_AG/37_cpx	57	100	469 (439–499)	983 (977–989)		
109	G/01_AE/02_AG/G/A-ancestral	35	100	176 (175–177)	298 (298–298)	844 (843–845)	959 (948–970)
110	G/G/A	82	100	451 (444–458)	983 (982–984)		
111	02_AG/A1/02_AG/A-ancestral	51	100	172 (171–173)	316 (313–319)	918 (916–920)	
112	G/02_AG/18_cpx	75	100	107 (104–110)	499 (495–503)	914 (913–915)	
114	G/01_AE/02_AG/G	94	100	176 (175–177)	298 (298–298)	918 (896–940)	
116	G/A-ancestral/G/18_cpx	89	100	107 (106–108)	406 (404–408)	953 (952–954)	
117	G/01_AE/02_AG/18_cpx	80	100	171 (169–173)	299 (294–304)	912 (911–913)	
125	A1/A1/U	66	91	476 (423–529)	916 (913–919)		
133	07_BC/B/07_BC/07_BC/C	45	100	184 (183–185)	291 (288–294)	733 (733–733)	932 (931–933)
135	A1/A-ancestral/A3	82	99	352 (349–355)	566 (549–583)		
145	G/A4/G/A4	32	100	116 (115–117)	406 (405–407)	958 (957–959)	
153	01_AE/B	82	100	725 (661–789)			
200	15_01B/B/15_01B	33	100	745 (736–754)	944 (942–946)		

Key: % support subtype (the % likelihood that the assigned subtypes and CRFs within the recombinant region sequenced are accurate; % support ISR (the % likelihood that true recombination events occur within the region sequenced); Breakpoint (the point in the sequence where a change from one subtype or CRF to another occurs); CI (confidence intervals).

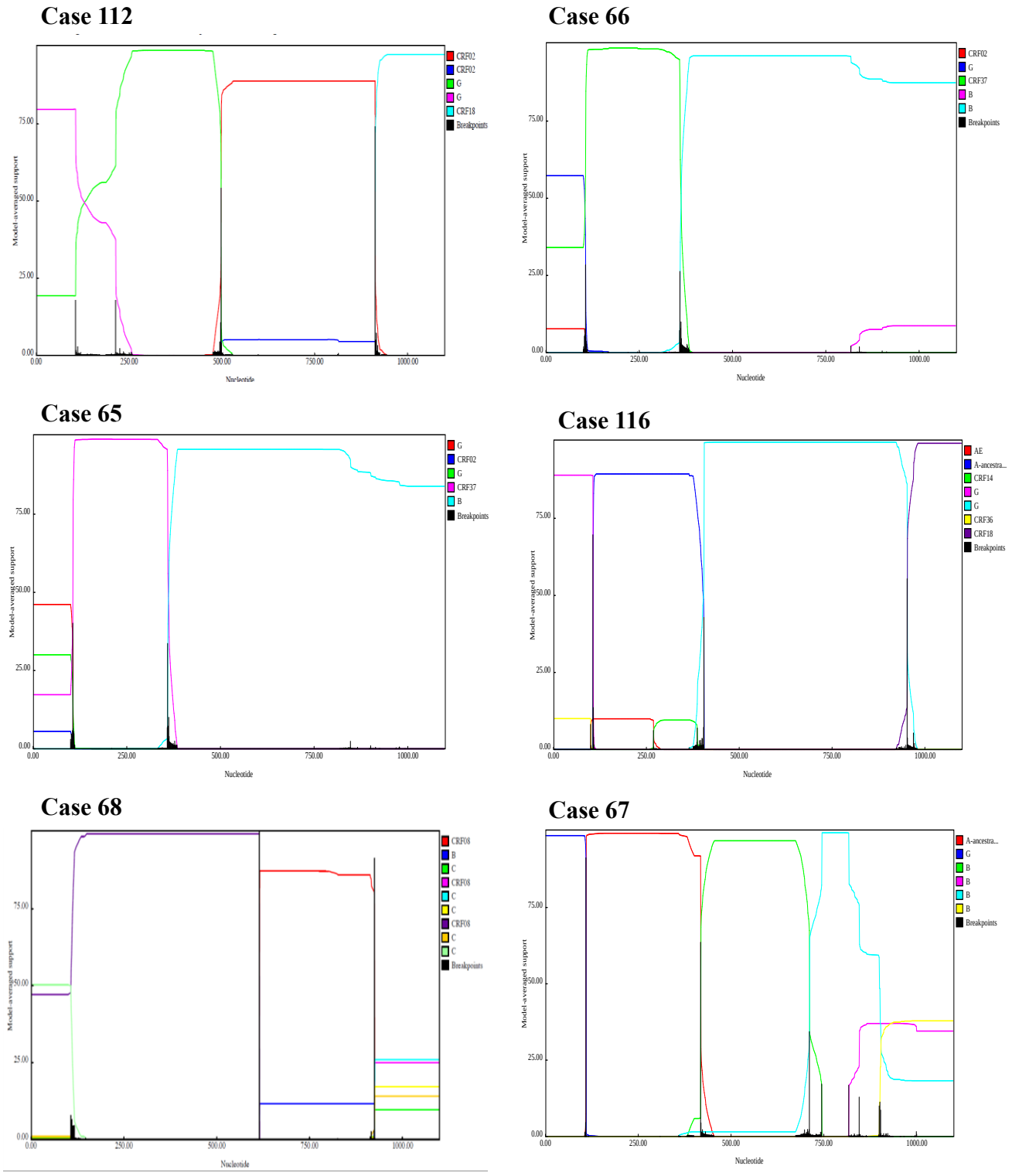


Figure 28. SCUEAL recombination plots for *pol* sequences assigned complex ISRs. Recombination breakpoints were identified at sites along *pol*-PR/RT that were not associated with known CRFs. Associated parental strains are listed on the right of the plot. The Y axis shows the model-averaged support for a given parental reference strain, the X axis shows the nucleotide position for the 1098bp *pol* region, which spans 2253-3350 of the HXB2 reference sequence.

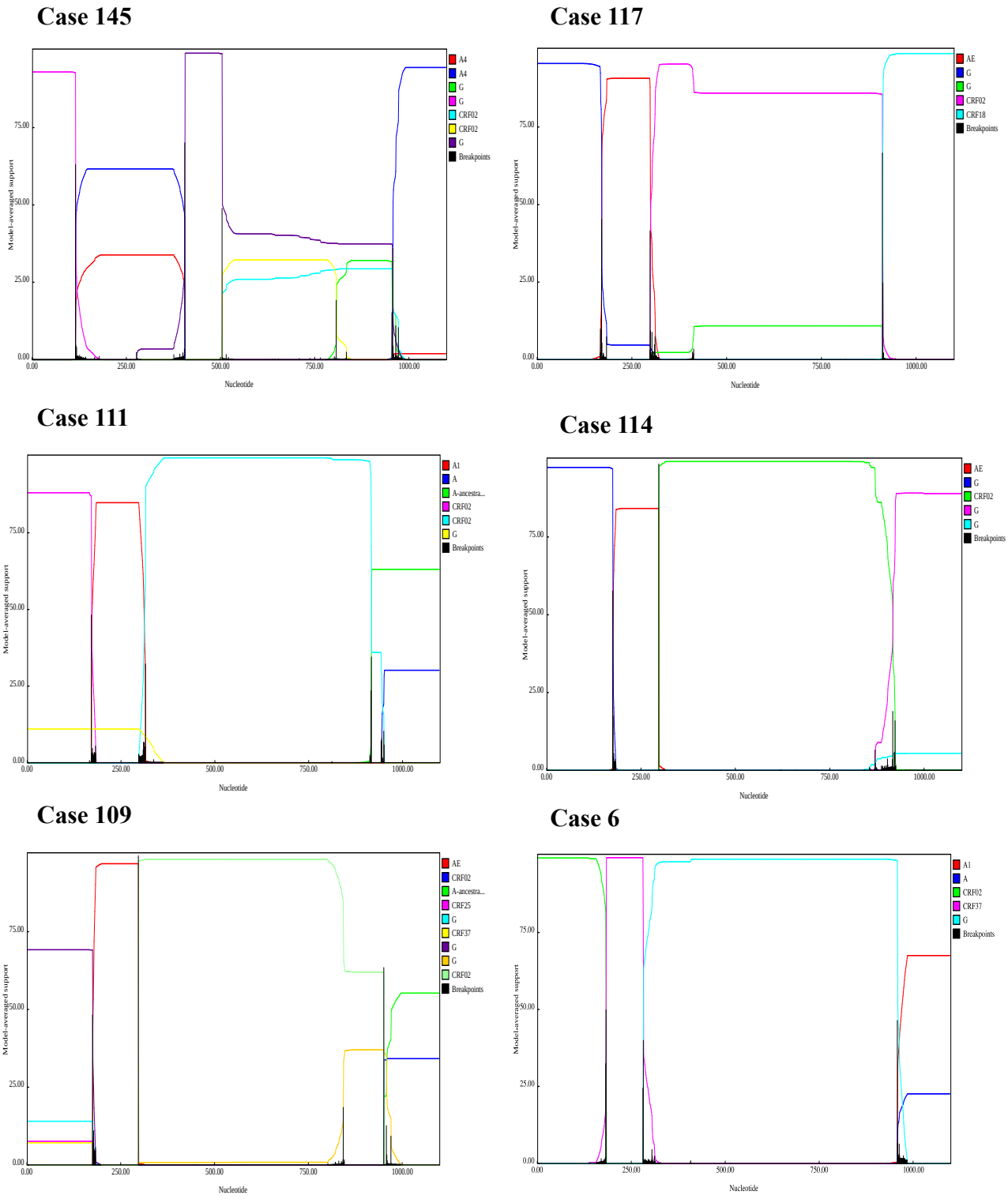


Figure 29a. SCUEAL recombination plots for *pol* sequences assigned complex ISRs. Recombination breakpoints were identified at sites along *pol*-PR/RT that were not associated with known CRFs. Associated parental strains are listed on the right of the plot. The Y axis shows the model-averaged support for a given parental reference strain, the X axis shows the nucleotide position for the 1098bp *pol* region, which spans 2253-3350 of the HXB2 reference sequence.

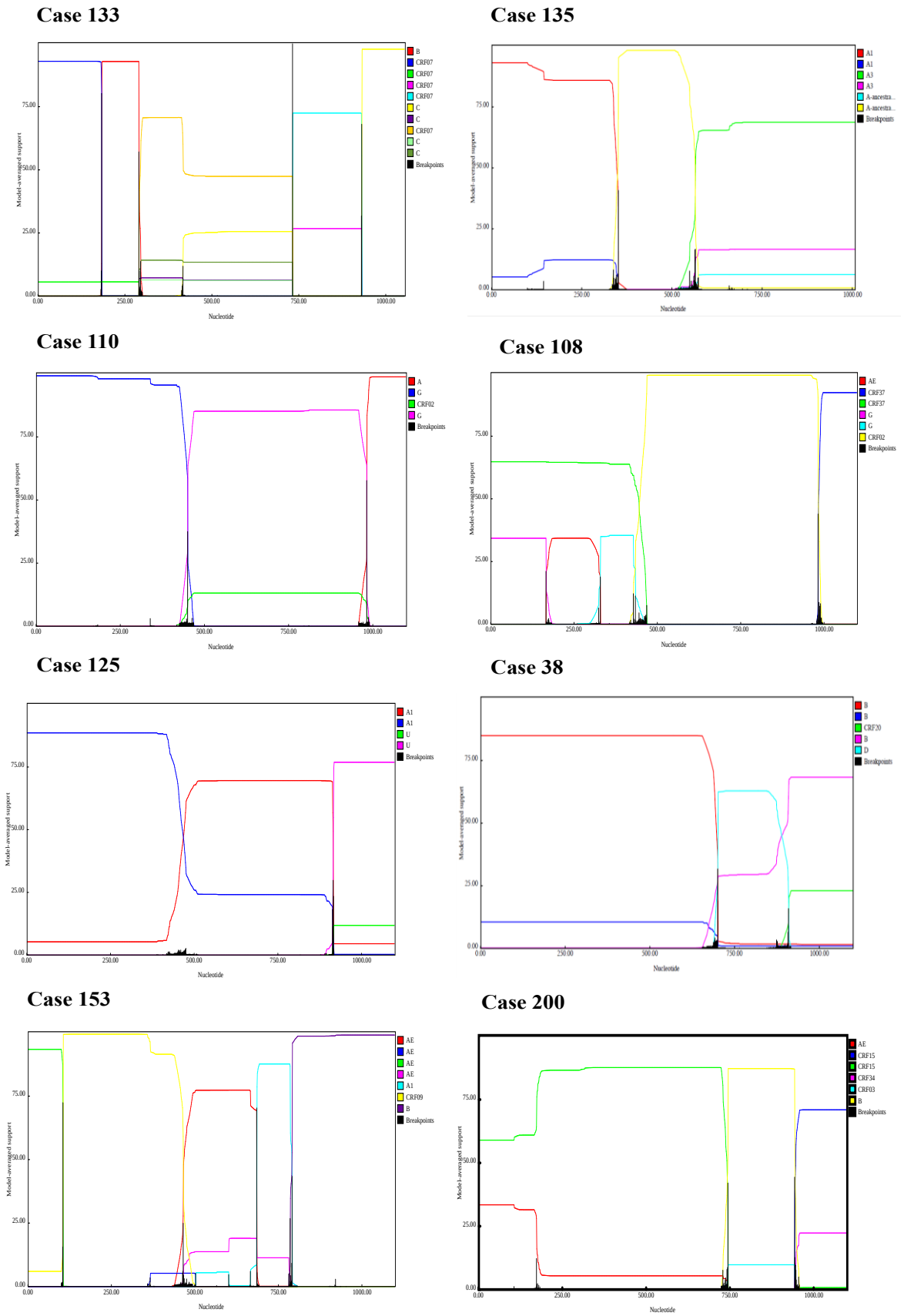


Figure 29b. SCUEAL recombination plots for *pol* sequences assigned complex ISRs. Recombination breakpoints were identified at sites along *pol*-PR/RT that were not associated with known CRFs. Associated parental strains are listed on the right of the plot. The Y axis shows the model-averaged support for a given parental reference strain, the X axis shows the nucleotide position for the 1098bp *pol* region, which spans 2253-3350 of the HXB2 reference sequence.

7.9 *pol* intrasubtype recombination

The following section also reports InSR using the SCUEAL tool, which identified 16 cases that carried multiple strains of the same pure subtype. The SCUEAL model's averaged support cutoff for evidence of InSR is ≥50%. Table 48 shows recombination characteristics and demographic information.

Demographic characterization shows that almost all cases (88%, 14/16) had *pol* InSR B sequences and these were all male cases. All but one acquired the infection in Australia and all but two acquired it via MSM transmission (including two MSM with IDU risk transmissions). Ten of the 16 (63%) cases were diagnosed before 2007.

The remaining two cases had *pol* InSR C sequences, one female adult and one female child, both born in Africa and acquired the infection overseas. LANL BLAST analysis found the *pol* sequences were most similar to sequences originating from Africa (≥95%). The two females were part of the same high reliability *pol* subtype C cluster (n=41), but not related by transmission. The female child was part of a *pol* transmission cluster of three, most likely a mother, father and child transmission, although the mother's and father's *pol* sequences showed no evidence of recombination. The female adult was not part of any transmission cluster.

Table 48. Intrasubtype recombination of 16 *pol* sequences assigned by SCUEAL

Case	*pol* SCUEAL	% support subtype	% support InSR	Age	Location acquired	Region of birth	Sex	Transmission Risk	Year dx	*pol* MEGA phy	*env* MEGA phy
22	B	48	52	38	Australia	Australia	M	MSM	2003	B	B
41	B	63	64	37	Australia	Australia	M	MSM & IDU	2006	B	B
186	B	39	69	31	Australia	N/A	M	MSM	2005	B	B
15	B	69	72	38	Australia	Australia	M	Heterosexual	2011	B	B
189	B	67	78	36	Australia	N/A	M	MSM	2005	B	B
56	B	81	81	35	Australia	Australia	M	MSM & IDU	2006	B	B
187	B	71	84	27	Australia	N/A	M	MSM	2006	B	B
185	B	84	87	46	Australia	Australia	M	MSM	2010	B	B
183	B	87	88	36	Australia	N/A	M	MSM	2005	B	B
35	B	90	94	32	Australia	Australia	M	MSM	2005	B	B
184	B	90	94	46	Australia	Europe	M	MSM	2008	B	B
54	B	94	95	34	Australia	Australia	M	MSM	2012	B	B
55	B	98	99	25	Australia	N/A	M	MSM	2000	B	B
219	B	100	100	34	O	Australia	M	Heterosexual	2005	B	B
74	C	70	79	7	O	Africa	F	MTCT	2007	C	C
82	C	96	100	28	O	Africa	F	Heterosexual	2001	C	C

7.10 Comparison with routine genotypic testing using Stanford CPR

As mentioned previously, the current practice in South Australia for genotypic drug resistance testing is to sequence a 1098bp *pol* fragment, then submit it to the Stanford CPR online tool. From all new diagnoses in South Australia between 2000 and 2012 that had routine genotyping conducting (n=513), the Stanford CPR tool detected 20 *pol*-ISRs, which equates to 4% of the total HIV infected population for that time period.

Thirteen of these Stanford CPR-identified *pol*-ISR sequences were included in the *pol/env* phylogenetic study, comprising 6% of the total 221 cases. Table 49 shows the subtype distribution using the Stanford CPR online tool compared with SCUEAL, and the inferred subtype using the online tool criteria for both *pol* and *env*. In total, 50 cases were identified as ISR by at least one tool. This was narrowed down by only classifying ISR cases as those with at least one online tool identifying an ISR and discordance between the other tools, and then examining results for each sequence carefully. Using this method, 27 *pol/env* intergene recombinants were identified. This equates to 12% of the cohort used for the phylogenetic study.

Identification of *pol*-ISR virus increased from 6% to 10% when the six online tools were used compared with only Stanford CPR, (p=0.06). The identification of recombinant virus using the online tool criteria to assess *pol* and *env* genes doubled from 6% to 12%, compared with only typing the *pol* gene using Stanford CPR, (p=0.03), Figure 30.

Table 49. Subtype distribution using routine surveillance information from Stanford CPR compared with other online subtyping tools.

Subtype	Stanford CPR *N (%)*	SCUEAL *N (%)*	Inferred *pol* (5/6 tools) *N (%)*	Inferred *env* (3/4 tools) *N (%)*	Final subtype assigned *pol* & *env* *N (%)*
01_AE	38 (17)	28 (13)	35 (16)	33 (15)	35 (16)
02_AG	12 (5)				1 (.5)
15_01B		6 (3)			
19_cpx		1 (.5)			
A	3 (1)	10 (5)	9 (4)	9 (4)	7 (3)
B	111 (50)	108 (49)	111 (50)	114 (52)	111 (50)
C	42 (19)	41 (19)	41 (19)	39 (18)	39 (18)
D	2 (1)	1 (.5)	2 (1)	2 (1)	1 (.5)
G		2 (1)			
InSR	13 (6)	24 (11)	23 (10)	24 (11)	
Possible URF					27 (12)
Total	221 (100)	221 (100)	221 (100)	221 (100)	221 (100)

Key: InSR (Intrasubtype recombinant [within a gene]), Possible URF (Evidence of intergene recombination without assignment to a known CRF reference sequence).

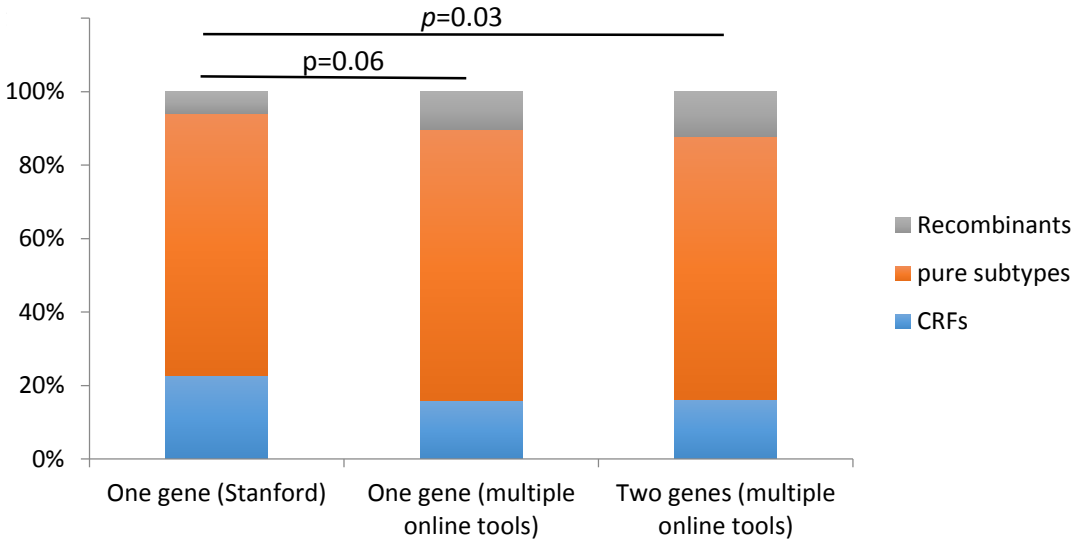


Figure 30. HIV-1 subtype distribution based on *pol* subtype or *pol*/*env* subtype. Column one is the proportion of *pol* subtypes as classified by the Stanford CPR online tool, Column two is the proportion *pol* subtypes as classified by five out of six concordant online tools, and Column three is the proportion of subtypes as classified by *pol/env* combined. *p* value shown for differences between column one and column two, and column one and column three.

7.11 Summary

The data demonstrated that using the chosen online tool criteria resulted in an accurate but conservative estimate of non-B viral prevalence in the South Australian cohort, with COMET, SCUEAL and jpHMM being the best tools for identifying non-B variants including unique and complex variants. The latter two also identified specific recombination breakpoints.

There was high concordance between MEGA phy and online tools when identifying subtype B, C and CRF01_AE, but low concordance for other non-B variants including complex and possibly unique recombinant forms. Previous studies have reported limitations when identifying non-B variants using rapid online tools and phy.[77,182,192,193,295,296]

Using the online tool criteria, 10% of sequences could not be confidently assigned to any previously defined subtype or CRF, and further analysis identified 12% of all 221 cases as intergene recombinant viruses including complex and possible unique variants.

This was double the proportion found using the Stanford CPR tool and *pol* region alone, but similar to proportions found in a UK study using multiple online tools.[193] These complex viruses were predominantly carried by people born overseas in high prevalence countries where HIV viral diversity is large and recombination is common. There was also evidence of intrasubtype recombination, predominantly among Australian-born people with subtype B virus.

7.12 Discussion

With the growing number of HIV infections imported from high prevalence countries, and a known increase in non-B and complex variants in Australia, accurate surveillance and identification of HIV subtypes is crucial.[69,182] Correct identification of HIV-1 subtypes is important both clinically and from a public health perspective to address issues such as

disease pathogenesis and transmissibility, sensitivity of diagnostic tools, vaccine development and clinical management, including drug resistance testing and appropriate treatment regimens.[77,196,295,296]. As described in Chapter Two, recombination occurs as a result of co- and super-infection with multiple viruses. These recombinant viruses are then spread through population mobility and can further recombine with one another. Accurate subtyping has become increasingly difficult because of the nature of this increased evolution and recombination of HIV-1.[182]

Phylogeny is the gold standard for subtyping, but it is not widely used to determine subtypes in clinical settings because of time constraints and the complexity of determining recombinant strains along a given fragment. Whole genome sequencing is the only true way a confident subtype can be assigned and full genomic sequencing techniques have improved over the years, with a growing number of whole genome reference sequences available for use in the public domain. However, it is often not practical or financially feasible to sequence the full genome, especially in LMICs. Historically, *gag* and *env* genes have been used to assign subtypes because these regions are highly variable, but viral subtyping is now commonly interpreted by using rapid online tools to test the *pol* sequences obtained following routine drug resistance testing.[193,297] The growing proportion of CRFs crossing international borders has, not surprisingly, led to the discovery of greater genetic complexity than previously thought,[203] with the new CRF and URF identifications occurring worldwide being monitored by LANL BLAST. It is now widely believed that many current HIV strains identified through partial genome sequencing may in fact be recognized as complex recombinant forms if full genome sequencing was performed.[295] For example, viruses first subtyped as clade G using *pol* sequences have now been identified as A/G recombinants using full genome sequencing ≤http://hiv.LANL BLAST.gov/content/hivdb/crfs/crfs.html≥. It is likely that the identification of genetic complexity will increase as whole genome

sequencing continues to evolve and becomes more cost effective.[295,298–300]

It is therefore imperative for accurate surveillance that subtyping tools are regularly updated and current reference sequences are accessed in order to detect variants. There are currently three main types of rapid online subtyping tools: similarity-based tools such as Stanford CPR,[187] statistical-based tools that use partial matching compression algorithms such as COMET[188] orHidden Markov Models such as jpHMM,[189] and phylogenetic-based tools such as REGA[184] and SCUEAL.[183].

Global surveillance figures for HIV-1 subtype distribution are influenced by the tool/s chosen to assign subtypes. Results gained from using a particular tool may not be replicated when a different tool is used, and with different tools in use worldwide it is difficult to ascertain true and accurate prevalence rates. Until a gold standard is accepted globally, molecular epidemiology surveillance reports should be interpreted with caution[182]

This is thought to be the first comprehensive comparative study to assess reliability of rapid HIV subtyping tools for global HIV infections diagnosed in Australia. It is also the first Australian study to characterize subtypes and ascertain the prevalence of genomic recombination based on *pol* and *env* genes.[192] The primary aims were to: 1) compare the performance of two statistical-based tools (COMET, jpHMM), two phylogenetic tools (SCUEAL, REGA) and two similarity-based tools (LANL BLAST, Stanford CPR) with phy using maximum likelihood in MEGA and an up-to-date reference sequence set (as of June 2013) from LANL BLAST ≤http://www.hiv.LANL BLAST.gov/≥,[183,187,199] and 2) identify the prevalence of B and non-B variants using devised online tool criteria to assess whether such criteria can be used as a gold standard for subtype surveillance in Australia.

7.12.1 Concordance between phy and online tools

Overall, concordance of subtype assignment between phy and online tools was low for *pol* sequences, with COMET and REGA (both 54%), BLAST (52%), Stanford CPR (37%),

SCUEAL (30%) and jpHMM (28%). The online tools with the highest concordance with *env* phy were COMET and BLAST (both 42%), followed by jpHMM (28%) and REGA (25%). REGA was unable to recognize CRFs or ISRs for the *env* region because of fragment size and jpHMM could not classify 14 of the 221 sequences because of issues with assigning subtypes at the 5' end of the HIV genome. When comparing online tools only, the concordance rate between them for *pol* and *env* sequences was 73% and 62% respectively and this concordance increased to 90% for both regions when the online tool criteria (5/6 tool concordance for *pol*, 3/4 for *env*) were implemented.

There was very high concordance (≥95%) between all tools and phy for *pol* and *env* subtype B sequences, followed by subtype C and 01_AE. The majority of cases in which there was low concordance between the tools were deemed to be carrying complex variants. Every online tool identified intragene recombination, but SCUEAL and COMET identified the highest proportion of *pol* and *env* intragene recombinants respectively. MEGA phy was superior in having the capability to calculate genetic distance and bootstrap support between query and reference sequences, but it could not identify complex variants and placed complex fragments near the most similar reference sequence, resulting in some incorrect subtype assignments. Future studies should identify possible recombinants using SCUEAL and exclude these from the dataset before MEGA phy is conducted, to minimize the risk of error when constructing the ancestral tree.

There were also sequences assigned as CRFs by MEGA phy that online subtyping tools assigned as variants similar to CRFs (e.g. B/C) rather than the CRF itself. This affected concordance rates between phy and the online tools. For example, *pol* sequences that clustered with CRFs 07_BC and 08_BC on the phylogenetic tree were assigned as 07_BC and 08_BC by REGA and COMET while the other online tools identified a B/C variant. Sequences that clustered with 23_BG on the phylogenetic tree were assigned by online tools

as B/G variants, while the *pol* sequence that clustered with 52_01B was subtyped as an A1/B or 01_AE/B variant by online tools. This happened because the 52_01B reference sequence had been typed as a combination of 01_AE and B at this region ≤http://www.hiv.LANL BLAST.gov/content/sequence/HIV/CRFs/CRFs.html≥. A number of sequences were assigned CRF01_AE by some tools while MEGA phy and other tools, especially SCUEAL, assigned 15_01B, and this is discussed later in the chapter. The same occurred for the *env* region. Sequences clustered with phy references 36_cpx and 45_cpx but were assigned as A1 or 'check for 02_AG' by online tools. The *env* region for both these reference sequences is a combination of A and 02_AG. Conversely, sequences clustering with 47_BF in the phylogenetic tree were all assigned subtype B by online tools, but the gp41 region of reference 47_BF is a combination of B and F ≤http://www.hiv.LANL BLAST.gov/content/sequence/HIV/CRFs/CRFs.html≥.

A partial explanation for the differences in subtyping is that not all online tools use up-to-date reference sequences or reference sequences from the same repository (the most common reference dataset is taken from LANL BLAST). The methods/algorithms used by each tool may also affect their ability to distinguish certain subtypes and CRFs from one another for that particular genome region. Maybe add something in here about the differences in algorithms (what other people have found).

Definitions of relationships between certain subtypes and CRFs can also lead to difficulties with accuracy of online tools. For example, researchers, including those who design online subtyping tools, have not reached consensus about the relationship between 02_AG and the parental strains A and G.[183,189] The correct assignment of several CRFs has been also disputed. Some regions thought to be recombinant may be so small that the phylogenetic signal is too weak, or many of the reference sequences are so similar in the particular gene region of interest that the tool assigns multiple reference sequences to a

region. It has also been suggested recently that re-analysis of current reference CRFs in the LANL BLAST database is needed, because the absence of breakpoints in certain regions can lead to misclassification. For example, 01_AE can be classified as 15_01B,[301] and G can be misclassified as CRF14_BG.[203]

When broken down by individual subtype or CRF, identification of subtype A1 was not uniform across the online tools and phy for the *pol* region, with some assigning A1 (jpHMM, COMET, MEGA phy) while others assigned A/01_AE ISRs, or more complex variants such as an A1/A-ancestral/A3 unique variant (Stanford CPR, SCUEAL). For the *env* region, jpHMM, REGA, COMET and phy were concordant in typing A1 sequences, but LANL BLAST assigned some sequences as A2, A3 or A1/D ISR. Subtype A has the second largest prevalence rate worldwide after subtype C. There are four sub-subtypes (A1, A2, A3, A4), each with different prevalence rates and geographic distribution. A1 variants are most common, having split into two distinct evolutionary lineages that have been transmitted worldwide. One line is from east central Africa and the other line is known as an IDU variant which circulates through the former Soviet Union.[302] There are scant data on subtypes A2–A4, but they are thought to circulate predominantly in Africa. A study in August 2015 found co-circulation of an A3/02_AG recombinant in Guinea-Bissau which is thought to lead to faster disease progression. In this study, most online tools and phy were unable to identify sub-subtypes of A, with Stanford CPR incorrectly identifying some A sequences as 01_AE. However, SCUEAL was quite good at distinguishing between the sub-subtypes, demonstrating that sequences previously assumed to be subtype A sequences were actually a diverse combination of A variants known to circulate in Africa. Further mosaic analysis of these sequences is warranted.[263]

All tools except jpHMM and SCUEAL were 100% concordant with phy for *pol* subtype D sequences, and both REGA and COMET were 100% concordant for *env*. jpHMM and

SCUEAL assigned one of the *pol* sequences as a complex variant and jpHMM and LANL BLAST did so for one of the *env* sequences. These sequences are discussed in the next section.

All six tools had high concordance with phy and each other for both *pol* and *env* assigned subtype B and C. However, jpHMM identified two *pol* subtype C sequences as subtype B and one as a C/B ISR, and LANL BLAST identified a fourth *pol* subtype C sequence as being most similar to an A1/C ISR. In the *env* region, jpHMM could not classify four subtype C sequences, and LANL BLAST identified one as being most similar to a B/C ISR. Further analysis of these sequences is warranted, using breakpoint locations to cut the sequences before analysis with MEGA phy to determine whether these are recombinant viruses or possibly co-infections.

COMET, REGA and LANL BLAST were concordant for the *pol* sequences that clustered with 07_BC and 08_BC on the phylogenetic tree, while jpHMM, Stanford CPR and SCUEAL identified B/C variants. COMET and LANL BLAST also identified these CRFs for *env*. Two different BC recombinants have been detected in intravenous drug users (IDU) in China. CRF07_BC and 08_BC were both isolated among IDUs in northwestern and southern China respectively.[261,303] The two parental strains B and C have been reported to co-circulate among IDUs in southwestern China, though the parental C strain was likely to have been imported from India around the early 1980's.[304] The origin of 07_BC and 08_BC can be traced back to Yunnan,[304] and two different routes of BC recombinants then spread throughout China, suggesting different founder effects in the Chinese IDU population.[261]

For CRF01_AE assigned sequences, REGA, LANL BLAST and COMET had high concordance with *pol* phy while jpHMM and SCUEAL had the lowest concordance because of the identification of complex variants. For *env*, jpHMM, COMET and LANL

BLAST had high concordance while REGA had none because of its inability to detect CRFs in short fragments. Two *pol* and two non-related *env* sequences clustered with 15_01B on the phylogenetic tree, but only SCUEAL assigned the *pol* sequences as 15_01B or a 01_AE/B variant while the other tools assigned 01_AE. The *env* sequences were typed as B by the online tools, which is the correct subtype for that part of the 15_01B reference sequence ≤http://www.hiv.LANL BLAST.gov/≥. However, SCUEAL, COMET and LANL BLAST did assign other *pol* and *env* sequences as 15_01B that were assigned as 01_AE (*pol*) or B (*env*) by phy and other online tools. Subtype B and 01_AE have co-circulated in Thailand and southeast Asia for over 15 years, although initially 01_AE was limited to IDU transmission and subtype B to heterosexual transmission.[305] Over time, the two strains have become intermixed and recombinant strains have formed, such as 15_01B, 33_01B, 48_01B, 52_01B, 53_01B, 54_01B and 58_01B, with these new viral variants freely circulating between the two high-risk populations.[13,306] It is now believed some of these CRFs are actually second generation recombinant descendants from initial CRFs, 48_01B was discovered in 2007 and found to be a CRF descended from the CRF33_01B lineage.[13,266] It is extremely plausible to assume that as the proportion of new CRF infections increase over time, so will viral complexity.[13,52,307]

The 15_01B strain is thought to constitute approximately 1.7% of HIV-1 infections across southeast Asia.[305] Sequences that were identified as 15_01B by at least one tool in this study should be examined further to determine whether they are a 01_AE virus, a subtype B virus, or a combination of the two.

A recent molecular epidemiological study in China found that subtype B, 01_AE, and CRF07_BC are currently co-circulating among MSM in China. These recombinants share a number of breakpoints and highlight the dynamic change of HIV and the importance of using full genome sequencing with MEGA phy and accurate subtyping tools to identify new

recombinant forms.[308]

Only REGA, Stanford CPR and COMET showed high concordance with phy for 02_AG clustered sequences. jpHMM and SCUEAL both showed no concordance, assigning each of the sequences as more complex variants that contained 02_AG, A and G along the two gene fragments. For *env*, only LANL BLAST had partial concordance while REGA, jpHMM and COMET all had no concordance, with the latter two identifying A, G or complex variants.

The inconsistency between these particular sequences is less likely to have occurred because of the shortcomings of the methods behind the tools and is more likely to result from trying to assign known viral types to highly divergent strains. This problem is especially magnified when a range of divergent subtypes co-circulate and the risk of co-infection/super-infection is high, as is the case for subtypes A, G and CRF02_AG variants in Africa.[7,309,310] It is thought that many recombinant events have occurred among these strains which have resulted in the merging of several lineages into a single diverse group.[193] This is supported by the *pol* and *env* phylogenetic trees. Although each of the sequences was genetically closest to 02_AG as determined by branch length, the sequences were not all part of the same monophyletic cluster.

An African study reported that the majority of *pol* 02_AG sequences in a cohort were actually recombinants comprised of several 02_AG strains. It has also been shown with *env*-gp41 and *pol*-PR/RT regions that these strains do not have a strong phylogenetic relationship.[265] It is therefore possible that *pol* genes have evolved from ancestral sequences and are undergoing complex recombination processes which create an array of strains that complicate its use for subtyping. The current data support this view.

The role of non-B variants, especially complex recombinants, in drug resistance (both primary and secondary) and disease progression is still evolving and it is therefore important to understand and accurately identify complex mosaics.[265]

7.12.2 Recombinant viruses in South Australia

Just over 85% of cases carried virus with subtype/CRF concordant *pol* and *env* regions, with the majority of cases carrying subtype B followed by subtype C, 01_AE, A1 and D. Further analysis of the remaining 14% showed *env* sequences were different to the matching *pol* sequences, supporting previous subtyping studies that used these two regions.[168] These 30 cases were deemed to carry an intergene recombinant strain. Three cases were assigned discordant pure subtypes at the *pol* and *env* region (A1/B, A1/D and A1/C). The remaining 27 cases carried a diverse range of complex recombinant forms that were not identified using previously defined CRFs. The complex mosaic structures were linked to region of birth and location the infection was acquired.

All but three of the complex viruses containing various 02_AG, G and A1 fragments were carried by African-born people who reported being infected overseas (one MTCT, others heterosexual) and who were diagnosed in South Australia between 2006 and 2011. Two people reported acquiring the infection in Australia, in 2010 and 2011 respectively. Three of the 27 cases were Australian-born people diagnosed with complex 02_AG/A1/G type virus. One female reported being infected overseas by heterosexual contact in 2004, while two Australian males reported MSM contact within Australia, diagnosed in 2009 and 2010 respectively. The three most recent diagnoses of these AG/A/G recombinants were in African-born and Australian-born people who became infected in South Australia. This is cause for concern because these infections suggest the start of local transmission of this complex virus, which may be more pathogenic than 02_AG alone and more readily transmitted than other viruses.[80,264]

There were two cases carrying complex D type virus; an African-born male diagnosed in 2004 who acquired an A/D complex virus overseas through heterosexual contact and an

MSM diagnosed in 2001 with region of birth unlisted who acquired a B/D complex virus within Australia. SCUEAL assigned the *pol* sequence of the A/D virus as A1/A-ancestral/A3. Subtypes A1 and D are the most common circulating subtypes in Kenya, and in 2004 a study found that the virus in patients co-infected with subtypes A and D generated and selected for A/D recombinant forms with A/D recombinants accounting for nearly half of all reported recombinant cases.[311] As mentioned previously, a recombinant of sub-subtype A3 and 02_AG has recently been implicated in faster disease progression in West Africa, and the characteristics of viruses that recombine have included higher potential for dispersal in the human population. This highlights the absolute necessity of continuous screening and surveillance of the HIV-1 epidemic.[80,264]

The initial introduction of HIV-1 into developed countries such as the US and Australia involved the importation of a single subtype B strain which spread rapidly through MSM communities.[312] Subtype D has been found to be closely related to subtype B and it has been suggested that it may be more appropriate to class subtype D viruses as sub-subtype B2.[51,313] Phy of subtype B and D sequences suggests the genetic diversity of 'D-like' viruses may be underrepresented, and that the D group of viruses may actually incorporate viruses that are intermediate between B and D subtypes. In the present study, some sequences were assigned subtype B by one tool but subtype D or a B/D recombinant by another. Given the similarity between these two types, the sequences identified in this study may be a representation of a new category of 'B/D like' viruses rather than a novel recombinant of the two types.[312]

There were two complex 01_AE/B cases, both European-born males. One reported an overseas medical procedure while the other reported heterosexual contact with IDU risk within Australia. Two cases were identified as carrying complex virus containing 07_BC, both acquired overseas (2001 and 2010) but only one person listed region of birth (Asia). Both cases reported heterosexual contact with IDU risk. It is likely all these viruses

originated from southeast Asia. This is consistent with the known transmission risk of 01_AE/B and B/C variant viruses that widely circulate in Asia in both the heterosexual and IDU populations.

The final two complex cases were children who were mentioned in the previous chapter as part of a cluster of three children all infected in the same region of Central Asia. Both children included in this subtyping study were determined to have a complex '02_AG like' virus. SCUEAL identified one of the *pol* regions as a G/A variant and the other as a 02_AG/A variant, and LANL BLAST classified both *env* regions as CRF63_02A1, identifying 97% similarity with a reference sequence from Uzbekistan. CRF63_02A1 is a recombinant virus of the 02_AG virus circulating through IDU in central Asia, and subtype A virus circulating through IDU in the former Soviet Union.[314] It was thought to have emerged around 2006, with rapid circulation around the Eastern Russian/Chinese border including east Kazakhstan and Uzbekistan between 2008 and 2009, which is consistent with the times (2009 and 2010) and location (Central Asia) these children were infected. Iatrogenic infection in central Asia is well documented and it is likely that these children were infected with this recombinant virus from contaminated blood or medical equipment.[62]

The data indicate that complex mosaic forms of HIV are being imported and the predominant population harboring these complex viruses are males and females infected in high prevalence African countries where co-infection with divergent strains occurs frequently.[193] There is also evidence of spread into the Australian-born population through overseas travel and local transmission.[168] From this subset of South Australian diagnosed cases, two broad mosaics have been locally transmitted to date: complex '02_AG like' strains originating from Africa and transmitted through MSM or heterosexually, and 01_AE/B or B/D like strains of unknown origin most likely transmitted through IDU. It is very likely that some of these complex variants will mature into CRFs, especially in the

presence of high-risk behaviour.[168] Complex recombinant viruses are especially prevalent in populations where multiple subtypes and CRFs co-circulate, such as sub-Saharan Africa.[7] With increasing geographic mobility, including migration from high prevalence African countries to Australia and Australian-born people travelling overseas, there is a corresponding increase in new HIV non-B diagnoses including these complex variants. This has also occurred in Spain, where the proportion of non-B variants has increased because of growing number of immigrants from sub-Saharan Africa.[295]

This growing global genetic diversity of intra- and inter-subtype CRFs indicates these complex mosaics may possess a possible fitness advantage compared with pure subtype strains in certain settings.[315–317] Research by Brown and colleagues also found recombinant viruses impact on phenotypic traits differently from parental strains, which may facilitate viral immune evasion, development of resistance to treatment, and have further reaching impacts on disease progression.[318]

It is crucial therefore that these non-B variants are reported accurately, which means improving access to testing, timely genotyping and regular updating of online reference repositories and subtyping tools.

Subtype distribution can be relatively stable within a geographic location if opportunities for recombination between diverse strains are rare, and historically this has been the case in Australia where subtype B predominates. However it is well established that intrasubtype recombination is a frequent occurrence within subtypes, especially in the presence of drug therapy.[193] The SCUEAL tool identified an additional 16 cases as carrying intrasubtype recombinant virus including 14 subtype B viruses, all diagnosed before 2008 when non-B variants were relatively uncommon. All but one of these subtype B viruses were acquired within Australia, and all were carried by Australian-born males who acquired the infection predominantly through MSM. One Australian-born male acquired his B infection overseas.

Two African-born females were also identified as carrying intrasubtype C recombinant virus, both infected overseas, one through MTCT and one through heterosexual contact. LANL BLAST analysis found the *pol* intrasubtype C sequences were most similar to subtype C strains originating from Africa (≥95%). The two females were part of the same high reliability *pol* subtype C cluster (n=41), but not related by transmission. The female child was part of a *pol* transmission cluster of three, most likely a mother, father and child transmission, although the mother's and father's *pol* sequences showed no evidence of intrasubtype recombination. The female adult was not part of any transmission cluster. Intrasubtype recombination is extensive through subtype C[59] and high rates of intrasubtype dual infection B have been found among the MSM population in the United States.[319] This is not surprising given that subtype C is one of the oldest and most prevalent HIV infections circulating in Africa, and historically the Australian epidemic has almost completely been comprised of subtype B.

7.12.3 Online tool considerations

Several factors must be considered when using multiple online tools. These include the proportion of non-B sequences in the cohort, the proportion of recombinant viruses, especially non 01_AE and 02_AG, the algorithm and methods used, the rate of false identification of B and non-B variants, and the reference sequences used. The high recombination rate of HIV combined with increasing population mobility also affects the performance of subtyping tools, the former by increasing the complexity and number of newly forming recombinants and the latter by introducing new recombinant strains and pure subtypes into different geographical areas and population groups.[216]. This means that reference sequences attached to subtyping tools should be regularly updated, and the algorithm and methods behind the tools should be regularly assessed as complexity grows, especially for similarity and statistical tools which were not designed to include intrinsic

biologically relevant evolutionary relationships.[183,184,188,203] The current study was an epidemiological and public health based explanatory analysis of subtype distribution using multiple online tools. However though exploratory analysis of the mathematical and algorithmic differences between the online was beyond the scope of this study, it would be a very worthwhile endeavor to assess the impact on subtype assignation. Other factors to consider when using subtyping tools are the operational characteristics, such as the time taken to analyze a whole sequence set. Phylogenetic-based tools require more time than statistical-based tools [183,188] and when analyzing a very large dataset this can be important. Only phylogenetic-based tools produce data on recombination breakpoints, which may be important when trying to ascertain whether the fragment is an URF.[183,320]

A study by Pineda-Pena *et al*. in 2013 found that no online tool could accurately subtype 100% of the time. Each tool performed slightly differently because of the method and algorithm behind it, its access to reference sequences and the region being typed. The findings from the current study support this. Phylogenetic-, statistical- and similarity-based tools worked well for identification of frequent subtypes such as B and C and CRF01_AE.[182,192,199,203] CRF02_AG was overrepresented by phy and some online tools but the statistical- and phylogenetic-based online tools were able to classify them as more complex forms. The best online tools for analysis of epidemics where there are many subtypes, CRFs and likely URFs being imported or circulating in the population are COMET or SCUEAL, and the latter has the most up-to-date reference dataset.[33,183,188,193,320,321]

REGA v.3 performed equally well as COMET and jpHMM, and was designed to identify most of the epidemic CRFs.[203] In this study the concordance between REGA and COMET was very high for all the pure subtypes and almost all of the CRFs, with both tools identifying a similar number of ISRs. When REGA could not confidently assign a specific subtype or CRF, it reported a 'like' subtype/CRF,[182] which alerted the user to examine these sequences

further.[203]

The Stanford database CPR tool is most widely used by clinicians for routine drug resistance testing and subtyping. When compared with phy and other online tools, Stanford CPR was very good at identifying subtype B, C, D and CRF01_AE. It showed perfect concordance with phy for assigning 02_AG, but other online tools showed these sequences to be more complex variants. The Stanford CPR tool does not have the capability to identify CRFs other than 01_AE and 02_AG, because it only uses 01_AE, 02_AG and pure subtype reference sequences. The CPR tool was able to identify different subtypes along the *pol* region such as B/C, but could not identify more complex variants or sub subtypes of A.

In the 513 newly diagnosed cases with routine genotype tests in South Australia between 2000 and 2012, the Stanford CPR tool detected 20 (4%) *pol*-intergene recombinants. Thirteen of these Stanford CPR ISR *pol* sequences were included in the present study of 221 cases, comprising 6% of total cases. Compared with Stanford CPR, 50 cases were identified as ISRs by at least one of the online tools, which increased the prevalence from 6% to 23%. A conservative estimate was then calculated using the online tool criteria, and a final prevalence rate of 12% was inferred. Although this did not reach statistical significance when compared with using the CPR tool alone, it was still double the CPR prevalence rate and suggests that CPR should not be used to infer subtype in populations with complex genetic diversity.

SCUEAL and COMET tools were considered superior at identifying complex recombinant forms, although SCUEAL's identification of 'complex' sequences are thought by some to be misclassifications, such as subtype B sequences being assigned as B/D recombinants because of the high similarity between the two pure subtypes in the *pol* region.[322] However, SCUEAL has been tested against a number of common online tools and was found to be superior at determining subtype, identifying inter and intra subtype

recombination, and mapping recombinant structures.[183] COMET was found to be equivalent to SCUEAL in identifying recombinants and excelled in identifying some known subtypes and CRFs.[188]

In summary, misclassification of HIV-1 subtypes and CRF strains can occur when using online tools or MEGA phy, and there are serious limitations on identification of complex variants that have epidemiological and clinical consequences. To report meaningful and accurate epidemiological information at a global level, a critical understanding of methods used by the subtyping tools is needed to control for error and misclassification.[192,193]

7.12.4 Considerations of partial versus full genome sequencing

In the present study, analysis of the PR and RT regions was conducted in one combined *pol* fragment.[192] Given that recombination was found in the *pol* region of a number of cases, splitting these regions and subtyping them separately may have led to clearer subtype assignations for some online tools and phy. However the PR/RT sequences were combined into one *pol* sequence because short fragments cannot always be confidently assigned subtypes.[192] Indeed, REGA and jpHMM were unable to assign some *env* fragments because of length.

To date, HIV subtyping in Australia has used the *pol* gene only. A recent study in Victoria found a non-B prevalence rate of 22% using the *pol* region, and identified six different CRFs and 12 URFs in a cohort of ~1000 people using Stanford CPR, REGA v2 and NCBI. Sequencing of two gene regions identified more variation than using the *pol* region alone. This supports reports from India in which an increased proportion of CRFs and URFs was noted when two genes were sequenced instead of one, with even more variation when three genes were sequenced.[168] In a Spanish cohort, subtyping of multiple genes using three rapid subtyping tools also found recombinant forms of HIV involving a variety of fragments that were different subtypes and CRFs with distinct breakpoints. The researchers

also found discrepant results between tools.[295]

To account for potential overestimation of non-B variants in the current study, online tool criteria were set conservatively. In fact, these criteria and the use of two gene regions rather than full genome sequencing may have led to an underestimation of variants. The *env* and *pol* regions were found to be discordant in more than 10% of cases, with the majority of them being quite complex recombinants. This was double the number reported from the first study in Chapter Three (see Appendix Two). These findings have implications for the way viruses are currently subtyped in South Australia. Currently, this relies predominantly on Stanford CPR and it is likely that further analysis of other genomic fragments or the entire genome would uncover even greater diversity[267,267,284,323] given that full genome sequencing has uncovered major heterogeneity of HIV in other parts of the world.[295] It is recommended therefore that full genome sequencing be conducted in South Australia to assess whether there is greater diversity in this population than shown in results based on two subgenomic fragments.

7.12.5 Strengths and Limitations

The study has some limitations. A subset of approximately half of all new diagnoses was analyzed because of the unavailability of samples originally used to sequence the *pol* region at time of diagnosis. This led to a number of subtype B cases diagnosed between 2000 and 2007 not being included, which may have slightly altered the prevalence rates of non-B variants and intrasubtype B cases. However, it was considered that this subset included fairly representative numbers of B and non-B variants diagnosed during the time period.[168]

Some online tools did not use up-to-date reference sequences and this affected correct assignation. However, MEGA phy used at least one reference sequence for each of the pure subtypes and CRFs circulating at the time the laboratory analysis was undertaken, downloaded from a frequently used and reliable source of full genome sequences

≤http://www.hiv.LANL BLAST.gov/≥. The SCUEAL tool also uses the most up-to-date reference subtypes and CRFs by linking directly with the LANL BLAST reference database.

Not all online subtyping tools were assessed and the inclusion of a wider selection may have produced different subtype results. However, the six online tools chosen are widely used and were drawn from similarity-based, phylogenetic-based and statistical-based tools. REGA, a statistical-based tool, has been recently updated and found to be accurate and comparable with SCUEAL and jpHMM.[203]

A major strength of this study was the incorporation of demographic and clinical information. This allowed comprehensive characterization of subtype B and non-B variants, especially complex variants, in order to identify origin of infection and determine whether these variants are circulating locally.[168]

7.12.6 Recommendations

In summary, the online tool criteria provided a reliable and conservative method with which to subtype a large and diverse HIV dataset and identify complex variants. A number of online tools can now classify multiple gene regions, although robust tools such as SCUEAL can currently only assign *pol*.[183,193] To ensure accurate HIV-1 subtype surveillance, there must be a gold standard subtyping protocol that uses tools which are capable of distinguishing genetic variants. In this study, use of the online tool criteria led to the unambiguous assignation of 90% of sequences to previously identified subtypes and CRFs, and using two gene fragments identified some as intergene recombinants. This high concordance was largely because of the consistency between tools for assigning subtypes B and C, which were the predominant clades among this cohort. However, the consistency of subtype assignment across tools was low for most other subtypes and CRFs with the exception of CRF01_AE. Discordance between tools occurred for three main reasons: 1) subtypes were highly similar within the sequenced region (i.e. B and D), 2) the region

examined contained a clade which is found at that particular region for multiple subtypes and CRFs (i.e. the gp41 region of *env* for A and 01_AE) and 3) the virus was a unique complex variant. It is also possible that some cases may have been co- or super-infections. It is recommended that sequences should be assessed in more detail with MEGA phy by using identified SCUEAL and jpHMM breakpoints, and comparing fragments with historical or newer sequences for the same patient.[297]

To control for the issues listed above and ensure the most robust assignation, it is recommended that more than one type of subtyping tool is used (phylogenetic, statistical and similarity) with conservative online tool criteria. It is very important that a wide variety of B and non-B sequences are added to online reference sets, and algorithms for online tools are continually improved to improve the prediction of correct variants, especially more complex forms.[192] Although time consuming, MEGA phy should also be used where possible, and this is still considered the gold standard for short sequences such as the gp41 region of *env* because of problems with low phylogenetic signal which can lead to incorrect classification. Although *pol* data derived from routine resistance testing provides a valuable opportunity for surveillance of HIV-1 genetic diversity, particularly in LMICs where resources are limited, it is strongly recommended that multiple genomic regions are sequenced or full genome sequencing be performed where possible.

7.12.7 Conclusion

This is thought to be the first study to complete an extensive comparison of subtyping tools and MEGA phy for the *pol* and *env* regions of an Australian dataset. This study has identified a higher rate of HIV-1 genetic diversity circulating in the South Australian population than previously described,[66] predominantly originating from different regions of Africa. By sequencing of *pol* and *env* genes and using multiple online subtyping tools and MEGA phy the study uncovered complex HIV variants that could not be linked to known

CRFs. These findings are in line with previous studies that have sequenced multiple genes [168] and assessed multiple online subtyping tools.[192]

Identification of genetic diversity in LMICs is crucial given the high rate of CRFs and continued recombination, which is increasingly spreading to regions such as Australia. Accurate surveillance will improve research efforts into strains widely circulating in high prevalence areas. However, accurate surveillance reports need accurate subtyping methods and a gold standard for confident and conservative assignment of subtypes, with an emphasis on identifying complex mosaics and CRFs rather than forcing sequences into groups that are most closely related. This will lead to the identification and epidemiological tracking of novel HIV-1 variants globally.

The sequencing of multiple gene regions and subtyping using an array of subtyping methods including MEGA phy will lead to more accurate representation of B and non-B variants diagnosed and circulating in South Australia. It is likely that full genome sequencing will reveal an even bigger increase in viral diversity. It is important that those conducting typing studies ensure that they are using the most up-to-date CRFs from LANL, and have an understanding of prevalent subtypes and CRFs in the region they are examining. This is important for both clinical and epidemiological reasons.[295] For clinical management of individual people, Stanford CPR is still the best tool to ascertain genotypic drug resistance quickly and it should remain the standard tool for resistance profiling.[182]

The recommendations from this study should be considered for clinical and surveillance purposes in South Australia. A gold standard subtyping protocol would pave the way for global accuracy of subtype and CRF surveillance. The combination of genetic and demographic information will inform public health departments and clinicians about how these variants are being transmitted, their prevalence rates and disease pathogenesis, and how they respond to treatment. This knowledge is crucial for the implementation and

monitoring of prevention and intervention strategies.[168]

CHAPTER 8: CONCLUSION

This thesis presents a comprehensive molecular characterization of HIV strains in people newly diagnosed with HIV in South Australia between 2000 and 2013. Molecular epidemiological and drug resistance analysis of all new diagnoses between these time points revealed an increasing genetic diversity of HIV strains being diagnosed in South Australia and beginning to circulate locally. Conversely the prevalence of TDR is decreasing in the South Australian community, with very few K103N resistant strains being recently diagnosed. However the overall TDR rates in both subtype B and non-B populations are still considered high and moderate respectively by WHO standards.

Phylogenetic and online tool analysis of the *pol* and *env* regions identified a larger proportion of recombinant viruses than were identified with the genotypic resistance tool used in routine surveillance. This included the identification of possible unique recombinant forms. The use of multiple subtyping tools with high specificities and sensitivities, and phylogeny, is recommended in order to accurately assign subtypes and CRFs, especially with the growing viral diversity occurring worldwide. Using both the *pol* and *env* region identified a larger number of high reliability and transmission pairs/clusters than only using one region alone. Subtype C and CRF01_AE infections were more likely to be found in high reliability *pol* and *env* clusters respectively, most likely due to evolutionary lineage pathways. Phylogeny uncovered heterosexual transmission pairs carrying non-B subtypes predominated, but there were also family clusters, and male/male and male/female subtype B pairs and clusters.

The data generated from these studies will be used in national research assessing subtype distribution circulating within Australia and will provide valuable information about the origin of HIV strains, transmission routes, local forward transmission of non-B subtypes, and identification of TDR prevalence among treatment naïve individuals.

REFERENCES

1. Zimmer C. The Lurker: How a virus hid in our genome for six millions years. PHENOMENA: A science salon hosted by National Geographic Magazine [Internet]. 2013 May 10; Available from: http://phenomena.nationalgeographic.com/2013/05/10/the-lurker-how-a-virus-hid-in-our-genome-for-six-million-years/

2. Pepin J. The origins of AIDS. Cambridge University Press; 2011.

3. Oelrichs R. The subtypes of human immunodeficiency virus in Australia and Asia. Sex Health. 2004;1:1–11.

4. Ndung'u T, Weiss RA. On HIV diversity: AIDS. 2012 Jun;26(10):1255–60.

5. Plantier JC, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemee V, et al. A new human immunodeficiency virus derived from gorillas. Nat Med. 2009 Aug;15:871–2.

6. Lihana RW, Ssemwanga D, Abimiku A, Ndembi N. Update on HIV-1 diversity in Africa: a decade in review. AIDS Rev. 2012;14(2):83–100.

7. Peeters M, Jung M, Ayouba A. The origin and molecular epidemiology of HIV. Expert Rev Anti Infect Ther. 2013 Sep;11(9):885–96.

8. Butler IF, Pandrea I, Marx PA, Apetrei C. HIV genetic diversity: biological and public health consequences. Curr HIV Res. 2007;5(1):23–45.

9. Lai A, Riva C, Marconi A, Balestrieri M, Razzolini F, Meini G, et al. Changing patterns in HIV-1 non-B clade prevalence and diversity in Italy over three decades. Hiv Med. 2010 Oct;11:593–602.

10. Russell JS, Chibo D, Kaye MB, Gooey ML, Carolan LA, Papadakis A, et al. Prevalence of transmitted HIV drug resistance since the availability of highly active antiretroviral therapy. Commun Dis Intell Q Rep. 2009;33(2):216–20.

11. Palmer C. HIV treatments and highly active antiretroviral therapy. Aust Prescr. 2003;26(3):59–61.

12. Ammaranond P, Cunningham P, Oelrichs R, Suzuki K, Harris C, Leas L, et al. Rates of transmission of antiretroviral drug resistant strains of HIV-1. J Clin Virol. 2003 Feb;26:153–61.

13. Cheong HT, Chow WZ, Takebe Y, Chook JB, Chan KG, Al-Darraji HAA, et al. Genetic Characterization of a Novel HIV-1 Circulating Recombinant Form (CRF74_01B) Identified among Intravenous Drug Users in Malaysia: Recombination History and Phylogenetic Linkage with Previously Defined Recombinant Lineages. Paraskevis D, editor. PLOS ONE. 2015 Jul 21;10(7):e0133883.

14. Easterbrook PJ, Smith M, Mullen J, O'Shea S, Chrystie I, de Ruiter A, et al. Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy.

J Int Aids Soc [Internet]. 2010 Feb;13. Available from: ://WOS:000296338900001

15. Singh K, Flores J, Kirby K, Neogi U, Sonnerborg A, Hachiya A, et al. Drug Resistance in Non-B Subtype HIV-1: Impact of HIV-1 Reverse Transcriptase Inhibitors. Viruses. 2014 Sep 24;6(9):3535–62.

16. Hogg RS, Bangsberg DR, Lima VD, Alexander C, Bonner S, Yip B, et al. Emergence of drug resistance is associated with an increased risk of death among patients first starting HAART. PLoS Med. 2006;3(9):e356.

17. Derache A, Traore O, Koita V, Sylla A, Tubiana R, Simon A, et al. Genetic diversity and drug resistance mutations in HIV type 1 from untreated patients in Bamako, Mali. Antivir Ther. 2007;12:123–9.

18. Pham QD, Wilson DP, Law MG, Kelleher AD, Zhang L. Global burden of transmitted HIV drug resistance and HIV-exposure categories: a systematic review and meta-analysis. AIDS. 2014 Nov 28;28(18):2751–62.

19. Hamers RL, Wallis CL, Kityo C, Siwale M, Mandaliya K, Conradie F, et al. HIV-1 drug resistance in antiretroviral-naive individuals in sub-Saharan Africa after rollout of antiretroviral therapy: a multicentre observational study. Lancet Infect Dis. 2011;11(10):750–9.

20. Wainberg MA. HIV-1 subtype distribution and the problem of drug resistance. Aids. 2004 Jun;18:S63–8.

21. Guy R, Lim MSC, Wang YH, Medland N, Anderson J, Rothe N, et al. A new surveillance system for monitoring HIV infection in Victoria, Australia. Sex Health. 2007 Sep;4:195–9.

22. Kuiken C, Thakallapalli R, Eskild A, de Ronde A. Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. Am J Epidemiol. 2000 Nov;152:814–22.

23. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull. 2001;58(1):19–42.

24. Cichutek K, Merget H, Norley S, Linde R, Kreuz W, Gahr M, et al. Development of a quasispecies of human immunodeficiency virus type 1 in vivo. Proc Natl Acad Sci. 1992;89(16):7365–9.

25. Smyth RP, Davenport MP, Mak J. The origin of genetic diversity in HIV-1. Virus Res. 2012 Nov;169(2):415–29.

26. Domingo E, Sheldon J, Perales C. Viral Quasispecies Evolution. Microbiol Mol Biol Rev. 2012 Jun 1;76(2):159–216.

27. Bar KJ, Li H, Chamberland A, Tremblay C, Routy JP, Grayson T, et al. Wide variation in the multiplicity of HIV-1 infection among injection drug users. J Virol. 2010;84(12):6241–7.

28. Li H, Bar KJ, Wang S, Decker JM, Chen Y, Sun C, et al. High multiplicity infection

by HIV-1 in men who have sex with men. PLoS Pathog. 2010;6(5):e1000890.

29. Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, Karita E, et al. Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. PLoS Pathog. 2009;5(1):e1000274.

30. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci. 2008;105(21):7552–7.

31. Abrahams M-R, Anderson J, Giorgi E, Seoighe C, Mlisana K, Ping L-H, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. J Virol. 2009;83(8):3556–67.

32. Hunter E. HIV Transmission – Lessons from Heterosexual Couples in Africa. In: Proceedings of the 2015 Australasian conference on HIV&AIDS. Brisbane, Australia; 2015.

33. Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med. 2012 Mar;18(3):182–92.

34. Kestens L. HIV variability [Internet]. Escort. 2005 [cited 2015 Sep 28]. Available from: http://www.itg.be/internet/e-learning/written_lecture_eng/8_hiv_variability.html

35. Yerly S, Jost S, Monnat M, Telenti A, Cavassini M, Chave J-P, et al. HIV-1 co/super-infection in intravenous drug users. AIDS. 2004;18(10):1413–21.

36. Blackard JT, Cohen DE, Mayer KH. Human immunodeficiency virus superinfection and recombination: current state of knowledge and potential clinical consequences. Clin Infect Dis. 2002;34(8):1108–14.

37. Osborn KG, Prahalada S, Lowenstine LJ, Gardner MB, Maul DH, Henrickson RV. The pathology of an epizootic of acquired immunodeficiency in rhesus macaques. Am J Pathol. 1984;114(1):94–103.

38. Clavel F, Guyader M, Guétard D, Sallé M, Montagnier L, Alizon M. Molecular cloning and polymorphism of the human immune deficiency virus type 2. Nature. 1986;324:691–5.

39. Chahroudi A, Bosinger SE, Vanderford TH, Paiardini M, Silvestri G. Natural SIV Hosts: Showing AIDS the Door. Science. 2012 Mar 8;335(6073):1188–93.

40. Los Alamos National Laboratory. HIV-1 Gene Map [Internet]. HIV Sequence Database. 2014 [cited 2015 Sep 28]. Available from: http://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

41. Fujita M, Nomaguchi M, Adachi A, Otsuka M. SAMHD1-Dependent and -Independent Functions of HIV-2/SIV Vpx Protein. Front Microbiol. 2012;3(297):1–7.

42. Lengauer T, Sing T. Bioinformatics-assisted anti-HIV therapy. Nat Rev Micro. 2006 Oct;4(10):790–7.

43. Freeman S, Herron JC. Evolutionary analysis [Internet]. Pearson Education Upper Saddle River (NJ); 2004 [cited 2014 Feb 14]. Available from: http://uncw.edu/people/bruce/hon%20210/pdfs/mccartneychaps.pdf

44. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004 Jan;5(1):52–61.

45. Kim JH, Hartley TL, Curran AR, Engelman DM. Molecular dynamics studies of the transmembrane domain of gp41 from HIV-1. Biochim Biophys Acta BBA - Biomembr. 2009 Sep;1788(9):1804–12.

46. Stages of HIV Infection [Internet]. AIDS.gov. 2015 [cited 2015 Nov 18]. Available from: https://www.aids.gov/hiv-aids-basics/just-diagnosed-with-hiv-aids/hiv-in-your-body/stages-of-hiv/

47. Zhu TF, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998 Feb;391:594–7.

48. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature. 2008;455(7213):661–4.

49. de Sousa JD. Pandemic HIV-1: Its Old Origin and Overlooked Mysteries. Aids Rev. 2009;11:52–52.

50. Sharp PM, Hahn BH. Origins of HIV and the AIDS Pandemic. Cold Spring Harb Perspect Med. 2011 Sep 1;1(1):a006841–a006841.

51. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, et al. HIV-1 nomenclature proposal. Science. 2000;288(5463):55–55.

52. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global trends in molecular epidemiology of HIV-1 during 2000–2007: AIDS. 2011 Mar;25(5):679–89.

53. Frost SD, Wrin T, Smith DM, Pond SLK, Liu Y, Paxinos E, et al. Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. Proc Natl Acad Sci U S A. 2005;102(51):18514–9.

54. Woo J, Robertson DL, Lovell SC. Constraints on HIV-1 diversity from protein structure. J Virol. 2010;84(24):12995–3003.

55. Travers SA, O'Connell MJ, McCormack GP, McInerney JO. Evidence for heterogeneous selective pressures in the evolution of the env gene in different human immunodeficiency virus type 1 subtypes. J Virol. 2005;79(3):1836–41.

56. Liu Y, Li L, Bao Z, Li H, Zhuang D, Liu S, et al. Identification of a Novel HIV Type 1 Circulating Recombinant Form (CRF52_01B) in Southeast Asia. AIDS Res Hum Retroviruses. 2012 Apr;28(10):1357–61.

57. Koelsch KK, Smith DM, Little SJ, Ignacio CC, Macaranas TR, Brown AJl, et al. Clade B HIV-1 superinfection with wild-type virus after primary infection with drug-resistant clade B virus. Aids. 2003;17(7):F11–6.

58. Taylor JE, Korber BT. HIV-1 intra-subtype superinfection rates: estimates using a structured coalescent with recombination. Infect Genet Evol. 2005;5(1):85–95.

59. Rousseau CM, Learn GH, Bhattacharya T, Nickle DC, Heckerman D, Chetty S, et al. Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. J Virol. 2007;81(9):4492–500.

60. Kiwelu IE, Novitsky V, Margolin L, Baca J, Manongi R, Sam N, et al. Frequent Intra-Subtype Recombination among HIV-1 Circulating in Tanzania. Liang C, editor. PLoS ONE. 2013 Aug 5;8(8):e71131.

61. Kouri V, Khouri R, Alemán Y, Abrahantes Y, Vercauteren J, Pineda-Peña A-C, et al. CRF19_cpx is an evolutionary fit HIV-1 variant strongly associated with rapid progression to AIDS in Cuba. EBioMedicine. 2015;2(3):244–54.

62. Thorne C, Ferencic N, Malyuta R, Mimica J, Niemiec T. Central Asia: hotspot in the worldwide {HIV} epidemic. Lancet Infect Dis. 2010;10(7):479–88.

63. Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci. 2007;104(47):18566–70.

64. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, et al. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000 Jun;288:1789–96.

65. Ajoge HO, Gordon ML, de Oliveira T, Green TN, Ibrahim S, Shittu OS, et al. Genetic Characteristics, Coreceptor Usage Potential and Evolution of Nigerian HIV-1 Subtype G and CRF02_AG Isolates. Carr J, editor. PLoS ONE. 2011 Mar 14;6(3):e17865.

66. Hawke KG, Waddell RG, Gordon DL, Ratcliff RM, Ward PR, Kaldor JM. HIV non-B subtype distribution: Emerging trends and risk factors for imported and local infections newly diagnosed in South Australia. AIDS Res Hum Retroviruses. 2013;29(2):311–7.

67. Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. Lancet Infect Dis. 2011;11(1):45–56.

68. Leoz M, Chaix M-L, Delaugerre C, Rivoisy C, Meyer L, Rouzioux C, et al. Circulation of multiple patterns of unique recombinant forms B/CRF02_AG in France: precursor signs of the emergence of an upcoming CRF B/02. AIDS. 2011 Jul;25(11):1371–7.

69. Chibo D, Birch C. Increasing Diversity of Human Immunodeficiency Virus Type 1 Subtypes Circulating in Australia. AIDS Res Hum Retroviruses. 2012 Jun;28(6):578–83.

70. Fox J, Castro H, Kaye S, McClure M, Weber JN, Fidler S, et al. Epidemiology of non-B clade forms of HIV-1 in men who have sex with men in the UK. Aids. 2010

Sep;24:2397–401.

71. Dwyer DE, Ge YC, Bolton WV, Wang B, Cunningham AL, Saksena NK. Subtype B isolates of human immunodeficiency virus type 1 detected in Australia. Ann Acad Med Singapore. 1996 Mar;25:188–91.

72. Descamps D, Chaix ML, Montes B, Pakianather S, Charpentier C, Storto A, et al. Increasing prevalence of transmitted drug resistance mutations and non-B subtype circulation in antiretroviral-naive chronically HIV-infected patients from 2001 to 2006/2007 in France. J Antimicrob Chemother. 2010 Dec;65:2620–7.

73. Carr JK, Osinusi A, Flynn CP, Gilliam BL, Maheshwari V, Zhao RY. Two Independent Epidemics of HIV in Maryland. Jaids-J Acquir Immune Defic Syndr. 2010 Jul;54:297–303.

74. Wheeler WH, Ziebell RA, Zabina H, Pieniazek D, Prejean J, Bodnar UR, et al. Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, US-2006. Aids. 2010 May;24:1203–12.

75. Thomson MM, Perez-Alvarez L, Najera R. Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect Dis. 2002 Aug;2:461–71.

76. Bhoopat L, Eiangleng L, Rugpao S, Frankel SS, Weissman D, Lekawanvijit S, et al. In vivo identification of Langerhans and related dendritic cells infected with HIV-1 subtype E in vaginal mucosa of asymptomatic patients. Mod Pathol. 2001 Dec;14:1263–9.

77. Baeten JM, Chohan B, Lavreys L, Chohan V, McClelland RS, Certain L, et al. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis. 2007 Apr;195:1177–80.

78. Kanki PJ, Hamel DJ, Sankale JL, Hsieh CC, Thior I, Barin F, et al. Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis. 1999 Jan;179:68–73.

79. Kaleebu P, Nankya IL, Yirrell DL, Shafer LA, Kyosiimire-Lugemwa J, Lule DB, et al. Relation between chemokine receptor use, disease stage, and HIV-1 subtypes A and D: results from a rural Ugandan cohort. JAIDS J Acquir Immune Defic Syndr. 2007;45(1):28–33.

80. Palm AA, Esbjornsson J, Mansson F, Kvist A, Isberg P-E, Biague A, et al. Faster Progression to AIDS and AIDS-Related Death Among Seroincident Individuals Infected With Recombinant HIV-1 A3/CRF02_AG Compared With Sub-subtype A3. J Infect Dis. 2014 Mar 1;209(5):721–8.

81. Taylor BS, Hammer SM. The challenge of HIV-1 subtype diversity (vol 358, pg 1590, 2008). N Engl J Med. 2008 Oct;359:1965–6.

82. Spira S, Wainberg MA, Loemba H, Turner D, Brenner BG. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J Antimicrob Chemother. 2003 Feb;51:229–40.

83. Quinn TC. Is an AIDS free generation feasible? Science vs Reality. In: Proceedings of the 2012 Australasian conference on HIV&AIDS. Melbourne, Australia; 2012.

84. Howat P, Jancey J. Policy-at-a-glance – HIV/AIDS Policy. Public Health Association Australia; 2011.

85. UNAIDS. How AIDS changed everything, MDG 6: 15 years, 15 lessons of hope from the AIDS response. UNAIDS; 2015.

86. The Kirby Institute. HIV, viral hepatitis and sexually transmissible infections in Australia Annual Surveillance Report 2015. The Kirby Institute, UNSW Australia, Sydney NSW 2052; 2015.

87. Joint United Nations Programme on HIV/AIDS. Report on the global HIV/AIDS epidemic. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS, UNAIDS; 2000.

88. AVERT. Avert - Worldwide HIV & AIDS Statistics [Internet]. [cited 2014 Apr 6]. Available from: http://www.avert.org/worldwide-hiv-aids-statistics.htm

89. Kilmarx PH. Global epidemiology of HIV: Curr Opin HIV AIDS. 2009 Jul;4(4):240–6.

90. McIntyre J. Antiretrovirals for reducing mother-to-child transmission of HIV infection. Reprod Contracept-Shanghai Inst Plan Parent Res. 2007;27(4):296.

91. Brody S, Gisselquist D, Potterat JJ, Drucker E. Evidence of iatrogenic HIV transmission in children in South Africa. BJOG Int J Obstet Gynaecol. 2003;110(5):450–2.

92. Ng KT, Ong LY, Lim SH, Takebe Y, Kamarulzaman A, Tee KK. Evolutionary History of HIV-1 Subtype B and CRF01_AE Transmission Clusters among Men Who Have Sex with Men (MSM) in Kuala Lumpur, Malaysia. PLoS ONE. 2013 Jun 20;8(6):e67286.

93. Phillips AF, Pirkle CM. Moving beyond behaviour: advancing HIV risk prevention epistemologies and interventions (A report on the state of the literature). Glob Public Health. 2011;6(6):577–92.

94. AIDS*gov. Pre-Exposure Prophylaxis (PrEP) [Internet]. AIDS.gov. 2015 [cited 2015 Sep 26]. Available from: https://www.aids.gov/hiv-aids-basics/prevention/reduce-your-risk/pre-exposure-prophylaxis/

95. Grant RM, Lama JR, Anderson PL, McMahan V, Liu AY, Vargas L, et al. Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men. N Engl J Med. 2010 Nov 23;363(27):2587–99.

96. Thigpen MC, Kebaabetswe PM, Paxton LA, Smith DK, Rose CE, Segolodi TM, et al. Antiretroviral Preexposure Prophylaxis for Heterosexual HIV Transmission in Botswana. N Engl J Med. 2012 Jul 11;367(5):423–34.

97. Choopanya K, Martin M, Suntharasamai P, Sangkum U, Mock PA, Leethochawalit M, et al. Antiretroviral prophylaxis for HIV infection in injecting drug users in

Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. The Lancet. 381(9883):2083–90.

98. Baeten JM, Donnell D, Ndase P, Mugo NR, Campbell JD, Wangisi J, et al. Antiretroviral Prophylaxis for HIV-1 Prevention among Heterosexual Men and Women. N Engl J Med. 2012 Aug 2;367(5):399–410.

99. Bourry O, Brochard P, Souquiere S, Makuwa M, Calvo J, Dereudre-Bosquet N, et al. Prevention of vaginal simian immunodeficiency virus transmission in macaques by postexposure prophylaxis with zidovudine, lamivudine and indinavir. Aids. 2009;23(4):447–54.

100. Otten RA, Smith DK, Adams DR, Pullium JK, Jackson E, Kim CN, et al. Efficacy of postexposure prophylaxis after intravaginal exposure of pig-tailed macaques to a human-derived retrovirus (human immunodeficiency virus type 2). J Virol. 2000;74(20):9771–5.

101. Cardo DM, Culver DH, Ciesielski CA, Srivastava PU, Marcus R, Abiteboul D, et al. A case–control study of HIV seroconversion in health care workers after percutaneous exposure. N Engl J Med. 1997;337(21):1485–90.

102. Australian Federation of AIDS organisations. HIV statistics in Australia [Internet]. Australian Federation of AIDS Organisations. 2015 [cited 2015 Sep 22]. Available from: https://www.afao.org.au/about-hiv/the-hiv-epidemic/hiv-statistics-australia#.VgZ0hVoVfzI

103. Guy RJ, McDonald AM, Bartlett MJ, Murray JC, Giele CM, Davey TM, et al. Characteristics of HIV diagnoses in Australia, 1993-2006. Sex Health. 2008;5:91–6.

104. Hocking J, Keenan C, Catton M, Breschkin A, Guy R, Hellard M. Rising HIV infection in Victoria: an analysis of surveillance data. Aust N Z J Public Health. 2004 Jun;28:217 – +.

105. Wilson D, Hamilton S. Reduced condom use and other STIs lead to more HIV in Australia's gay men [Internet]. UNSW Newsroom. 2008 [cited 2015 Sep 22]. Available from: https://newsroom.unsw.edu.au/news/health/reduced-condom-use-and-other-stis-lead-more-hiv-australias-gay-men

106. McDonald AM, Gertig DM, Crofts N, Kaldor JM. A national surveillance system for newly acquired HIV-infection in Australia. Am J Public Health. 1994 Dec;84:1923–8.

107. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance, Joint United Nations Programme on HIV/AIDS, World Health Organization. Guidelines for second generation HIV surveillance: an update : know your epidemic [Internet]. 2013 [cited 2015 Sep 27]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK158982/

108. Chen JHK, Wong KH, Chen ZW, Chan K, Lam HY, To SWC, et al. Increased Genetic Diversity of HIV-1 Circulating in Hong Kong. Plos One. 2010 Aug;5(8):e12198.

109. Sullivan PS, Hamouda O, Delpech V, Geduld JE, Prejean J, Semaille C, et al.

Reemergence of the HIV Epidemic Among Men Who Have Sex With Men in North America, Western Europe, and Australia, 1996-2005. Ann Epidemiol. 2009;19:423–31.

110. Foxman B, Riley L. Molecular epidemiology: focus on infection. Am J Epidemiol. 2001;153(12):1135–41.

111. Balode D, Westman M, Kolupajeva T, Rozentale B, Albert J. Low prevalence of transmitted drug resistance among newly diagnosed HIV-1 patients in Latvia. J Med Virol. 2010;82(12):2013–8.

112. Chow WZ, Takebe Y, Syafina NE, Prakasa MS, Chan KG, Al-Darraji HAA, et al. A Newly Emerging HIV-1 Recombinant Lineage (CRF58_01B) Disseminating among People Who Inject Drugs in Malaysia. PLoS ONE. 2014 Jan 22;9(1):e85250.

113. Bertagnolio S, World Health Organization, World Health Organization, Department of HIV/AIDS. WHO HIV drug resistance report, 2012 [Internet]. Geneva, Switzerland: World Health Organization; 2012 [cited 2015 Sep 17]. Available from: http://apps.who.int/iris/bitstream/10665/75183/1/9789241503938_eng.pdf

114. HIV/AIDS JUNP on, others. 2013 report on the global AIDS epidemic [Internet]. Unaids; 2013 [cited 2014 Sep 3] p. 198. Available from: http://www.unaids.org/en/media/unaids/contentassets/documents/epidemiology/2013/gr2013/unaids_global_report_2013_en.pdf

115. Bartmeyer B, Kuecherer C, Houareau C, Werning J, Keeren K, Somogyi S, et al. Prevalence of Transmitted Drug Resistance and Impact of Transmitted Resistance on Treatment Success in the German HIV-1 Seroconverter Cohort. Fouchier RAM, editor. PLoS ONE. 2010 Oct 7;5(10):e12718.

116. National Institutes of Health. Starting antiretroviral treatment early improves outcomes for HIV-infected individuals [Internet]. News & Events. 2015. Available from: http://www.nih.gov/news/health/may2015/niaid-27.htm

117. Ambrosioni J, Sued O, Nicolas D, Parera M, López-Diéguez M, Romero A, et al. Trends in Transmission of Drug Resistance and Prevalence of Non-B Subtypes in Patients with Acute or Recent HIV-1 Infection in Barcelona in the Last 16 Years (1997-2012). Paraskevis D, editor. PLOS ONE. 2015 Jun 3;10(6):e0125837.

118. Kuritzkes DR. HIV Drug Resistance: New Insight and Updated Practices. PRN Noteb. 2004 Sep;9(3):9–13.

119. Lazzarin A, Campbell T, Clotet B, Johnson M, Katlama C, Moll A, et al. Efficacy and safety of TMC125 (etravirine) in treatment-experienced HIV-1-infected patients in DUET-2: 24-week results from a randomised, double-blind, placebo-controlled trial. The Lancet. 2007;370(9581):39–48.

120. Johnson LB, Saravolatz LD. Etravirine, a Next-Generation Nonnucleoside Reverse-Transcriptase Inhbitor. Clin Infect Dis. 2009 Apr 15;48(8):1123–8.

121. Ward MJ, Lycett SJ, Kalish ML, Rambaut A, Brown AJL. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. J Virol.

2013;87(4):1967–73.

122. Boyd M, Pett S. HIV fusion inhibitors: a review. Aust Prescr. 2008 Jun;31(3):66–9.

123. Hicks C, Gulick RM. Raltegravir: the first HIV type 1 integrase inhibitor. Clin Infect Dis. 2009;48(7):931–9.

124. Temesgen Z, Siraj DS. Raltegravir: first in class HIV integrase inhibitor. Ther Clin Risk Manag. 2008;4(2):493.

125. Youle M, Wainberg MA. Pre-exposure chemoprophylaxis (PREP) as an HIV prevention strategy. J Int Assoc Physicians AIDS Care JIAPAC. 2003;2(3):102–5.

126. Penazzato M, Prendergast A, Muhe L, Tindyebwa D, Abrams E. Optimisation of antiretroviral therapy in HIV-infected children under 3 years of age (Review). Cochrane Libr. 2014;(5):1–69.

127. Myers JE, Taylor BS, Rojas Fermín RA, Reyes EV, Vaughan C, José L, et al. Transmitted Drug Resistance Among Antiretroviral-Naive Patients with Established HIV Type 1 Infection in Santo Domingo, Dominican Republic and Review of the Latin American and Caribbean Literature. AIDS Res Hum Retroviruses. 2012 Jul;28(7):667–74.

128. Bennett DE, Myatt M, Bertagnolio S, Sutherland D, Gilks CF. Recommendations for surveillance of transmitted HIV drug resistance in countries scaling up antiretroviral treatment. Antivir Ther. 2008;13:25.

129. Nii-Trebi NI, Ibe S, Barnor JS, Ishikawa K, Brandful JAM, Ofori SB, et al. HIV-1 Drug-Resistance Surveillance among Treatment-Experienced and -Naïve Patients after the Implementation of Antiretroviral Therapy in Ghana. Barbour JD, editor. PLoS ONE. 2013 Aug 19;8(8):e71972.

130. Frentz D, Boucher C, Van De Vijver D. Temporal changes in the epidemiology of transmission of drug-resistant HIV-1 across the world. AIDs Rev. 2012;14(1):17–27.

131. Stanford University. HIV-1 Drug Resistance in ARV-naive Populations [Internet]. Stanford University HIV Drug Resistance Database. 2014 [cited 2015 Oct 9]. Available from: http://hivdb.stanford.edu/surveillance/map/

132. Shafer RW, Kantor R, Gonzales MJ. The genetic basis of HIV-1 resistance to reverse transcriptase and protease inhibitors. AIDS Rev. 2000;2(4):211.

133. Kantor R, Katzenstein DA, Efron B, Carvalho AP, Wynhoven B, Cane P, et al. Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: Results of a global collaboration. Plos Med. 2005 Apr;2:325–37.

134. Langs-Barlow A, Paintsil E. Impact of Human Immunodeficiency Virus Type-1 Sequence Diversity on Antiretroviral Therapy Outcomes. Viruses. 2014 Oct 20;6(10):3855–72.

135. Sendagire H, Easterbrook PJ, Nankya I, Arts E, Thomas D, Reynolds SJ. The challenge of HIV-1 antiretroviral resistance in Africa in the era of HAART. AIDS Rev. 2009;11(2):59–70.

136. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, et al. Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. Nixon DF, editor. PLoS ONE. 2009 Mar 6;4(3):e4724.

137. Stanford University. NRTI Resistance Notes [Internet]. Stanford University HIV Drug Resistance Database. 2014. Available from: http://hivdb.stanford.edu/DR/NRTIResiNote.html

138. Stanford University. Surveillance Drug Resistance Mutation (SDRM) Worksheet: NNRTIs [Internet]. Stanford University HIV Drug Resistance Database. 2014. Available from: http://hivdb.stanford.edu/pages/SDRM.worksheet.NNRTI.html

139. Kantor R, Katzenstein D. Polymorphism in HIV-1 non-subtype B protease and reverse transcriptase and its potential impact on drug susceptibility and drug resistance evolution. AIDS Rev. 2003;5(1):25–35.

140. Smith D, Moini N, Pesano R, Cachay E, Aiem H, Lie Y, et al. Clinical utility of HIV standard genotyping among antiretroviral-naive individuals with unknown duration of infection. Clin Infect Dis. 2007;44(3):456–8.

141. Shet A, Berry L, Mohri H, Mehandru S, Chung C, Kim A, et al. Tracking the prevalence of transmitted antiretroviral drug-resistant HIV-1: a decade of experience. JAIDS J Acquir Immune Defic Syndr. 2006;41(4):439–46.

142. Novitsky V, Wang R, Bussmann H, Lockman S, Baum M, Shapiro R, et al. HIV-1 Subtype C-Infected Individuals Maintaining High Viral Load as Potential Targets for the "Test-and-Treat'' Approach to Reduce HIV Transmission. Plos One. 2010;5(4):e10148.

143. Marconi VC, Sunpath H, Lu Z, Gordon M, Koranteng-Apeagyei K, Hampton J, et al. Prevalence of HIV-1 drug resistance after failure of a first highly active antiretroviral therapy regimen in KwaZulu Natal, South Africa. Clin Infect Dis. 2008;46(10):1589–97.

144. Turner D, Shahar E, Katchman E, Kedem E, Matus N, Katzir M, et al. Prevalence of the K65R resistance reverse transcriptase mutation in different HIV-1 subtypes in Israel. J Med Virol. 2009;81(9):1509–12.

145. Brenner BG, Coutsinos D. The K65R mutation in HIV-1 reverse transcriptase: genetic barriers, resistance profile and clinical implications. HIV Ther. 2009;3(6):583–94.

146. Orrell C, Walensky RP, Losina E, Pitt J, Freedberg KA, Wood R. HIV-1 clade C resistance genotypes in naïve patients and after first virological failure in a large community ART programme. Antivir Ther. 2009;14(4):523.

147. Doualla-Bell F, Avalos A, Brenner B, Gaolathe T, Mine M, Gaseitsiwe S, et al. High prevalence of the K65R mutation in human immunodeficiency virus type 1 subtype C isolates from infected patients in Botswana treated with didanosine-based regimens. Antimicrob Agents Chemother. 2006;50(12):4182–5.

148. Gupta R, Chrystie I, O'Shea S, Mullen J, Kulasegaram R, Tong C. K65R and Y181C are less prevalent in HAART-experienced HIV-1 661 subtype A patients. Aids.

2005;19(16):1916–9.

149. Brenner BG, Oliveira M, Doualla-Bell F, Moisi DD, Ntemgwa M, Frankel F, et al. HIV-1 subtype C viruses rapidly develop K65R resistance to tenofovir in cell culture. Aids. 2006;20(9):F9–13.

150. Eshleman SH, Guay LA, Wang J, Mwatha A, Brown ER, Musoke P, et al. Distinct patterns of emergence and fading of K103N and Y181C in women with subtype A vs. D after single-dose nevirapine: HIVNET 012. JAIDS J Acquir Immune Defic Syndr. 2005;40(1):24–9.

151. Eshleman SH, Mracna M, Guay LA, Deseyve M, Cunningham S, Mirochnick M, et al. Selection and fading of resistance mutations in women and infants receiving nevirapine to prevent HIV-1 vertical transmission (HIVNET 012). Aids. 2001;15(15):1951–7.

152. Stanford University. Surveillance Drug Resistance Mutation (SDRM) Worksheet: PIs [Internet]. Stanford University HIV Drug Resistance Database. 2014. Available from: http://hivdb.stanford.edu/pages/SDRM.worksheet.PI.html

153. Ariyoshi K, Matsuda M, Miura H, Tateishi S, Yamada K, Sugiura W. Patterns of point mutations associated with antiretroviral drug treatment failure in CRF01_AE (subtype E) infection differ from subtype B infection. J Acquir Immune Defic Syndr 1999. 2003;33(3):336–42.

154. Blower SM, Aschenbach AN, Gershengorn HB, Kahn JO. Predicting the unpredictable: transmission of drug-resistant HIV. Nat Med. 2001;7(9):1016–20.

155. Cambiano V, Bertagnolio S, Jordan MR, Lundgren JD, Phillips A. Transmission of drug resistant HIV and its potential impact on mortality and treatment outcomes in resource-limited settings. J Infect Dis. 2013;207(suppl 2):S57–62.

156. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73.

157. Wilkinson E. Molecular characterization of non-subtype C and recombinant HIV-1 viruses from Cape Town, South Africa [Masters]. [South Africa]: Stellenbosch University; 2008.

158. Hartfield M, Murall CL, Alizon S. Clinical applications of pathogen phylogenies. Trends Mol Med. 2014 Jul;20(7):394–404.

159. Ratcliff RM. Genetic and Functional Studies of the Mip Protein of Legionella [PhD]. [Adelaide, Australia]: University of Adelaide; 2000.

160. Hall BG. Building Phylogenetic Trees from Molecular Data with MEGA. Mol Biol Evol. 2013 Mar 13;30(5):1229–35.

161. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. JAIDS J Acquir Immune Defic Syndr. 2008;49(1):9–16.

162. Hué S, Clewley JP, Cane PA, Pillay D. HIV-1 pol gene variation is sufficient for

reconstruction of transmissions in the era of antiretroviral therapy. Aids. 2004;18(5):719–28.

163. Lubelchek RJ, Hoehnen SC, Hotton AL, Kincaid SL, Barker DE, French AL. Transmission Clustering Among Newly Diagnosed HIV Patients in Chicago, 2008 to 2011: Using Phylogenetics to Expand Knowledge of Regional HIV Transmission Patterns. JAIDS J Acquir Immune Defic Syndr. 2015;68(1):46–54.

164. Poon AFY, Joy JB, Woods CK, Shurgold S, Colley G, Brumme CJ, et al. The Impact of Clinical, Demographic and Risk Factors on Rates of HIV Transmission: A Population-based Phylogenetic Analysis in British Columbia, Canada. J Infect Dis. 2015 Mar 15;211(6):926–35.

165. Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. Bmc Infect Dis. 2010;10:262.

166. Babiker A, Darby S, De Angelis D, Kwart D, Porter K, Beral V, et al. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. Lancet. 2000;355:1131–7.

167. Buonaguro L, Tornesello ML, Buonaguro FM. Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications. J Virol. 2007 Oct 1;81(19):10209–19.

168. Neogi U, Bontell I, Shet A, De Costa A, Gupta S, Diwan V, et al. Molecular Epidemiology of HIV-1 Subtypes in India: Origin and Evolutionary History of the Predominant Subtype C. Salemi M, editor. PLoS ONE. 2012 Jun 29;7(6):e39819.

169. Salemi M, Vandamme A-M. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press; 2003.

170. Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach. Nachdr. Malden, Mass.: Blackwell Publ; 2009. 346 p.

171. Li W-H, Graur D. Fundamentals of Molecular Evolution. Sinauer Associates; 1991.

172. Jukes TH, Cantor CR. Evolution of protein molecules. Mamm Protein Metab. 1969;3:21–132.

173. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16(2):111–20.

174. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160–74.

175. Tavaré S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci. 1986;17:57–86.

176. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.

177. Baldauf SL. Phylogeny for the faint of heart: a tutorial. Trends Genet. 2003 Jun;19(6):345–51.

178. Singh M. Phylogeny. Lecture presented at: Topics in Computational Molecular Biology; 1999 Oct; Princeton University, Princeton, New Jersey, USA.

179. Holmes S. Bootstrapping phylogenetic trees: theory and methods. Stat Sci. 2003;241–55.

180. Berry V, Gascuel O. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol Biol Evol. 1996;13(7):999–1011.

181. Campbell N, Reece J, Urry L, Cain M, Wasserman S, Minorsky P, et al. Biology, 8th Edition. 8th ed. Benjamin Cummings; 2007.

182. Yebra G, de Mulder M, Martín L, Pérez-Cachafeiro S, Rodríguez C, Labarga P, et al. Sensitivity of seven HIV subtyping tools differs among subtypes/recombinants in the Spanish cohort of naïve HIV-infected patients (CoRIS). Antiviral Res. 2011 Jan;89(1):19–25.

183. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AFY, Hughes G, et al. An Evolutionary Model-Based Algorithm for Accurate Phylogenetic Breakpoint Mapping and Subtype Prediction in HIV-1. Fraser C, editor. PLoS Comput Biol. 2009 Nov 26;5(11):e1000581.

184. Alcantara LCJ, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, et al. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res. 2009 Jul 1;37(Web Server):W634–42.

185. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, et al. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics. 2005 Oct 1;21(19):3797–800.

186. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

187. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. Clin Infect Dis. 2006;42(11):1608–18.

188. Struck D, Lawyer G, Ternes A-M, Schmit J-C, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 2014 Oct 13;42(18):e144–e144.

189. Zhang M, Schultz A-K, Calef C, Kuiken C, Leitner T, Korber B, et al. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. Nucleic Acids Res. 2006 Jul 1;34(Web Server):W463–5.

190. Lynch RM, Shen T, Gnanakaran S, Derdeyn CA. Appreciating HIV type 1 diversity: subtype differences in Env. AIDS Res Hum Retroviruses. 2009;25(3):237–48.

191. Cleary JG, Witten IH. Data compression using adaptive coding and partial string

matching. Commun IEEE Trans On. 1984;32(4):396–402.

192. Holguin A, Lopez M, Soriano V. Reliability of Rapid Subtyping Tools Compared to That of Phylogenetic Analysis for Characterization of Human Immunodeficiency Virus Type 1 Non-B Subtypes and Recombinant Forms. J Clin Microbiol. 2008 Oct 8;46(12):3896–9.

193. Gifford R, de Oliveira T, Rambaut A, Myers RE, Gale CV, Dunn D, et al. Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. Aids. 2006;20(11):1521–9.

194. Gifford RJ, Oliveira T, Rambaut A, Pybus OG, Dunn D, Vandamme AM, et al. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. J Virol. 2007 Dec;81:13050–6.

195. Wilkinson E, Holzmayer V, Jacobs GB, de Oliveira T, Brennan CA, Hackett J, et al. Sequencing and Phylogenetic Analysis of Near Full-Length HIV-1 Subtypes A, B, G and Unique Recombinant AC and AD Viral Strains Identified in South Africa. AIDS Res Hum Retroviruses. 2015;31(4):412–20.

196. Garcia F, Perez-Cachafeiro S, Guillot V, Alvarez M, Perez-Romero P, Perez-Elias MJ, et al. Transmission of HIV drug resistance and non-B subtype distribution in the Spanish cohort of antiretroviral treatment naive HIV-infected individuals (CoRIS). Antiviral Res. 2011 Aug;91:150–3.

197. Herring BL, Ge YC, Wang B, Ratnamohan M, Zheng F, Cunningham AL, et al. Segregation of human immunodeficiency virus type 1 subtypes by risk factor in Australia. J Clin Microbiol. 2003 Oct;41:4600–4.

198. Rhee S-Y. Human immunodeficiency virus reverse transcriptase and protease sequence database. Nucleic Acids Res. 2003 Jan 1;31(1):298–303.

199. Gifford RJ, Liu TF, Rhee S-Y, Kiuchi M, Hue S, Pillay D, et al. The calibrated population resistance tool: standardized genotypic estimation of transmitted HIV-1 drug resistance. Bioinformatics. 2009;25(9):1197–8.

200. Nisar L, Qadir MI, Malik SA, Tabassum N. Characterization of the immunodominant regions within gp41 of env gene of HIV in Pakistan. J Chem Soc Pak. 2011;33(4):545–8.

201. Poveda E, Rodés B, Toro C, Martín-Carbonero L, Gonzalez-Lahoz J, Soriano V. Evolution of the gp41 env region in HIV-infected patients receiving T-20, a fusion inhibitor. Aids. 2002;16(14):1959–61.

202. Spang R, Rehmsmeier M, Stoye J. A novel approach to remote homology detection: jumping alignments. J Comput Biol. 2002;9(5):747–60.

203. Pineda-Peña A-C, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol. 2013 Oct;19:337–48.

204. Adungo FO, Gicheru MM, Adungo NI, Matilu MM, Lihana RW, Khamadi SA. Diversity of Human Immunodeficiency Virus Type-1 Subtypes in Western Kenya. World J AIDS. 2014;04(04):365–72.

205. Russell KL, Carcamo C, Watts DM, Sanchez J, Gotuzzo E, Euler A, et al. Emerging genetic diversity of HIV-1 in South America. Aids. 2000 Aug;14:1785–91.

206. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AMJ, de Vijver DAV, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology. 2009;6(1):49.

207. Mallitt K-A, Wilson DP, McDonald A, Wand H. HIV incidence trends vary between jurisdictions in Australia: an extended back-projection analysis of men who have sex with men. Sex Health. 2012;9:138–43.

208. Del Amo J, Likatavicius G, Perez-Cachafeiro S, Hernando V, Gonzalez C, Jarrin I, et al. The epidemiology of HIV and AIDS reports in migrants in the 27 European Union countries, Norway and Iceland: 1999-2006. Eur J Public Health. 2010 Nov 4;21(5):620–6.

209. Asher AK, Hahn JA, Couture M-C, Maher K, Page K. People Who Inject Drugs, {HIV} Risk, and {HIV} Testing Uptake in Sub-Saharan Africa. J Assoc Nurses AIDS Care. 2013;24(6):e35–44.

210. Zhou Y-H, Yao Z-H, Liu F-L, Li H, Jiang L, Zhu J-W, et al. High Prevalence of HIV, HCV, HBV and Co-Infection and Associated Risk Factors among Injecting Drug Users in Yunnan Province, China. Tillmann H, editor. PLoS ONE. 2012 Aug 16;7(8):e42937.

211. Blumental S, Ferster A, Van den Wijngaert S, Lepage P. HIV Transmission Through Breastfeeding: Still Possible in Developed Countries. Pediatrics. 2014;134(3):e875–9.

212. Noori T, Pharris A. Migrant health: Sexual transmission of HIV within migrant groups in the EU/EEA and implications for effective interventions. Stockholm, Sweden: European Centre for Disease Prevention and Control; 2013 p. 1–45.

213. Jewkes RK, Dunkle K, Nduna M, Shai N. Intimate partner violence, relationship power inequity, and incidence of HIV infection in young women in South Africa: a cohort study. The Lancet. 2010;376(9734):41–8.

214. Iordanskiy S, Waltke M, Feng YJ, Wood C. Subtype-associated differences in HIV-1 reverse transcription affect the viral replication. Retrovirology. 2010;7(1):85.

215. Salemi M. Toward a robust monitoring of HIV subtypes distribution worldwide: AIDS. 2011 Mar;25(5):713–4.

216. Zhang J, Guo Z, Yang J, Pan X, Jiang J, Ding X, et al. Genetic diversity of HIV-1 and transmitted drug resistance among newly diagnosed individuals with HIV infection in Hangzhou, China. J Med Virol. 2015;87:1668–76.

217. Kiwelu IE, Novitsky V, Kituma E, Margolin L, Baca J, Manongi R, et al. HIV-1 pol

Diversity among Female Bar and Hotel Workers in Northern Tanzania. Liang C, editor. PLoS ONE. 2014 Jul 8;9(7):e102258.

218. Zhou Y-H, Liang Y-B, Pang W, Qin W-H, Yao Z-H, Chen X, et al. Diverse forms of HIV-1 among Burmese long-distance truck drivers imply their contribution to HIV-1 cross-border transmission. BMC Infect Dis. 2014;14(1):463.

219. Dalai SC, de Oliveira T, Harkins GW, Kassaye SG, Lint J, Manasa J, et al. Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. AIDS Lond Engl. 2009;23(18):2523.

220. Condon J. How to access HIV care and treatment in Australia [Internet]. National Association of People with HIV Australia (NAPWHA). 2014 [cited 2015 Oct 7]. Available from: http://napwha.org.au/health-treatment/hiv-treatment/how-access-hiv-care-and-treatment-australia

221. ASHM. Drug-Resistance Testing [Internet]. Antiretroviral guidelines. 2014 [cited 2015 Oct 7]. Available from: http://arv.ashm.org.au/arv-guidelines/laboratory-testing/drug-resistance-testing

222. Stringer EM, Chi BH, Chintu N, Creek TL, Ekouevi DK, Coetzee D, et al. Monitoring effectiveness of programmes to prevent mother-to-child HIV transmission in lower-income countries. Bull World Health Organ. 2008;86(1):57–62.

223. Krakower DS, Jain S, Mayer KH. Antiretrovirals for Primary HIV Prevention: the Current Status of Pre-and Post-exposure Prophylaxis. Curr HIV/AIDS Rep. 2015;12(1):127–38.

224. Kiwanuka N, Laeyendecker O, Robb M, Kigozi G, Arroyo M, McCutchan F, et al. Effect of human immunodeficiency virus type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. J Infect Dis. 2008 Mar;197:707–13.

225. Horyniak D, Stoove M, Yohannes K, Breschkin A, Carter T, Hatch B, et al. The impact of immigration on the burden of HIV infection in Victoria, Australia. Sex Health. 2009;6:123–8.

226. McPherson ME, McMahon T, Moreton RJ, Ward KA. Using HIV notification data to identify priority migrant groups for HIV prevention, New South Wales, 2000-2008. Commun Dis Intell Q Rep. 2011;35(2):185.

227. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The Challenge of HIV-1 Subtype Diversity. N Engl J Med. 2008 Apr 10;358(15):1590–602.

228. Ong LY, Razak SNH, Lee YM, Sri La Sri Ponnampalavanar S, Syed Omar SF, Azwa RI, et al. Molecular diversity of HIV-1 and surveillance of transmitted drug resistance variants among treatment Naïve patients, 5 years after active introduction of HAART in Kuala Lumpur, Malaysia. J Med Virol. 2014 Jan 1;86(1):38–44.

229. Jain V, Liegler T, Vittinghoff E, Hartogensis W, Bacchetti P, Poole L, et al. Transmitted Drug Resistance in Persons with Acute/Early HIV-1 in San Francisco, 2002-2009. Chen Z, editor. PLoS ONE. 2010 Dec 10;5(12):e15510.

230. Hamers RL, Oyomopito R, Kityo C, Phanuphak P, Siwale M, Sungkanuparph S, et al. Cohort Profile: The PharmAccess African (PASER-M) and the TREAT Asia (TASER-M) Monitoring Studies to Evaluate Resistance--HIV drug resistance in sub-Saharan Africa and the Asia-Pacific. Int J Epidemiol. 2010 Nov 10;41(1):43–54.

231. Poz Health, Life and HIV. HIV Drug Chart [Internet]. Avita Pharmacy; 2015 [cited 2015 Sep 22]. Available from: http://www.aidsmeds.com/articles/DrugChart_10632.shtml

232. Akinsete O, Hirigoyen D, Cartwright C, Schut R, Kantor R, Henry K. K103N Mutation in Antiretroviral Therapy—Naive African Patients Infected with HIV Type 1. Clin Infect Dis. 2004;39(4):575–8.

233. Guo W, Li H, Zhuang D, Jiao L, Liu S, Li L, et al. Impact of Y181C and/or H221Y mutation patterns of HIV-1 reverse transcriptase on phenotypic resistance to available non-nucleoside and nucleoside inhibitors in China. BMC Infect Dis. 2014;14(1):237.

234. Nachega JB, Hislop M, Dowdy DW, Gallant JE, Chaisson RE, Regensberg L, et al. Efavirenz versus nevirapine-based initial treatment of HIV infection: clinical and virological outcomes in Southern African adults: AIDS. 2008 Oct;22(16):2117–25.

235. Lefebvre E, Schiffer CA. Resilience to Resistance of HIV-1 Protease Inhibitors: Profile of Darunavir. AIDS Rev. 2008;10(3):131–42.

236. Chan PA, Huang A, Kantor R. Low prevalence of transmitted K65R and other tenofovir resistance mutations across different HIV-1 subtypes: implications for pre-exposure prophylaxis. J Int AIDS Soc. 2012 Oct 15;15(2):17701.

237. Young TP, Parkin NT, Stawiski E, Pilot-Matias T, Trinh R, Kempf DJ, et al. Prevalence, mutation patterns, and effects on protease inhibitor susceptibility of the L76V mutation in HIV-1 protease. Antimicrob Agents Chemother. 2010;54(11):4903–6.

238. Magiorkinis G, Paraskevis D, Schmidt HA, Hatzakis A. The phylogenetic information profile of HIV-1 and the degradation effect of recombination. Infect Genet Evol. 2008 Mar;8(2):139–45.

239. Bollerup AR, Donoghoe MC, Lazarus JV, Nielsen S, Matic S. Access to highly active antiretroviral therapy (HAART) in the WHO European Region 2003—2005. Scand J Public Health. 2008;36(2):183–9.

240. Joint United Nations Programme on HIV/AIDS (UNAIDS). UNAIDS report on the global AIDS epidemic 2013, 2013: Geneva [Internet]. Switzerland: UNAIDS; 2013 [cited 2015 Aug 21] p. 1–106. Available from: http://www.unaids.org/sites/default/files/media_asset/UNAIDS_Global_Report_2013_en_1.pdf

241. Brenner BG, Roger M, Moisi DD, Oliveira M, Hardy I, Turgel R, et al. Transmission networks of drug resistance acquired in primary/early stage HIV infection. AIDS Lond Engl. 2008;22(18):2509.

242. Bezemer D, van Sighem A, Lukashov VV, van der Hoek L, Back N, Schuurman R,

et al. Transmission networks of HIV-1 among men having sex with men in the Netherlands. Aids. 2010;24(2):271–82.

243. Chmiel JS, Detels R, Kaslow RA, Van Raden M, Kingsley LA, Brookmeyer R, et al. Factors associated with prevalent human immunodeficiency virus (HIV) infection in the Multicenter AIDS Cohort Study. Am J Epidemiol. 1987;126(4):568–75.

244. Di Stefano M, Favia A, Monno L, Lopalco P, Caputi O, Scardigno AC, et al. Intracellular and cell-free (infectious) HIV-1 in rectal mucosa. J Med Virol. 2001;65(4):637–43.

245. Harding AK, Gray LA, Neal M. Confidentiality Limits With Clients Who Have HIV: A Review of Ethical and Legal Guidelines and Professional Policies. J Couns Dev. 1993;71(3):297–305.

246. Golden MR, Stekler J, Kent JB, Hughes JP, Wood RW. An Evaluation of HIV Partner Counseling and Referral Services Using New Disposition Codes: Sex Transm Dis. 2009 Feb;36(2):95–101.

247. Resik S, Lemey P, Ping L-H, Kouri V, Joanes J, Perez J, et al. Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba. Aids Res Hum Retroviruses. 2007 Mar;23:347–56.

248. Aldous JL, Pond SK, Poon A, Jain S, Qin H, Kahn JS, et al. Characterizing HIV Transmission Networks Across the United States. Clin Infect Dis. 2012 Oct 15;55(8):1135–43.

249. Brenner BG, Wainberg MA. Future of Phylogeny in HIV Prevention: JAIDS J Acquir Immune Defic Syndr. 2013 Jul;63:S248–54.

250. Castley A, Gaudieri S, James I, Gizzarelli L, Guelfi G, John M, et al. Longitudinal trends in Western Australian HIV-1 sequence diversity and viral transmission networks and their influence on clinical parameters: 2000–2014. AIDS Res Hum Retroviruses. 2015;(ja).

251. Dennis AM, Hué S, Hurt CB, Napravnik S, Sebastian J, Pillay D, et al. Phylogenetic insights into regional HIV transmission: AIDS. 2012 Sep;26(14):1813–22.

252. Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc Natl Acad Sci U S A. 2005;102(12):4425–9.

253. Mehta SR, Wertheim JO, Brouwer KC, Wagner KD, Chaillon A, Strathdee S, et al. HIV Transmission Networks in the San Diego–Tijuana Border Region. EBioMedicine. 2015 Oct;2(10):1456–63.

254. Ragonnet-Cronin M, Ofner-Agostini M, Merks H, Pilon R, Rekart M, Archibald CP, et al. Longitudinal Phylogenetic Surveillance Identifies Distinct Patterns of Cluster Dynamics: JAIDS J Acquir Immune Defic Syndr. 2010 Sep;55(1):102–8.

255. Birch CJ, McCaw RF, Bulach DM, Revill PA, Carter JT, Tomnay J, et al. Molecular analysis of human immunodeficiency virus strains associated with a case of criminal

transmission of the virus. J Infect Dis. 2000;182(3):941–4.

256. Cameron S. HIV, Crime and the Law in Australia: Options for Policy Reform – a law reform advocacy kit. Australian Federation of AIDS Organisations; 2011.

257. Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proc Natl Acad Sci. 2010 Dec 14;107(50):21242–7.

258. Bernard EJ, Azad Y, Vandamme A-M, Weait M, Geretti AM. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission*. HIV Med. 2007;8(6):382–7.

259. O'Byrne P, Willmore J, Bryan A, Friedman DS, Hendriks A, Horvath C, et al. Nondisclosure prosecutions and population health outcomes: examining HIV testing, HIV diagnoses, and the attitudes of men who have sex with men following nondisclosure prosecution media releases in Ottawa, Canada. BMC Public Health. 2013;13(1):94.

260. Millett GA, Peterson JL, Flores SA, Hart TA, Jeffries WL, Wilson PA, et al. Comparisons of disparities and risks of HIV infection in black and other men who have sex with men in Canada, UK, and USA: a meta-analysis. The Lancet. 2012 Jul;380(9839):341–8.

261. Piyasirisilp S, McCutchan FE, Carr JK, Sanders-Buell E, Liu W, Chen J, et al. A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. J Virol. 2000;74(23):11286–95.

262. Ng KT, Ng KY, Khong WX, Chew KK, Singh PK, Yap JK, et al. Phylodynamic Profile of HIV-1 Subtype B, CRF01_AE and the Recently Emerging CRF51_01B among Men Who Have Sex with Men (MSM) in Singapore. PLoS ONE. 2013 Dec 2;8(12):e80884.

263. Meloni ST, Kim B, Sankalé J-L, Hamel DJ, Tovanabutra S, Mboup S, et al. Distinct human immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. J Virol. 2004;78(22):12438–45.

264. Palm AA, Esbjörnsson J, Månsson F, Biague A, da Silva ZJ, Norrgren H, et al. Cocirculation of Several Similar But Unique HIV-1 Recombinant Forms in Guinea-Bissau Revealed by Near Full-Length Genomic Sequencing. AIDS Res Hum Retroviruses. 2015;31(9):938–45.

265. Sagoe KW, Dwidar M, Adiku TK, Arens MQ. HIV-1 CRF 02 AG polymerase genes in Southern Ghana are mosaics of different 02 AG strains and the protease gene cannot infer subtypes. Virol J. 2009;6(1):27.

266. Li Y, Tee KK, Liao H, Hase S, Uenishi R, Li X-J, et al. Identification of a novel second-generation circulating recombinant form (CRF48_01B) in Malaysia: a descendant of the previously identified CRF33_01B. JAIDS J Acquir Immune Defic Syndr. 2010;54(2):129–36.

267. Sanabani S, Neto WK, Kalmar EM, Diaz RS, Janini LM, Sabino EC. Analysis of the near full length genomes of HIV-1 subtypes B, F and BF recombinant from a cohort of 14 patients in Sao Paulo, Brazil. Infect Genet Evol. 2006;6(5):368–77.

268. Taylor S, Cane P, Hué S, Xu L, Wrin T, Lie Y, et al. Identification of a transmission chain of HIV type 1 containing drug resistance-associated mutations. AIDS Res Hum Retroviruses. 2003;19(5):353–61.

269. Goldwater PN. Iatrogenic Blood-borne Viral Infections in Refugee Children from War and Transition Zones. Emerg Infect Dis [Internet]. 2013 Jun [cited 2015 Oct 15];19(6). Available from: http://wwwnc.cdc.gov/eid/article/19/6/12-0806_article.htm

270. Leitner T, Kumar S, Albert J. Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J Virol. 1997;71(6):4761–70.

271. Gao F, Morrison SG, Robertson DL, Thornton CL, Craig S, Karlsson G, et al. Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. The WHO and NIAID Networks for HIV Isolation and Characterization. J Virol. 1996;70(3):1651–67.

272. Pieniazek D, Janini LM, Ramos A, Tanuri A, Schechter M, Peralta JM, et al. HIV-1 patients may harbor viruses of different phylogenetic subtypes: implications for the evolution of the HIV/AIDS pandemic. Emerg Infect Dis. 1995;1(3):86.

273. Patel MB, Hoffman NG, Swanstrom R. Subtype-specific conformational differences within the V3 region of subtype B and subtype C human immunodeficiency virus type 1 Env proteins. J Virol. 2008;82(2):903–16.

274. Xu L, Hué S, Taylor S, Ratcliffe D, Workman JA, Jackson S, et al. Minimal variation in T-20 binding domain of different HIV-1 subtypes from antiretroviral-naive and-experienced patients. Aids. 2002;16(12):1684–6.

275. Carmona R, Perez-Alvarez L, Munoz M, Casado G, Delgado E, Sierra M, et al. Natural resistance-associated mutations to Enfuvirtide (T20) and polymorphisms in the gp41 region of different HIV-1 genetic forms from T20 naive patients. J Clin Virol. 2005;32(3):248–53.

276. Bunnik EM, Pisas L, van Nuenen AC, Schuitemaker H. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. J Virol. 2008;82(16):7932–41.

277. Choisy M, Woelk CH, Guégan J-F, Robertson DL. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol. 2004;78(4):1962–70.

278. Koonin EV. The logic of chance: the nature and origin of biological evolution. FT press; 2011.

279. Hall HI, Frazier EL, Rhodes P, Holtgrave DR, Furlow-Parmley C, Tang T, et al.

Differences in Human Immunodeficiency Virus Care and Treatment Among Subpopulations in the United States. JAMA Intern Med. 2013 Jul 22;173(14):1337.

280. Hurt CB, Beagle S, Leone PA, Sugarbaker A, Pike E, Kuruc J, et al. Investigating a Sexual Network of Black Men Who Have Sex With Men: Implications for Transmission and Prevention of HIV Infection in the United States. JAIDS J Acquir Immune Defic Syndr. 2012 Dec;61(4):515–21.

281. Van de Peer Y. Phylogenetic inference based on distance methods. Phylogenetic Handb. 2009;142–60.

282. Pasquier C, Millot N, Njouom R, Sandres K, Cazabat M, Puel J, et al. HIV-1 subtyping using phylogenetic analysis of pol gene sequences. J Virol Methods. 2001;94(1):45–54.

283. Chow WZ, Ong LY, Razak SH, Lee YM, Ng KT, Yong YK, et al. Molecular Diversity of HIV-1 among People Who Inject Drugs in Kuala Lumpur, Malaysia: Massive Expansion of Circulating Recombinant Form (CRF) 33_01B and Emergence of Multiple Unique Recombinant Clusters. PLoS ONE. 2013 May 7;8(5):e62560.

284. Njouom R, Pasquier C, Sandres-Sauné K, Harter A, Souyris C, Izopet J. Assessment of HIV-1 subtyping for Cameroon strains using phylogenetic analysis of pol gene sequences. J Virol Methods. 2003 Jun;110(1):1–8.

285. Starcich BR, Hahn BH, Shaw GM, McNeely PD, Modrow S, Wolf H, et al. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. Cell. 1986;45(5):637–48.

286. Yahi N, Fantini J, Tourres C, Tivoli N, Koch N, Tamalet C. Use of drug resistance sequence data for the systematic detection of non-B human immunodeficiency virus type 1 (HIV-1) subtypes: how to create a sentinel site for monitoring the genetic diversity of HIV-1 at a country scale. J Infect Dis. 2001;183(9):1311–7.

287. Brown AE, Gifford RJ, Clewley JP, Kucherer C, Masquelier B, Porter K, et al. Phylogenetic Reconstruction of Transmission Events from Individuals with Acute HIV Infection: Toward More-Rigorous Epidemiological Definitions. J Infect Dis. 2009 Feb;199(3):427–31.

288. Oster AM, Pieniazek D, Zhang X, Switzer WM, Ziebell RA, Mena LA, et al. Demographic but not geographic insularity in HIV transmission among young black MSM: AIDS. 2011 Nov;25(17):2157–65.

289. Smith DM, May SJ, Tweeten S, Drumright L, Pacold ME, Pond SLK, et al. A public health model for the molecular surveillance of HIV transmission in San Diego, California: AIDS. 2009 Jan;23(2):225–32.

290. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al. The Global Transmission Network of HIV-1. J Infect Dis. 2014 Jan 15;209(2):304–13.

291. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, et al. Global epidemiology of HIV infection in men who have sex with men. The Lancet.

2012 Jul;380(9839):367–77.

292. Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, et al. Transmission Clustering Drives the Onward Spread of the HIV Epidemic Among Men Who Have Sex With Men in Quebec. J Infect Dis. 2011 Oct 1;204(7):1115–9.

293. Sullivan EA, Koro S, Tabrizi S, Kaldor J, Poumerol G, Chen S, et al. Prevalence of sexually transmitted diseases and human immunodeficiency virus among women attending prenatal services in Apia, Samoa. Int J Std Aids. 2004;15:116–9.

294. Sullivan EA, Abel M, Tabrizi S, Garland SM, Grice A, Poumerol G, et al. Prevalence of sexually transmitted infections among antenatal women in Vanuatu, 1999-2000. Sex Transm Dis. 2003;30:362–6.

295. Holguín A, Lospitao E, López M, de Arellano ER, Pena MJ, del Romero J, et al. Genetic characterization of complex inter-recombinant HIV-1 strains circulating in Spain and reliability of distinct rapid subtyping tools. J Med Virol. 2008 Mar;80(3):383–91.

296. Sacktor N, Nakasujja N, Skolasky RL, Rezapour M, Robertson K, Musisi S, et al. HIV Subtype D Is Associated with Dementia, Compared with Subtype A, in Immunosuppressed Individuals at Risk of Cognitive Impairment in Kampala, Uganda. Clin Infect Dis. 2009 Sep;49(5):780–6.

297. Ntemgwa M, Gill MJ, Brenner BG, Moisi D, Wainberg MA. Discrepancies in Assignment of Subtype/Recombinant Forms by Genotyping Programs for HIV Type 1 Drug Resistance Testing May Falsely Predict Superinfection. AIDS Res Hum Retroviruses. 2008 Jul;24(7):995–1002.

298. Konings FA, Haman GR, Xue Y, Urbanski MM, Hertzmark K, Nanfack A, et al. Genetic analysis of HIV-1 strains in rural eastern Cameroon indicates the evolution of second-generation recombinants to circulating recombinant forms. JAIDS J Acquir Immune Defic Syndr. 2006;42(3):331–41.

299. Takebe Y, Motomura K, Tatsumi M, Lwin HH, Zaw M, Kusagawa S. High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: geographical hot spot of extensive recombination. Aids. 2003;17(14):2077–87.

300. Yang R, Kusagawa S, Zhang C, Xia X, Ben K, Takebe Y. Identification and characterization of a new class of human immunodeficiency virus type 1 recombinants comprised of two circulating recombinant forms, CRF07_BC and CRF08_BC, in China. J Virol. 2003;77(1):685–95.

301. Tovanabutra S, Sanders EJ, Graham SM, Mwangome M, Peshu N, McClelland RS, et al. Evaluation of HIV Type 1 Strains in Men Having Sex with Men and in Female Sex Workers in Mombasa, Kenya. Aids Res Hum Retroviruses. 2010 Feb;26:123–31.

302. Riva C, Romano L, Saladini F, Lai A, Carr JK, Francisci D, et al. Identification of a Possible Ancestor of the Subtype A1 HIV Type 1 Variant Circulating in the Former Soviet Union. AIDS Res Hum Retroviruses. 2008 Oct;24(10):1319–25.

303. Su L, Graf M, Zhang Y, von Briesen H, Xing H, Köstler J, et al. Characterization of

a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B′) recombinant strain in China. J Virol. 2000;74(23):11367–76.

304. Tee KK, Pybus OG, Li X-J, Han X, Shang H, Kamarulzaman A, et al. Temporal and Spatial Dynamics of Human Immunodeficiency Virus Type 1 Circulating Recombinant Forms 08_BC and 07_BC in Asia. J Virol. 2008 Sep 15;82(18):9206–15.

305. Tovanabutra S, Watanaveeradej V, Viputtikul K, De Souza M, Razak MH, Suriyanon V, et al. A new circulating recombinant form, CRF15_01B, reinforces the linkage between IDU and heterosexual epidemics in Thailand. AIDS Res Hum Retroviruses. 2003;19(7):561–7.

306. Tee KK, Li X-J, Nohtomi K, Ng KP, Kamarulzaman A, Takebe Y. Identification of a novel circulating recombinant form (CRF33_01B) disseminating widely among various risk populations in Kuala Lumpur, Malaysia. JAIDS J Acquir Immune Defic Syndr. 2006;43(5):523–9.

307. Tee KK, Pon CK, Kamarulzaman A, Ng KP. Emergence of HIV-1 CRF01_AE/B unique recombinant forms in Kuala Lumpur, Malaysia. Aids. 2005;19(2):119–26.

308. Li Z, Li J, Feng Y, Kalish ML, Lu H, Yin L, et al. Genomic characterization of two novel HIV-1 unique (CRF01_AE/B) recombinant forms among men who have sex with men in Beijing, China. AIDS Res Hum Retroviruses. 2015;(ja).

309. Peeters M, Toure-Kane C, Nkengasong JN. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. Aids. 2003;17(18):2547–60.

310. Osmanov S, Pattou C, Walker N, Schwardländer B, Esparza J. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr 1999. 2002;29(2):184–90.

311. Songok EM, Lwembe RM, Kibaya R, Kobayashi K, Ndembi N, Kita K, et al. Active Generation and Selection for HIV Intersubtype A/D Recombinant Forms in a Coinfected Patient in Kenya. AIDS Res Hum Retroviruses. 2004 Feb;20(2):255–8.

312. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, et al. US Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. J Virol. 2003;77(11):6359–66.

313. Van Der Auwera G, Janssens W, Heyndrickx L, Van Der Groen G. Reanalysis of full-length HIV type 1 group M subtype K and sub-subtype F2 with an MS-DOS bootscanning program. AIDS Res Hum Retroviruses. 2001;17(2):185–9.

314. Shcherbakova NS, Shalamova LA, Delgado E, Fernández-García A, Vega Y, Karpenko LI, et al. Short Communication: Molecular Epidemiology, Phylogeny, and Phylodynamics of CRF63_02A1, a Recently Originated HIV-1 Circulating Recombinant Form Spreading in Siberia. AIDS Res Hum Retroviruses. 2014 Sep;30(9):912–9.

315. Aulicino PC, Holmes EC, Rocco C, Mangano A, Sen L. Extremely rapid spread of human immunodeficiency virus type 1 BF recombinants in Argentina. J Virol.

2007;81(1):427–9.

316. Tee KK, Pybus OG, Parker J, Ng KP, Kamarulzaman A, Takebe Y. Estimating the date of origin of an HIV-1 circulating recombinant form. Virology. 2009;387(1):229–34.

317. Liao H, Tee KK, Hase S, Uenishi R, Li X-J, Kusagawa S, et al. Phylodynamic analysis of the dissemination of HIV-1 CRF01_AE in Vietnam. Virology. 2009;391(1):51–6.

318. Brown RJP, Peters PJ, Caron C, Gonzalez-Perez MP, Stones L, Ankghuambom C, et al. Intercompartmental Recombination of HIV-1 Contributes to env Intrahost Diversity and Modulates Viral Tropism and Sensitivity to Entry Inhibitors. J Virol. 2011 Jun 15;85(12):6024–37.

319. Wagner GA, Pacold ME, Pond SLK, Caballero G, Chaillon A, Rudolph AE, et al. Incidence and prevalence of intrasubtype HIV-1 dual infection in at-risk men in the United States. J Infect Dis. 2013;jit633.

320. Schultz AK, Stanke M. A Jumping Profile HMM for Remote Protein Homology Detection [Internet]. N/A; n.d; N/A. Available from: http://gobics.de/anne/poster.pdf

321. Myers RE, Gale CV, Harrison A, Takeuchi Y, Kellam P. A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). Bioinformatics. 2005 Jul 26;21(17):3535–40.

322. Leitner T, Escanilla D, Marquina S, Wahlberg J, Broström C, Hansson HB, et al. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. Virology. 1995;209(1):136–46.

323. Westesson O, Holmes I. Accurate Detection of Recombinant Breakpoints in Whole-Genome Alignments. Regev A, editor. PLoS Comput Biol. 2009 Mar 20;5(3):e1000318.

APPENDIX 1 – SUBTYPE ASSIGNMENT

		***pol*-PR/RT**							***env*-gp41**					***pol/env***
Case	**Sample**	**jpHMM**	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**	**MEGA phy**	**jpHMM**	**REGA**	**COMET**	**LANL BLAST**	**MEGA phy**	**Inferred subtype**
1	1928302	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
10	2192987	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
126	12162512	AE	AE^	AE/AE	AE	AE	15_01B^	AE	AE	C	AE	AE^	AE	AE/AE
127	14528451	AE	AE^	AE/AE	15_01B	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
129	99986042	A1	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
130	14220742	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
131	21830432	AE	AE^	AE/AE	AE	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
132	19640250	AE	AE^	AE/AE	AE	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
143	18468564	A1	AE^	AE/AE	AE^	AE	AE^	AE	-	C	AE	AE^	AE	AE/AE
149	16056024	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
150	16056823	AE/A1	AE^	AE/AE	AE^	15_01B	AE^	AE	AE	C	AE	AE^	AE	AE/AE
151	16256286	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
152	2168896	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
154	23141944	AE	AE^	AE/AE	15_01B	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
155	2505123	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
156	30417794	AE	AE^	AE/AE	15_01B^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
158	6603507	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
159	21014077	AE	AE^	AE/AE	15_01B	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
160	32092120	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
163	34032805	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
164	11274164	AE	AE^	AE/AE	AE	AE	AE^	AE	AE	C	AE	AE^	AE	AE/AE
165	31038285	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
166	31718054	AE	AE^	AE/AE	AE	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
167	19640140	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
168	20158678	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
169	6706716	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	G	AE	AE^	AE	AE/AE
170	9028891	AE	AE^	AE/AE	AE^	AE	15_01B^	AE	AE	C	AE	AE^	AE	AE/AE
172	99139547	A1	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
161	16001679	AE	AE^	AE/AE	CISR	AE	AE^	15_01B	AE	J	AE	AE^	AE	AE/AE
171	99090558	AE	AE^	AE/AE	15_01B^	AE	AE^	15_01B	AE	A1	AE	AE^	AE	AE/AE
120	2991453	AE	AE^	AE/AE	AE^	AE	AE^	A1	AE	G	AE	AE^	AE	AE/AE
121	19008295	AE	AE^	AE/AE	15_01B	AE	AE^	A1	AE	G	AE	AE^	AE	AE/AE
128	31725169	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE	A1	AE	AE^	AE	AE/AE
157	6092993	AE	AE^	AE/AE	AE^	15_01B	AE^	AE	AE/B	A1	AE	AE^	AE	AE/U
162	30024539	AE	AE^	AE/AE	AE^	AE	AE^	AE	AE/B	C	AE	AE^	AE	AE/U
122	20008873	A1	A1^	A/AE	A1^	A1	AE^	AE	A1	A1^	A1	A3^	A1	A1/A1
123	20008852	A1	A1^	A/AE	A1^	A1	AE^	AE	A1	A1^	A1	A2^	A1	A1/A1
141	19701541	A1	A1^	A/AE	A1^	A1	AE^	AE	A1	A1^	A1	A1^	A1	A1/A1
118	9075229	A1	A1^	X/AE	A1^	A1	A1^	A1	A1	A1^	A1	A1^	A1	A1/A1
119	99237176	A1	A1^	A/AE	A1^	A1	A1^	A1	A1	A1^	A1	A1^	A1	A1/A1
138	14676267	A1	A1^	A/A	A1^	A1	A1^	A1	A1	A1^	A1	A2^	A1	A1/A1
142	20823470	A1	A1^	A/A	A1^	A1	A1^	A1	A1	A1^	A1	A1^	A1	A1/A1
136	2618286	A1	A1^	A/AE	A1^	A1	A1^	A1	B	B^	B	U	B	A1/B
135	5545224	A1	A1^	A/A	CISR	A1	A1^	A1	D	D^	D	D^	D	A1/D
63	32048145	B	B^	B/B	CISR	B	B^	47_BF	B	B^	B	B^	B	B/B
57	2029353	B	B^	B/B	B^	B	B^	48_01B	B	B^^	B	B^	B	B/B
58	20258479	B	B-like^	B/B	B^	B	B^	48_01B	B	B^^	B	B^	B	B/B
60	15894934	B	B^	B/B	B^	B	B^	48_01B	B	B^	B	B^	29_BF	B/B
61	20055854	B	B/F1	B/B	B^	B	B^	48_01B	B	B^	15_01B	B^	29_BF	B/B
193	1451936	B	B^	B/B	B^	B	B^	48_01B	B	B^	B	B^	03_AB	B/B

		pol-PR/RT							*env*-gp41					*pol/env*
Case	**Sample**	**jpHMM**	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**	**MEGA phy**	**jpHMM**	**REGA**	**COMET**	**LANL BLAST**	**MEGA phy**	**Inferred subtype**
195	2996268	B	B^	B/B	B^	B	B^	48_01B	B	B^	B	SHIV	B	B/B
197	5970878	B	B^	B/B	B^	B	B^	48_01B	B	B^^	B	B^	B	B/B
146	20055850	B	B^	B/B	B^	B	B^	51_01B	B	B^	B	B^	B	B/B
147	20786291	B	B^	B/B	B^	B	B^	51_01B	B	B^	B	B^	B	B/B
196	5513449	B	B^	X/B	B^	B	B^	51_01B	B	B^	B	B^	B	B/B
205	12953353	B	B^	B/B	B^	B	B^	51_01B	B	B^	B	B^	B	B/B
4	3863952	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
5	5094482	B	B^	B/X	B^	B	B^	B	B	B^	B	B^	B	B/B
7	9140126	B	B^	B/B	B	CISR	B^	B	B	B^^	B	B^	B	B/B
9	6707171	B	B^	B/B	B	B	B^	B	-	B^	B	B^	B	B/B
11	6096140	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
12	20008784	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
13	16001672	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
14	92457483	B	B^	B/B	B	B	B^	B	B	B^^	B	B^	B	B/B
15	19640144	B	B^	B/B	B	CISR	B^	B	B	B^^	B	B^	B	B/B
16	15837375	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
17	99141685	B	B^	B/B	B	B	B^	B	B	B^^	B	B^	B	B/B
18	99172140	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
19	16001681	B	B-like^	B/B	B	B	B^	B	B	B^^	B	B^	B	B/B
20	19640123	B	B^	B/B	B^	B	B^	B	B	B^^	B	SHIV	B	B/B
21	5973887	B	B^	B/B	CISR	B	B^	B	B	B^	B	B^	B	B/B
22	6941144	B	B^	B/B	B	B	B^	B	-	B^^	15_01B	B^	B	B/B
23	3569864	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
24	3842619	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
25	99500052	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
26	91453154	B	B^	B/B	B^	CISR	B^	B	B	B^^	B	SHIV	B	B/B
27	4146887	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
28	23476192	B	B^	B/B	B^	CISR	B^	B	B	B^^	B	B^	B	B/B
29	20250438	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
30	30040121	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
31	2028523	B	B^	B/B	B^	B	B^	B	B	B^^	B	SHIV	B	B/B
32	5799925	B	B^	B/B	B^	B	B^	B	B	B^^	15_01B	B^	B	B/B
33	2771676	B	B^	B/B	B^	B	B^	B	B	B^	B	SHIV	B	B/B
34	2817947	B	B^	B/B	B^	B	B^	B	B	B^	B	SHIV	B	B/B
35	2508761	B	B^	B/B	B	B	B^	B	B	B^^	B	B^	B	B/B
36	30591960	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
37	31768998	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
39	20786165	B	B^	B/B	B^	B	B^	B	-	B^	B	B^	B	B/B
40	2204141	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
41	13530637	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
42	20169865	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
43	20006543	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
44	19640143	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
45	22990320	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
46	21986259	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
47	2203990	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
48	90017957	B	B^	B/B	B^	B	B^	B	-	B^	B	B^	B	B/B
49	12545668	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
50	2932525	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
51	6707173	B	B^	B/B	B^	B	B^	B	-	B^	B	U	B	B/B
52	13009412	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
53	20055849	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
54	33694639	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
55	1922272	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
56	13044300	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
59	20851044	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
62	17142811	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
134	9232177	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
137	6027259	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
140	5256171	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	15_01B	B/B
144	20823325	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
148	6276448	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
173	14720791	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
174	11468126	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
175	20651815	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
176	2168597	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
177	22103235	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
178	22333559	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
179	90130737	B	B^	B/B	B^	B	B^	B	-	B^	B	B^	B	B/B
180	90144520	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
181	11274406	B	B^	B/B	B	B	B^	B	B	B^^	B	B^	B	B/B
182	11972330	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
183	1533024	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
184	15894939	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
185	19640209	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
186	2713816	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
187	3087881	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
188	5973961	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
189	9051502	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
190	18668983	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
191	21830650	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
192	22026640	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	47_BF	B/B
194	14665046	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
198	20798036	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	28_BF	B/B
199	12545640	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
201	12760485	B	B^	B/B	B	B	B^	B	B	B	B	B^	47_BF	B/B
202	20111954	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	47_BF	B/B
203	11846811	B	B^	B/B	B	B	B^	B	B	B^	B	B^	B	B/B
204	12760740	B	B^	B/B	B^	B	B^	B	B	B^	B	U	14_BG	B/B
206	19640194	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
207	19640342	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
208	20959768	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
209	2183643	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
210	23997325	B	B^	B/B	B^	B	B^	B	B	B^^	B	B^	B	B/B
211	2476955	B	B^	B/B	B^	B	B^	B	B	B^^	B	SHIV	B	B/B
212	2477024	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	15_01B	B/B
213	2816307	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
214	2991479	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
215	30630697	B	B^	B/B	CISR	B	B^	B	B	B^^	B	AE/B^	B	B/B
216	30673601	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
217	30812542	B	B^	D/B	B^	B	B^	B	B	B^	B	B^	B	B/B
218	3896556	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
219	4513800	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
220	5410195	B	B^	B/B	B^	B	B^	B	B	B^	B	B^	B	B/B
221	90131987	B	B^	B/B	B^	B	B^	B	-	B^	B	B^	B	B/B
91	18028751	C	C^	C/C	C^	C	C^	C	A1	A1^	A1	A1/D^	A1	C/A1
3	2476903	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
8	20843896	C	C^	C/C	C^	C	A1/C^	C	C	C^	C	C^	C	C/C

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
69	2618143	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
70	2991463	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
71	2991464	B	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
72	14791830	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
73	13932499	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
74	14815292	C	C^	C/C	C^	C	C^	C	C	C^^	C	C^	C	C/C
75	30408245	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
76	30408244	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
77	20785294	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
78	4364138	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
79	16256696	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
80	4294160	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
81	9023702	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
82	2007570	C	C^	C/C	C^	C	C^	C	-	C^	C	C^	C	C/C
83	5496316	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
84	9077520	B	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
85	99172132	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
86	6707162	C	C^	C/C	C^	C	C^	C	-	C^^	C	C^	C	C/C
87	2618226	C	C^	C/C	C	C	C^	C	C	C^	C	C^	C	C/C
88	21376678	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
89	12020855	C/B	C^	C/C	C	C	C^	C	C	C^	C	C^	C	C/C
90	31187450	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
92	16001620	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
93	18317169	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
94	19640326	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
95	20650791	C	C^	C/C	C^	C	C^	C	-	C^	C	C^	C	C/C
96	90060379	C	C^	C/C	C^	C	C^	C	-	C^	C	C^	C	C/C

		pol-PR/RT							*env*-gp41					*pol/env*
Case	Sample	jpHMM	REGA	Stanford CPR	SCUEAL	COMET	LANL BLAST	MEGA phy	jpHMM	REGA	COMET	LANL BLAST	MEGA phy	Inferred subtype
97	20857905	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
98	20106524	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
100	19080945	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
101	19640284	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
102	14648212	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
103	19640289	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
104	23141923	C	C^	C/C	C	C	C^	C	C	C^	C	C^	C	C/C
105	90415995	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
106	32092009	C	C^	C/C	C^	C	C^	C	C	C^	C	B/C^	C	C/C
107	23953761	C	C^	C/C	C^	C	C^	C	C	C^	C	C^	C	C/C
99	21237697	C	C^	C/C	C^	C	C^	C	C	C^	C/check for 07	B/C^	C	C/U
64	22990325	D	D^	D/D	19_cpx^	D	D^	D	D	D^	D	D^	D	D/D
2	16921534	B	D^	D/D	D^	D	D^	D	-	D^	D	A/D^	D	D/U
153	22976612	A1/B	A1/B	AE/B	CISR	CISR	AE/B^	52_01B	AE	A1	AE	AE^	AE	U/AE
125	21032866	A1	A1^	AE/AE	CISR	A1	A1^	A1	A1	A1^	A1	A1^	A1	U/A1
200	12242272	A1/B	A1-like^	AE/AE	CISR	CISR	AE^	AE	B	B^	B	B^	B	U/B
38	2204038	B	B-like^	B/B	CISR	B	B^	B	B	B^	B	B^	B	U/B
6	90203162	A1/B/A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1^	A1/check for AG	AG^	AG	U/U
65	15989247	G/B	B/G	AG/B	CISR	CISR	B^	23_BG	A1	A1^	A1/check for AG	AG^	AG	U/U
66	20835856	G/B	B/G	AG/B	CISR	CISR	B^	23_BG	A1	A1^	A1/check for AG	AG^	AG	U/U
67	30417851	G/B	B/G	AG/B	CISR	CISR	B^	23_BG	A1	A1^	A1/check for AG	AG^	AG	U/U
68	2818947	C/B	08_BC^	C/C	CISR	08_BC	08_BC^	08_BC	C	C^	08_BC	08_BC^	08_BC	U/U
108	22111189	G/A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1^	A1/check for AG	AG^	AG	U/U

		pol-PR/RT							*env*-gp41					*pol/env*
Case	**Sample**	**jpHMM**	**REGA**	**Stanford CPR**	**SCUEAL**	**COMET**	**LANL BLAST**	**MEGA phy**	**jpHMM**	**REGA**	**COMET**	**LANL BLAST**	**MEGA phy**	**Inferred subtype**
109	22108496	A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1	A1/check for AG	AG^	AG	U/U
110	92006411	G/A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1^	A1/check for AG	63_02A1^	A1	U/U
111	90119902	G/B/G	AG^	AG/AG	CISR	AG	AG/A1^	AG	-	A1^	A1/check for AG	63_02A1^	A1	U/U
112	99874530	B	G^	AG/AG	CISR	AG	AG^	AG	A1/B	A1	AG	AG^	AG	U/U
113	9213856	G/A1	AG^	AG/AG	G^	AG	AE^	AG	A1	A1^	AG	AG^	AG	U/U
114	13876993	G	AG^	AG/AG	CISR	AG	AG/U^	AG	A1	A1^	A1/check for AG	AG^	AG	U/U
115	23624532	G/A1	AG^	AG/AG	G^	AG	AG^	AG	A1	A1^^	A1/check for AG	A1^	AG	U/U
116	90285489	G/A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1^	A1/check for AG	AG^	36_cpx	U/U
117	31130046	G/A1	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1	A1/check for AG	AG^	AG	U/U
124	31130019	A1	A1/K	K/AE	A1^	A1	A1^	AE	A1	A1^	A1/check for AG	AG/A1^	45_cpx	U/U
133	6603741	B/C/B	07_BC^	B/C	CISR	07_BC	07_BC^	07_BC	C	C^	07_BC	07_BC^	07_BC	U/U
139	16010109	A1	A1-like^	AG/AE	A2	A1	AG/22_01A1^	A1	A1	A1^	A1/check for AG	A3^	A1	U/U
145	22391692	B	AG^	AG/AG	CISR	AG	AG^	AG	A1	A1	A1/check for AG	A1^	AG	U/U

Key: ^ (under REGA, SCUEAL and LANL BLAST = bootstrap value ≥70%), ^^ (under REGA = bootstrap value ≥70% but REGA didn't assign pure subtype because it did not sub-cluster with a pure subtype, ** (under MEGA phy = part of transmission cluster with bootstrap value ≥98% and genetic diversity ≤1.5%), N (under REGA = no subtype assigned with a bootstrap ≥70%), U (Unassigned), SHIV (under LANL heading = Unassigned, Simian-Human Immunodeficiency Virus) CISR (complex intersubtype recombinant form), and – (Not available)

APPENDIX 2 – CONTRIBUTIONS

Name	Author/Speaker	Year	Location	Type
Time trends and risk factors for non-B clade infections in South Australia.	Hawke	2011	Australasian HIV/AIDS conference **Canberra**	Oral presentation
HIV subtype and TDRM trends in South Australia.	Hawke	2012	International Union for Sexually Transmitted Infections **Melbourne**	Oral presentation
HIV subtype and TDRM trends in South Australia.	Hawke	2012	Australasian Virology conference **Adelaide**	Poster
HIV non-B subtype distribution: Emerging trends and risk factors for imported and local infections newly diagnosed in South Australia.	Hawke et al	2012	AIDS Research and Human Retroviruses **Journal**	Journal article
A decade of HIV surveillance.	Hawke	2012	Gilead Sciences **Adelaide**	Sponsored oral presentation

Name	Author/Speaker	Year	Location	Type
Social determinants of Health – HIV.	Hawke	2013-2014	Flinders University **Adelaide**	Lecture
An epidemic in transition: impacts of migration and local networks on HIV sequence diversity and infection transmission in Australia 2005-2012.	Castley et al	2014	International AIDS conference **Melbourne**	Oral presentation
Molecular characterization of HIV-1 in South Australia; 14 years of subtype and drug resistance surveillance.	Hawke	2014	International AIDS conference **Melbourne**	Poster presentation
The Challenges of Diversity: HIV-1 Subtype Distribution and Transmission Networks within the Australian Molecular Epidemiology Network-HIV 2005-2012	Castley et al	2015	World STI and HI Congress **Brisbane**	Oral presentation

APPENDIX 3 – PUBLICATIONS

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 28, Number 11, 2012

DOI: 10.1089/aid.2012.0082

HIV Non-B Subtype Distribution: Emerging Trends and Risk Factors for Imported and Local Infections Newly Diagnosed in South Australia

Karen G. Hawke,[1] Russell G. Waddell,[2] David L. Gordon,[3] Rodney M. Ratcliff,[4] Paul R. Ward,[1] and John M. Kaldor[5]

Abstract

Monitoring HIV subtype distribution is important for understanding transmission dynamics. Subtype B has historically been dominant in Australia, but in recent years new clades have appeared. Since 2000, clade data have been collected as part of HIV surveillance in South Australia. The aim of this study was to evaluate the prevalence of and risk factors for HIV-1 non-B subtypes. The study population was composed of newly diagnosed, genotyped HIV subjects in South Australia between 2000 and 2010. We analyzed time trends and subtype patterns in this cohort; notification data were aggregated into three time periods (2000–2003, 2004–2006, and 2007–2010). Main outcome measures were number of new non-B infections by year, exposure route, and other demographic characteristics. There were 513 new HIV diagnoses; 425 had information on subtype. The majority (262/425) were in men who have sex with men (MSM), predominantly subtype B and acquired in Australia. Infections acquired in Australia decreased from 77% (2000–2003) to 64% (2007–2010) ($p=0.007$) and correspondingly the proportion of subtype B declined from 85% to 68% ($p=0.002$). Non-B infections were predominantly (83%) heterosexual contacts, mostly acquired overseas (74%). The majority (68%) of non-B patients were born outside of Australia. There was a nonsignificant increase from 1.6% to 4.2% in the proportion of locally transmitted non-B cases ($p=0.3$). Three non-B subtypes and two circulating recombinant forms (CRFs) were identified: CRF_AE ($n=41$), C ($n=36$), CRF_AG ($n=13$), A ($n=9$), and D ($n=2$). There has been a substantial increase over the past decade in diagnosed non-B infections, primarily through cases acquired overseas.

Introduction

HIV-1 is divided into distinct lineages, Major (M), Outlier (O), New (N), and P, most likely reflecting four separate introductions of simian immunodeficiency viruses into humans.[1–3] Analysis of nucleotide sequence variations of the reverse transcriptase and protease regions of the HIV *pol* gene is routinely conducted to determine resistance to antiretroviral drugs. Thus this region can be also exploited to define HIV into subtypes, and track HIV evolution and diversity.[4] Over 99% of the HIV pandemic is attributable to the M lineage[4] and there are at least nine phylogenetically distinct subtypes: A–D, F–H, J, and K.[5,6] Viruses of different subtypes can also recombine and create hybrid or circulating recombinant forms (CRFs) of which there are currently 52.[4,7]

Historically, HIV subtypes and CRFs have been broadly linked with geographic location and risk group.[8] However, subtype distribution of the global HIV pandemic has diversified extensively through mutation and recombination, partly driven by a combination of population mobility, sexual mixing, and the impact of antiretroviral therapies.[3,9] Subtype and CRF differences have been linked to disease progression, transmission route, pathogenicity, transmissibility, accuracy of current diagnostic assays, response to therapy, and development of drug resistance mutations.[3,5,10,11] Other factors can also influence these characteristics, which makes it difficult to

[1]Discipline of Public Health, Flinders University, Adelaide, Australia.
[2]Clinic 275, Royal Adelaide Hospital, Adelaide, Australia.
[3]Department of Microbiology and Infectious Diseases, SA Pathology at Flinders Medical Centre, and Flinders University, Adelaide, Australia.
[4]Department of Microbiology and Infectious Diseases, SA Pathology at Institute of Medical and Veterinary Science, and School of Biomedical Science, University of Adelaide, Adelaide, Australia.
[5]Kirby Institute, University of NSW, Sydney, Australia.

establish the independent effects of subtype, but this variability potentially has major clinical and epidemiological significance.[3,12]

Surveillance systems have been in place since the beginning of the global pandemic, and in recent years have incorporated molecular epidemiology as a tool both for surveillance of HIV-1 genetic diversity and to monitor transmission and geographic pathways of genetic variants.[3,13–15] For example, recent mapping of HIV strains in Asia revealed a large genetic diversity, including two new CRFs and transmission of new recombinant HIV-1 subtypes.[14,16]

Historically, subtype B has predominated in Western countries[5,6] where transmission is primarily through male-to-male sex.[6,8,17] However, subtype B accounts for only 11% of global HIV infections,[9] and the prevalence of non-B infections in Western countries is increasing. Recent studies in France have found non-B prevalence rates of 42–48% in newly diagnosed HIV infections.[18] In Italy, non-B prevalence rates rose from 25% in 2000 to over 60% in 2008, with African ethnicity and heterosexual acquisition as independent predictors.[6] In a Washington cohort the non-B prevalence rate was 13%,[13] and in one broad population-based study in the United States a national non-B prevalence rate of 5.1% was reported in newly diagnosed infections.[19] In a very recently published Australian study, Chibo and Birch (2012) found a non-B prevalence rate of 22% in a Victorian cohort.[20]

Parallel to this genetic diversity, there are increasing data on subtype-specific differences, related to genotyping, transmission efficiency, disease progression, vaccine development, and drug therapy.[12,21] Though subtype B has always accounted for a relatively small percentage of the total pandemic, it has historically been the predominant global reference clade for assay development, drug resistance testing, and antiretroviral susceptibility.[1,5,13]

In 2000, South Australia became the first state to integrate drug resistance testing as part of the routine HIV reporting and surveillance system. The resulting data provide the first analysis of Australian trends and molecular epidemiology of HIV subtype distribution over the past decade. We report a pattern of increasing non-B subtypes in South Australia, though subtype B still characterizes the predominant HIV infection in this cohort.

Materials and Methods

Surveillance system for HIV in South Australia

AIDS and HIV notification was commenced in South Australia in 1985 and 1991, respectively. For each new person diagnosed, a standardized form is completed, which includes demographic, epidemiological, and clinical information. Where possible, an in-depth interview is also conducted.

In 2000, South Australia became the first jurisdiction to conduct routine genotypic and drug resistance testing as part of an enhanced surveillance system. This genotype information is housed on a separate database, but is linked to the notification system via patient number. The South Australian Health Department is the custodian of both databases. For the current study, the two databases were merged to create a combined dataset. Patient identifiers such as name and address were removed, and limited demographic, epidemiological, and clinical data were retained, including gender, age, reported continent of birth, reported exposure route, and reported location of infection acquired (overseas/Australia). This study was approved by both the South Australian Health and Flinders University Research Ethics Committees.

Study population and design

Five hundred and thirteen people newly diagnosed with HIV between 2000 and 2010 were identified from the South Australian HIV notification database, and 425 were retrospectively selected from this dataset according to the following inclusion criteria: no previously documented positive diagnosis and a plasma-derived RNA *pol* sequence available for genotyping, taken within 12 months of diagnosis. For each patient, notification data, including location where the infection was acquired, were collected through a standardized form and interview at the time of diagnosis, as part of the routine notification protocol.

HIV-1 genotyping

Blood was collected for routine drug resistance testing, viral load, and $CD4^+$ cell count. Due to the number of routine samples, patient plasma was stored at $-20°C$ until genotyping. Past experience has demonstrated negligible degradation of virus nucleic acid. Viral RNA was extracted and a 1098 nucleotide fragment of the *pol* gene that encompasses the protease and reverse transcriptase genes was sequenced in both directions, using RT-PCR and dye terminator sequencing with standard commercial reagents. Sequences were assembled and proofread to obtain a contiguous sequence using Kodon 2.4 (Applied Maths, Sint-Martens-Latem, Belgium). The entire sequence was submitted to the Stanford HIV Drug Resistance Database for the determination of virus subtype and for drug resistance interpretation. Although phylogenetic and subtype analysis was carried out on the entire 1098 nucleotide fragment of the *pol* gene, only the protease region was used in subsequent epidemiological analyses for simplicity.

Statistical methods

HIV genetic and notification data were linked and analyzed using subtype as the dependent variable, and year, country of origin, where infection was acquired, reported risk exposure, and age as explanatory variables. Notification data were aggregated into three time periods (2000–2003, 2004–2006, and 2007–2010) of relatively equal numbers and with significant power to conduct statistical tests. Categorical variables were analyzed using chi square tests-for-trend and Fishers exact test to identify subtype-specific characteristics. Multivariate analysis was performed using logistic regression. Variables used are described in Table 1. Significance levels were set at $p \leq 0.05$. All data were analyzed using the software package Stata 10.1 (StataCorp LP, College Station, TX).

Results

South Australian HIV population; genotype

There were 513 reported diagnoses between 2000 and 2010, and 425 (83%) had genotypes determined. The annual number of diagnoses has remained relatively stable over the past 11 years (mean = 47/year). Demographic and other characteristics of those for whom genotypes were obtained ($n = 425$) were very similar to those for the total diagnosed population

TABLE 1. CHARACTERISTICS OF NEWLY DIAGNOSED HIV-INFECTED PATIENTS IN SOUTH AUSTRALIA 2000–2010

Characteristics	*All reported diagnoses (n=513)*	*Genotype obtained (n=425)*	*Number (%) B (n=324)*
Gender			
Male	437 (85)	371 (87)	306 (83)
Female	76 (15)	54 (13)	18 (33)
Age at HIV diagnosis (years)			
<25	54 (11)	45 (11)	27 (60)
25+	458 (89)	380 (89)	297 (78)
Unknown[a]	1 (0.1)		
Region of birth			
Australia/Oceania	277 (54)	240 (57)	211(88)
Africa	53 (10)	43 (10)	3 (7)
Asia	38 (7)	24 (6)	8 (33)
America	8 (2)	7 (2)	6 (86)
Europe	39 (8)	26 (6)	21 (81)
Not reported[a]	98 (19)	85 (20)	75 (88)
Risk exposure			
Heterosexual	180 (35)	146 (34)	62 (43)
MSM	314 (61)	262 (62)	255 (97)
Other	16 (3)	3 (1)	4 (100)
Unknown[a]	3 (1)	14 (3)	3 (23)
Location HIV acquired			
Australia	344 (67)	305 (72)	290 (95)
Overseas	163 (32)	118 (28)	32 (27)
Not stated[a]	6 (1)	2 (.2)	2 (100)

[a]Unknown: insufficient information in the database.

Data represents number (%) of subjects. MSM, men who have sex with men; Percentages in last column are proportion B compared to non-B. Genotyped sample was representative of the total population diagnosed.

($n=513$, Table 1). There were 101 cases (24%) of HIV non-B subtypes among these genotyped specimens (Fig. 1). The most common non-B infection was CRF_AE (41) followed by subtype C (36), CRF_AG (13), subtype A (9), and subtype D (2).

Though new diagnoses in South Australia have remained stable over time, there has been a significant change in subtype distribution within newly diagnosed individuals. The proportion of non-B infections increased from 15% (2000–2003) to 21% (2004–2006) to 32% (2007–2010) ($p=0.002$). In 2010, the proportion of non-B infections was 47% (Fig. 2).

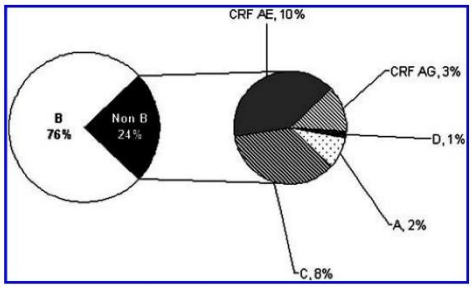


FIG. 1. Proportion HIV-1 subtypes in new diagnoses between 2000 and 2010.

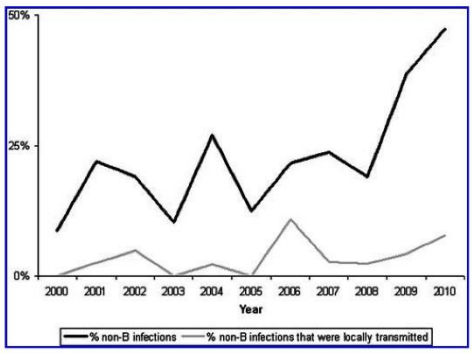


FIG. 2. Proportion of non-B diagnoses in Australia between 2000 and 2010, including local transmission.

Location infection acquired

As well as an increase in non-B subtypes over time, there was a significant increase in the proportion of infections acquired overseas from 23% (2000–2003) to 37% (2007–2010) ($p=0.006$). When analyzed by subtype, 85% of the 101 non-B infections were reported to be acquired overseas and were composed of 95% (34/36) of the total subtype C infections, 81% (36/41) of the CRF_AEs, 77% (10/13) of the CRF_AGs, 89% (8/9) of the As, and 50% of the (1/2) Ds.

In contrast to non-B infections, only 10% of the 324 B infections were acquired overseas ($p<0.0001$).

Among the 72% of cases (305/425) determined to have been acquired within Australia, 95% were clade B (Table 1). However, new local non-B diagnoses have increased from two cases in the first time period to six in the second and seven in the third (Fig. 2). Of these 15 local transmissions, nine were female and six reported being born in Australia. Of the six males who acquired HIV locally, four reported being born in Australia.

Reported risk exposure, gender, and age at diagnosis

The majority of genotyped subjects were men who have sex with men (MSM) (62%), although the proportion has dropped from 72% (2000–2003) to 61% (2004–2006) and 55% in the most recent time period (2007–2010). Correspondingly, the proportion of reported heterosexually acquired HIV increased from 27% (2000–2003) to 35% (2004–2006) and 40% (2007–2010).

Of the 262 genotyped MSM infections, nearly all (97%) were subtype B, and 90% were acquired in Australia. The remaining 10% acquired overseas were nearly all subtype B, but three had CRF_AE. Of the 146 genotyped heterosexual infections, 43% were subtype B (Table 1) and 43% (63/146) were acquired in Australia. The remaining 57% acquired overseas were nearly all (89%) non-B (36% C, 35% CRF_AE, 18% A/D/CRF_AG). The proportion of female infections was significantly higher in the non-B cohort (36% or 36/101) compared to the B cohort (6% or 18/324) ($p<0.0001$).

Approximately one-tenth of genotyped patients (45/425) were under the age of 25 years when diagnosed (32 males and 13 females). The observed frequency of non-B infections in the

<25 years cohort was 64% greater than expected, compared to B infections in the same cohort (20% smaller than expected) ($p=0.01$). The majority (77%) of the 13 young females had a non-B infection, in comparison to 25% of the young males.

Region of origin

People born in Australia constituted 57% (240/425) of the genotyped cohort and subtype B accounted for 88% (211/240) of these infections (Table 1). Of these Australian born subtype B patients, 77% (163/211) were acquired in Australia. In contrast, of the 12% (29/240) of Australian born people with a non-B infection, 66% acquired their infection overseas.

One hundred (24%) people in the genotyped cohort were born overseas while 20% had no recorded country of birth. In the overseas born white population, subtype B accounted for 81% of European born people (21/26), of which 91% were acquired in Australia. Subtype B accounted for 86% of American born people (6/7), but half of these infections were acquired overseas. All non-B infections in both European and American born individuals were acquired overseas.

Nearly all (40/43) African born people had a non-B infection and 98% of these were acquired overseas. The predominant non-B infection was C (51%), followed by CRF_AG (21%), A (16%), CRF_AE (2%), and D (2%). Two-thirds (16/24) of Asian born people had a non-B infection and 75% of these were acquired overseas. The predominant non-B infections were CRF_AE (63%), C (19%), CRF_AG (13%), and A (5%).

Subtype associations with comparator variables

Multivariate logistic regression analyses were conducted, taking subtype B infections as the baseline group for comparison and controlling for other independent variables (acquired overseas, heterosexual, overseas born, and age <25 years).

As shown in Table 2, there was a strong association between subtype C and overseas acquisition, with subtype C infections 34 times more likely than subtype B to be acquired overseas (95% CI 5.8–192.0, $p<0.0001$). Non-B infections were 17 times more likely to be acquired overseas compared to B infections (95% CI 6.9–42.6, $p<0.0001$). Non-B infections were almost 21 times more likely than B infections to be acquired through reported heterosexual transmission (95% CI 7.5–57.0, $p<0.0001$), which rose to 44 times more likely for subtype C cases (95% CI 5.0–384.6, $p=0.001$). Non-B infections were almost four times more likely (95% CI 1.5–9.9, $p=0.006$) among people born overseas, though this was not significant for CRF_AE infections alone (95% CI 0.7–7.9, $p=0.146$). The age at diagnosis of HIV infection was not different between B and non-B patients, except in the "Other" category, which was subtype A, D, and CRF_AG ($p<0.05$). There were 24 patients with A, D, or CRF_AG: four born in Australia, 17 born in Africa, and three born in Asia. Of these, 38% were under 25 years of age at detection of infection; 75% (3/4) from Australia, 24% (4/17) from Africa, and 66% (2/3) from Asia.

Discussion

This is the first Australian study to investigate genotypic diversity and trends in subtype distribution of the HIV epidemic over the past 10 years. At present, no formal national HIV-1 molecular surveillance program exists in Australia. However, South Australia has implemented routine baseline drug resistance testing, providing viral sequence data combined with notification information to yield an enhanced comprehension of South Australian HIV molecular epidemiology.[21] As HIV continues to evolve, migration patterns change, and tourism increases, it is important to monitor this geographic diversity in order to understand and respond to transmission patterns.[5,17]

Despite ongoing programs and improved access to testing and treatment, the rate of new HIV infections remains stable. However, differences were identified in subtype prevalence, place of acquisition, region of birth, gender, and age at diagnosis. The main finding of this study is that the proportion of non-B infections has increased in Australia over the past decade. Non-B patients represented approximately a quarter of the cases in this study, but this proportion had risen to nearly half of all new diagnoses in 2010. This is similar to a European report in which more than 60% of new infections were non-B in 2008.[6]

The current study demonstrates that patients infected with B or non-B subtypes represent highly distinct populations. The majority of B-infected people were MSM, while non-B infections were mainly heterosexually acquired and just over a third of were female. Most B infections were acquired in Australia by people born in Australia while the majority of non-B infections were acquired overseas by people born overseas or people traveling to areas of high HIV prevalence.[5,6] These figures are comparable to findings by Chalmet *et al.* (2010) in Belgium, in a similar sized cohort.[22]

Table 2. Multivariate Odds Ratios in Newly Diagnosed HIV Patients, 2000–2010, When Compared to the Proportion of Subtype B Infections

	Characteristics[a] *(comparator group)*			
Clade	*Acquired overseas (acquired Australia)*	*Heterosexual (MSM)*	*Overseas born (Australian born)*	*Age <25 (≥25)*
Non-B	17.1 (6.9–42.6)	20.6 (7.5–57.0)	3.8 (1.5–9.9)	3.8 (0.9–15.1)
CRF_AE	17.1 (5.9–49.1)	17.0 (5.1–57.0)	2.4 (0.7–7.9)	2.0 (0.1–14.3)
C	33.5 (5.8–192.0)	44.0 (5.0–384.6)	5.4 (1.2–24.2)	3.4 (0.3–42.8)
Other	7.3 (1.2–46.4)	44.1 (4.6–420.7)	22.5 (2.9–175.4)	27.6 (2.6–295.8)

[a]95% CIs are indicated in parentheses.

Proportion of subtype B infections was used as the baseline for comparison. All associations in the first three columns were statistically significant at $p<0.05$ except born overseas and CRF_AE ($p<0.146$). In the age column, only the association between other and age <25 was significant ($p<0.05$). Non-B includes all non-B subtypes and CRFs. Other consists of subtypes A, D, and CRF_AG. MSM, men who have sex with men.

White MSM with locally acquired subtype B continue to represent the largest HIV risk group in South Australia, supporting epidemiological trend analyses in Australia, Canada, Europe, and the United States.[15,21–25] There is also a small subset of MSM acquiring non-B infections overseas and locally, reflecting findings in the U.K.[8] and the recent Victorian study.[20] Though national HIV incidence has remained stable in the MSM population over the past decade, the proportion of undiagnosed infections in South Australia is estimated to be around 20%.[26]

On a global scale, heterosexual transmission is the major route of HIV infection.[27] In South Australia we are seeing a shift in the epidemic toward this: a decrease in the local representation of MSM and an increase in the representation of heterosexual men and women, including a large proportion from sub-Saharan Africa. This shift has also been noted in Europe.[28] Since 2000, heterosexually acquired infections have doubled to nearly half of all new diagnoses, not quite as high as the 4-fold increase seen in the U.K. since 1996.[21] Most non-B patients contracted HIV by heterosexual contact, and a small number through male-to-male sex. The proportion of women infected with subtype B was very low; this can be partly explained by the high representation of MSM in the B cohort, while the majority of females acquired their infection overseas where non-B subtypes are more frequently circulating. Despite the apparent shift in the South Australian epidemic toward increasing heterosexually acquired non-B HIV infections, MSM are still a critical at-risk population that justifies a continued prevention and intervention focus for scientists, program experts, and policy makers.[15]

Corresponding with an increase in heterosexual infections, there has been a significant increase in imported non-B infections, through migration or travel to countries where there is a high prevalence of HIV. Though the numbers are small, there is evidence of non-B local transmission, predominantly found in females. A significantly higher proportion of non-B patients was diagnosed under the age of 25 compared to the B cohort, though when location acquired, country of birth, and risk behavior were controlled for, this was significant only for subtype A, D, and CRF_AG patients. These clades are common in Africa and Asia, where two-thirds of these young patients were born. Patients born in Africa reported sexual contact as their risk exposure while the two patients born in Asia were both under 10 years of age, with overseas medical procedure cited as the risk exposure.

Transmission of non-B subtypes and CRFs is rapidly expanding geographically, and the rise in non-B diagnoses may be a marker of more recent transmission events—some attributed to tourism and some to importation by people born in high prevalence countries where multiple subtypes and CRFs circulate.

Over half of the global HIV population is infected with subtype C,[4,29,30] which is dangerously uncontrolled in Africa and India. Subtypes A and B follow, then CRF_AG and CRF_AE, the latter predominantly found in Asia.[9,30] In our cohort, subtype C accounted for fewer than 10% of total infections but over a third of non-B infections. Nearly all were reportedly acquired overseas, and though information on a specific country was not available, a large percentage of these patients originated from Africa.

The predominant non-B infection was CRF_AE, with prevalence in the genotyped cohort twice that of the 5% global average, and almost all cases reportedly were acquired overseas. Hemelaar *et al.* found global CRF infections increased by over 50% between 2000 and 2007,[9,30] and the current study reflects this temporal CRF increase; almost half the AE infections were diagnosed in 2007–2010. Unlike subtype C, nearly half the patients were born in Australia and were most likely the result of acquisition during overseas travel to Asia.

Subtypes A, D, and CRF_AG are predominantly found in Africa, with a combined global prevalence rate of 10%.[9] Though prevalence was relatively low in our cohort, nearly all were diagnosed in the latter time period, and were acquired overseas by people of African origin, possibly reflecting the increase in Australian migration from this region. A number of Australian born heterosexual men and women also imported or locally acquired subtypes D and CRFs AE and AG.

These findings have public health implications, both for targeting specific at-risk populations and assessing the potential increase of non-B subtypes within the domestic HIV-1 epidemic.[1]

There are scarce data available on subtype differences and even fewer data available on non-B subtypes in developing countries where they are the major infection type. There is growing evidence, however, that suggests HIV strains do differ from each other in terms of virulence, transmission, or rate of progression.[4] A 10-year prospective study in Senegal found female sex workers with a non-A subtype had a significantly shorter AIDS-free survival time.[31] A 2010 London study found a CD4 cell decline 4-fold faster in subtype D patients, and a higher virological rebound at 6 months, after adjustment for baseline, gender, and ethnicity.[21] A study of Kenyan women found a >2-fold higher risk of mortality and faster rate of CD4 cell decline in D patients compared with A, after adjustment for viral load,[32] and in a Ugandan cohort, subtype D patients tended to develop AIDS earlier.[33] In Rakai, the median time to onset and risk of progression to death were significantly shorter for subtype D and CRF patients compared with A.[11] Each of these studies concluded that HIV disease progression is affected by subtype and that this may have an impact on decision and policy making in terms of initiation of therapy and future vaccine trials.[11,21,32]

Understanding genetic diversity is very important for the treatment of non-B subtypes. Many researchers now agree that though it appears subtypes and CRFs are equally sensitive to treatment, transmitted polymorphisms present before therapy may affect subtype-specific pathways of secondary resistance.[12] This combined with suboptimal therapy and poor adherence in developing countries makes them a prime target for accelerated drug resistance, both acquired and transmitted.[34] Current drug regimens targeted against subtype B may not be equally effective long term for non-B subtypes and may lead to faster drug resistance.[27]

Interpreting and reporting surveillance data can be problematic. Reporting newly acquired infections does not demonstrate true reductions or increases in the wider community; HIV diagnoses represent only the subgroup of people who have willingly been tested and had an HIV-positive result. These individuals are often those who have easy access to medical health services and are concerned about their own risk behavior.[35] Immigrants, visa holders, and refugees still face barriers to accessing health services for screening and treatment of HIV, arising from stigma, financial restrictions,

limited support systems and English skills, and residency concerns.[28,35] Refugees in particular may be difficult to reach because of traumatic life experiences prior to arrival in Australia.

This is a major concern as the UN recognizes migrants as one of the groups most vulnerable to HIV, and overseas born people now comprise a third of HIV notifications in Australia.[28,36] These issues along with a continuing influx of new arrivals from high HIV prevalence and low/middle income countries are likely to lead to an underestimate of HIV infections in these populations, a possible increase in local transmission of non-B subtypes, and poor treatment adherence, which could lead to transmitted drug resistance.[1,36]

The global spread of HIV diversity is highly dynamic with regard to epidemiological factors such as risk group and geographic location; it continually generates through mutation and recombination, and then travel and migration assist in the transfer of this diversity between populations.[21] Our analyses focused exclusively on *pol*, as it is routinely used for drug resistance testing, and provided the largest possible reference dataset of B and non-B subtype sequences. We note that further subtype validation should be conducted with alternative HIV genes, such as *env*.[37] Ongoing surveillance and a deeper understanding of HIV variation, including factors and molecular mechanisms that affect transmission, replication, and resistance, are crucial for the development of appropriately targeted subtype-specific prevention and treatment options for populations most at risk.[5,9,21,29] Further evidence of subtype differences could drastically change the way we respond to the HIV epidemic.

Acknowledgments

The authors would like to thank Michael Kidd, Peter McDonald, Stuart Shapiro, and Bob Seamark for helpful discussions and comments during the course of this project. K.G.H. is supported by an Australian Post Graduate Award. Presented in part at the Australasian Society for HIV Medicine Annual Scientific meeting Canberra, ACT, Australia, September 2011.

Author Disclosure Statement

No competing financial interests exist.

References

1. Oelrichs R: The subtypes of human immunodeficiency virus in Australia and Asia. Sex Health 2004;1:1–11.
2. Plantier JC, Leoz M, Dickerson JE, *et al.*: A new human immunodeficiency virus derived from gorillas. Nat Med 2009; 15:871–872.
3. Lihana RW, Ssemwanga D, Abimiku A, and Ndembi N: Update on HIV-1 diversity in Africa: A decade in review. AIDS Rev 2012;14:83–100.
4. Ndung'u T and Weiss RA: On HIV diversity. AIDS 2012; 26:1255–1260.
5. Butler IF, Pandrea I, Marx PA, and Apetrei C: HIV genetic diversity: Biological and public health consequences. Curr HIV Res 2007;5:23–45.
6. Lai A, Riva C, Marconi A, *et al.*: Changing patterns in HIV-1 non-B clade prevalence and diversity in Italy over three decades. HIV Med 2010;11:593–602.
7. Liu Y, Li L, Bao Z, *et al.*: Identification of a novel HIV type 1 circulating recombinant form (CRF52_01B) in Southeast Asia. AIDS Res Hum Retroviruses 2012;28(10): 1357–1361.
8. Fox J, Castro H, Kaye S, *et al.*: Epidemiology of non-B clade forms of HIV-1 in men who have sex with men in the UK. AIDS 2010;24:2397–2401.
9. Hemelaar J, Gouws E, Ghys PD, Osmanov S, and WHO/UNAIDS Network for HIV Isolation and Characterisation: Global trends in molecular epidemiology of HIV-1 during 2000–2007. AIDS 2011;25:679–689.
10. Kantor R, Katzenstein DA, Efron B, *et al.*: Impact of HIV-1 subtype and antiretroviral therapy on protease and reverse transcriptase genotype: Results of a global collaboration. Plos Med 2005;2:325–337.
11. Kiwanuka N, Laeyendecker O, Robb M, *et al.*: Effect of human immunodeficiency virus type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. J Infect Dis 2008;197:707–713.
12. Taylor BS and Hammer SM: The challenge of HIV-1 subtype diversity. N Engl J Med 2008;359:1965–1966.
13. Carr JK, Osinusi A, Flynn CP, Gilliam BL, Maheshwari V, and Zhao RY: Two independent epidemics of HIV in Maryland. J Acquir Immune Defic Syndr 2010;54:297–303.
14. Chen JHK, Wong KH, Chen ZW, *et al.*: Increased genetic diversity of HIV-1 circulating in Hong Kong. Plos One 2010; 5:e12198.
15. Sullivan PS, Hamouda O, Delpech V, *et al.*: Reemergence of the HIV epidemic among men who have sex with men in North America, Western Europe, and Australia, 1996–2005. Ann Epidemiol 2009;19:423–431.
16. Lau KA, Wang B, and Saksena NK: Emerging trends of HIV epidemiology in Asia. AIDS Rev 2007;9:218–229.
17. Dwyer DE, Ge YC, Bolton WV, Wang B, Cunningham AL, and Saksena NK: Subtype B isolates of human immunodeficiency virus type 1 detected in Australia. Ann Acad Med Singapore 1996;25:188–191.
18. Descamps D, Chaix ML, Montes B, *et al.*: Increasing prevalence of transmitted drug resistance mutations and non-B subtype circulation in antiretroviral-naive chronically HIV-infected patients from 2001 to 2006/2007 in France. J Antimicrob Chemother 2010;65:2620–2627.
19. Wheeler WH, Ziebell RA, Zabina H, *et al.*: Prevalence of transmitted drug resistance associated mutations and HIV-1 subtypes in new HIV-1 diagnoses, US-2006. AIDS 2010;24: 1203–1212.
20. Chibo D and Birch C: Increasing diversity of human immunodeficiency virus type 1 subtypes circulating in Australia. AIDS Res Hum Retroviruses 2012;28:578–583.
21. Easterbrook PJ, Smith M, Mullen J, *et al.*: Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. J Int AIDS Soc 2010;13:4.
22. Chalmet K, Staelens D, Blot S, *et al.*: Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis 2010;10:262.
23. Gifford RJ, Oliveira T, Rambaut A, *et al.*: Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. J Virol 2007;81:13050–13056.
24. Guy RJ, McDonald AM, Bartlett MJ, *et al.*: Characteristics of HIV diagnoses in Australia, 1993–2006. Sex Health 2008;5: 91–96.

25. Paraskevis D, Pybus O, Magiorkinis G, *et al.*: Tracing the HIV-1 subtype B mobility in Europe: A phylogeographic approach. Retrovirology 2009;6:49.
26. Mallitt K-A, Wilson DP, McDonald A, and Wand H: HIV incidence trends vary between jurisdictions in Australia: An extended back-projection analysis of men who have sex with men. Sex Health 2012;9:138–143.
27. Kilmarx PH: Global epidemiology of HIV. Curr Opin HIV AIDS 2009;4:240–246.
28. Del Amo J, Likatavicius G, Perez-Cachafeiro S, *et al.*: The epidemiology of HIV and AIDS reports in migrants in the 27 European Union countries, Norway and Iceland: 1999–2006. Eur J Public Health 2011;21: 620–626.
29. Iordanskiy S, Waltke M, Feng YJ, and Wood C: Subtype-associated differences in HIV-1 reverse transcription affect the viral replication. Retrovirology 2010;785.
30. Salemi M: Toward a robust monitoring of HIV subtypes distribution worldwide. AIDS 2011;25:713–714.
31. Kanki PJ, Hamel DJ, Sankale JL, *et al.*: Human immunodeficiency virus type 1 subtypes differ in disease progression. J Infect Dis 1999;179:68–73.
32. Baeten JM, Chohan B, Lavreys L, *et al.*: HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. J Infect Dis 2007;195:1177–1180.
33. Kaleebu P, Ross A, Morgan D, *et al.*: Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. AIDS 2001;15:293–299.
34. Derache A, Traore O, Koita V, *et al.*: Genetic diversity and drug resistance mutations in HIV type 1 from untreated patients in Bamako, Mali. Antivir Ther 2007;12:123–129.
35. Horyniak D, Stoove M, Yohannes K, *et al.*: The impact of immigration on the burden of HIV infection in Victoria, Australia. Sex Health 2009;6:123–128.
36. McPherson ME, McMahon T, Moreton RJ, and Ward KA: Using HIV notification data to identify priority migrant groups for HIV prevention, New South Wales, 2000–2008. Commun Dis Intell 2011;35:185–191.
37. Dalai SC, de Oliveira T, Harkins GW, *et al.*: Evolution and molecular epidemiology of subtype C HIV-1 in Zimbabwe. AIDS 2009;23:2523–2532.

Address correspondence to:
Karen Hawke
Discipline of Public Health
Flinders University
GPO Box 2100
Adelaide 5000
South Australia

E-mail: karen.hawke@health.sa.gov.au