Molecular Epidemiology of HIV in Canada

Manon Lily Ragonnet

Supervisor: Dr. Stéphane Aris-Brosou

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
University of Ottawa
In partial fulfillment of the requirements for a
Master's degree from the
Ottawa-Carleton Institute of Biology

Thèse soumise à la
Faculté des Etudes Supérieures et Postdoctorales
Université d'Ottawa
En vue de l'obtention de la maîtrise
L'Institut de Biologie d'Ottawa-Carleton

Acknowledgements

For my thesis work, I have had the extraordinary good fortune to work with two experts in their fields: Stéphane Aris-Brosou and James Brooks. I cannot overstate how indebted I am to both of them.

Stéphane is a mine of information on every topic and has offered me constructive guidance in every step of this thesis. I could never have learned so much in the last year without him. His patience has known no bounds, and he has helped me tackle everything from complicated concepts to superficial details.

To James, I am grateful for taking me under his wing for the last three years and acting as my mentor. I would not be continuing on to a PhD in this field if it were not for him. He opened the world of HIV research and Public Health to me, and generously shared his extensive wisdom on these topics and many other unrelated ones. I have enjoyed working in his laboratory tremendously.

I appreciate the efforts of both Stéphane and James in encouraging me to present my work and sending me to so many conferences, making my research experience fun and fulfilling. I also thank the members of my committee, Drs Robert Smith? and Nicolas Corradi for their time and advice.

I am also extremely grateful to my friend and colleague Anna van Weringh. She has substantially contributed to all of my work by helpful discussions and by reading everything I wrote as many times as I have myself. Anna has also supported me morally throughout my Masters', and it would not have been as enjoyable without her. I hope we can continue to share ideas even after we go our separate ways. I thank both Anna and Nick Petronella for so generously writing wonderful Perl scripts for me.

I would like to extend my gratitude to many other members of the National HIV and Retrovirology Laboratory, for taking me in and being so generous with their time, knowledge and friendship. None of this work would have been possible without all the lab work carried out by Harriet Merks and Isabelle Joanisse. Harriet, in particular, has been incredibly thorough in addressing every technical issue I have come across during my work, locating sequences and data that I would have given up on. I would also like to thank Richard Pilon and Paul Sandstrom, who have helped and supported me throughout the last three years.

Of course I thank my parents for their love and encouragement to do whatever I want, as well as their financial support throughout my education. I am also thankful to my partner Dave for standing by me when I was excited about my work and when I wasn't. He also helped me with math and programming along the way, as well as with some last minute graphic design. I am grateful to his family for sharing their home and holidays with me while I have been so far from my own.

Finally I would like to acknowledge funding from NSERC, the Canada Foundation for Innovation, and the Federal Initiative on HIV/AIDS in Canada, as well from the Association of Commonwealth universities who made this last year so much more comfortable than the previous one thanks to their generous scholarship.

Abstract

With over 35 million people currently infected, the World Health Organization considers HIV a global pandemic. HIV is characterized by a high mutation rate, which allows it to evade the host immune system and develop resistance to drugs. However, this extraordinary adaptive ability may also be the key to HIV's demise. Through the field of phylodynamics, the evolutionary behavior of the virus is being studied in an attempt to control the epidemic. In this thesis, three papers are presented in which we analyze sequences generated through the Canadian HIV Strain and Drug Resistance Surveillance program. In chapter 2 we validate a classifier which distinguishes between recent and established infections based on the proportion of mixed bases observed in population-based *pol* sequences. Our results will help identify recent infections and improve incidence calculations. In chapter 3, we investigate immune-induced patterns in HIV that are shared by patients of the same ethnicity. An understanding of the forces shaping HIV evolution is instrumental to the development of a vaccine relevant to the Canadian epidemic. In chapter 4, we present preliminary results of a historical reconstruction of HIV across the provinces of Canada. This analysis will highlight strategies that have succeeded or failed in controlling the epidemic. Furthermore, our work will establish whether non-B subtypes of HIV are an increasing threat to Canadian public health. Overall, this thesis provides the first country-wide evolutionary and phylogenetic analysis of the HIV epidemic.

Résumé

Avec plus de 35 millions de personnes infectées, l'Organisation Mondiale de la Santé considère le VIH comme étant une pandémie mondiale. Le taux de mutation du VIH est particulièrement élevé, ce qui lui permet d'échapper au système immunitaire de l'hôte et de développer une résistance aux traitements. Cependant, cette extraordinaire capacité d'adaptation pourrait également être son talon d'Achille. Au cours de recherches phylodynamiques, le comportement évolutif du virus est étudié avec pour but de contrôler l'épidémie. Dans cette thèse, trois papiers sont présentés dans lesquels nous analysons les séquences générées par le programme sur les Souches VIH et la Pharmacorésistance Primaire au Canada. Dans le chapitre 2, nous validons un classificateur qui distingue entre les infections récentes et chroniques en fonction de la proportion de bases mixtes observées dans les séquences *pol*. Nos résultats permettront d'identifier les infections récentes et d'améliorer les calculs d'incidence. Dans le chapitre 3, nous étudions les mutations induites dans les séquences *pol* par le système immunitaire de patients de la même ethnie. Une bonne compréhension des pressions sélectives entraînant l'évolution du VIH est essentielle pour le développent d'un vaccin pertinent à l'épidémie Canadienne. Dans le chapitre 4, nous présentons des résultats préliminaires d'une reconstruction historique du VIH à travers les provinces du Canada. Cette analyse mettra en évidence les stratégies de contrôle de l'épidémie qui ont réussi ou qui ont échoué. Par ailleurs, cette étude établira si les souches non-B du VIH sont une menace pour la santé publique du Canada. Globalement, cette thèse fournit la première analyse évolutive et phylogénétique du VIH sur l'ensemble du pays.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AB	Alberta
AIDS	Acquired Immune Deficiency Syndrome
ART	Antiretroviral Therapy
AZT	Azidothymidine
BC	British Columbia
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
BED	Calypte Enzyme Immunoassay (CEIA), developed for HIV subtypes B, E and D
BF	Bayes' Factor
bp	base pairs
BSP	Bayesian Skyline Plot
CRF	Circulating Recombinant Form
CTL	Cytotoxic T Lymphocyte Response
DNA	Deoxyribonucleic Acid
DRM	Drug Resistant Mutation
dsDNA	double stranded DNA
HAART	Highly Active Antiretroviral Therapy
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
IDU	Intravenous Drug Users/ Intravenous Drug Use
IgG	Immunoglobulin G
MB	Manitoba
MCMC	Markov chain Monte Carlo
ML	Maximum Likelihood
MRCA	Most Recent Common Ancestor
MSM	Men who have Sex with Men
MTCT	Mother to child transmission
NEP	Needle Exchange Program
NHRL	National HIV and Retrovirology Laboratory
NLHG	National Laboratory for HIV Genetics
NLHRS	National Laboratory for HIV Reference Services
NNRTI	Non-Nucleoside Reverse Transcriptase Inhibitor
N	Non-Synonymous
NS	Nova Scotia
PHAC	Public Health Agency of Canada
PCR	Polymerase Chain Reaction
PR	Protease
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
S	Synonymous

SDR	Strain and Drug Resistance
SIV	Simian Immunodeficiency Virus
SK	Saskatchewan
STI	Sexually Transmitted Infection
TDR	Transmitted Drug Resistance
WHO	World Health Organization

Chapter 1

Background

1.1 HIV Epidemiology

1.1.1 Worldwide distribution of HIV

The Human Immunodeficiency Virus (HIV) is the etiologic agent of Acquired Immunodeficiency Syndrome (AIDS). The immune system of individuals infected with HIV is progressively destroyed by the virus until they are no longer able to fight off opportunistic infections. The World Health Organization (WHO) considers HIV a global pandemic, with 33.4 million people infected in 2009, an estimated 2.7 million new infections every year and more than 2 million people dying each year from AIDS [1]. HIV infections are unequally distributed across the globe with low and middle income countries bearing the heaviest of the burden. Two thirds of infections, totaling 22.4 million people, exist within Sub-Saharan Africa. The number of people living with HIV continues to rise, although the ascent may have peaked in 1996 when an unparalleled 3.5 million people were newly infected. In addition to being a leading cause of death in many developing countries, HIV is a major economic and social burden, greatly reducing working life expectancy and leaving millions of children orphaned. Moreover, HIV is a highly stigmatized condition due to its associations with homosexuality, drug use and race.

Despite being the most intensely studied infectious agent, thirty years into the epidemic, HIV has no cure and continues to be a major global heath priority. Important advances have nevertheless been made in treating HIV and preventing transmission (section 1.3).

1.1.2 HIV in Canada

In developed countries HIV is concentrated within risk groups, specifically amongst men who have sex with men (MSM) and intravenous drug users (IDU). In Canada, approximately 73,000 people are living with HIV (min 43,000 max 110,000) [2]. Aboriginals are disproportionately affected, comprising only 4% of the population but 8% of prevalent infections. Aboriginal women accounted for 40.3% diagnoses among females in 2008. The latest HIV statistics reveal that MSM account for the largest proportion of positive HIV test reports (45.1%) followed by IDU (19.1%) [3]. The Canadian epidemic is costly, but most importantly marginalizes already stigmatized groups. The recent criminalization of HIV, with the prosecution of several individuals living with HIV for exposing their partners to significant risk, adds another layer to the HIV debate.

1.1.3 HIV transmission

The dominant route of HIV transmission is through sex. HIV can also be transmitted from mother to child (MTCT), and through contact with infected blood or blood products. In sub-Saharan Africa the main mode of transmission is heterosexual sex, while in most other regions the predominant risk factors are sex between men and the use of contaminated needles.

1.1.3.1 Sexual transmission of HIV

During intercourse, the receptive partner is at higher risk for HIV than the insertive partner [4] making women at higher risk in heterosexual relationships than men [5]. Similarly in MSM, the receptive partner is also at higher risk.

Longitudinal analyses of heterosexual sero-discordant couples (where one partner is positive and the other negative) in Uganda have measured the risk of HIV transmission per sexual act to be around 1 in 1,000 [6]. A recent meta-analysis however, calculated estimates 2-8 times higher across developing countries [7]. In the developed world, risk of sexual transmission from female to male in the developed world was 0.04%, and from male to female, 0.08% [7]. It is generally accepted that during intercourse MSM have a higher risk of contracting HIV than heterosexuals, in part due to the higher risk of transmission during anal sex, 1.7% in pooled estimates [7]. Moreover, the higher prevalence of HIV

among MSM and possibly increased risk-taking behavior in this group [8] also contribute to this increased risk.

The risk of sexual transmission of HIV is further increased in the presence of other sexually transmitted infections [9]. If one partner has genital ulcer disease for example, HIV transmission risk is increased fivefold [7]. In addition, transmission of HIV is highly dependent on the viral load of the infected partner [10]. In turn, viral load is dependent on numerous factors, including stage of infection [11]. In particular viral load is very high during the first few weeks following infection, thus a large proportion of transmissions are believed to originate from recently infected patients. For all epidemiological studies on HIV transmission it is essential to distinguish between newly diagnosed infections that are recent and those that are chronic. In Chapter 2, I develop a method for classifying infections as recent or chronic based on within host genetic diversity.

1.1.3.2 Mother to child transmission of HIV

MTCT may occur at three stages: during pregnancy, during delivery or through breastfeeding. In the absence of treatment, the total risk across these stages is approximately 25% [12], with MTCT accounting for ~10% of new infections annually. In developed countries, MTCT has been practically eliminated by controlling the mother's HIV infection with antiretroviral treatments (further discussed in section 1.3).

1.1.3.3 Transmission of HIV through infected blood and needle sharing

Blood transfusion is the most efficient route of HIV transmission; in a retrospective analysis, 90% of HIV+ blood recipients became infected [13]. Before testing of donor blood was implemented this resulted in 14,000 new infections in Europe and the USA [14]. Hemophiliacs were overrepresented among these infections [15]. In countries where donor blood is tested, HIV transmission through the blood supply is nearly non-existent. Nevertheless, there are countries where blood remains untested and continues to contribute to new HIV infections. Across the African continent for example, an estimated 5-10% of new infections occur as a consequence of unsafe blood transfusions [16].

Exchange of blood through sharing needles among IDU is a very common route of HIV transmission. IDU, their partners and children account for one third of AIDS cases in the USA [17]. Worldwide, 10% of new infections occur through IDU, although in some regions it encompasses a much more significant portion,

as in Eastern Europe where over 80% of infections are related to drug use (http://www.usaid.gov/our_work/global_health/aids).

1.2 Molecular virology

1.2.1 Disease progression

Two aspects of HIV biology make it a particularly challenging infection. First, HIV has a long period of clinical latency during which patients show no symptoms of being infected. This interval lasts up to ten years without treatment, a period during which many onward transmissions can take place. Second, the virus directly attacks the CD4 T cells responsible for fighting off infection, leading to the progressive failure of the immune system.

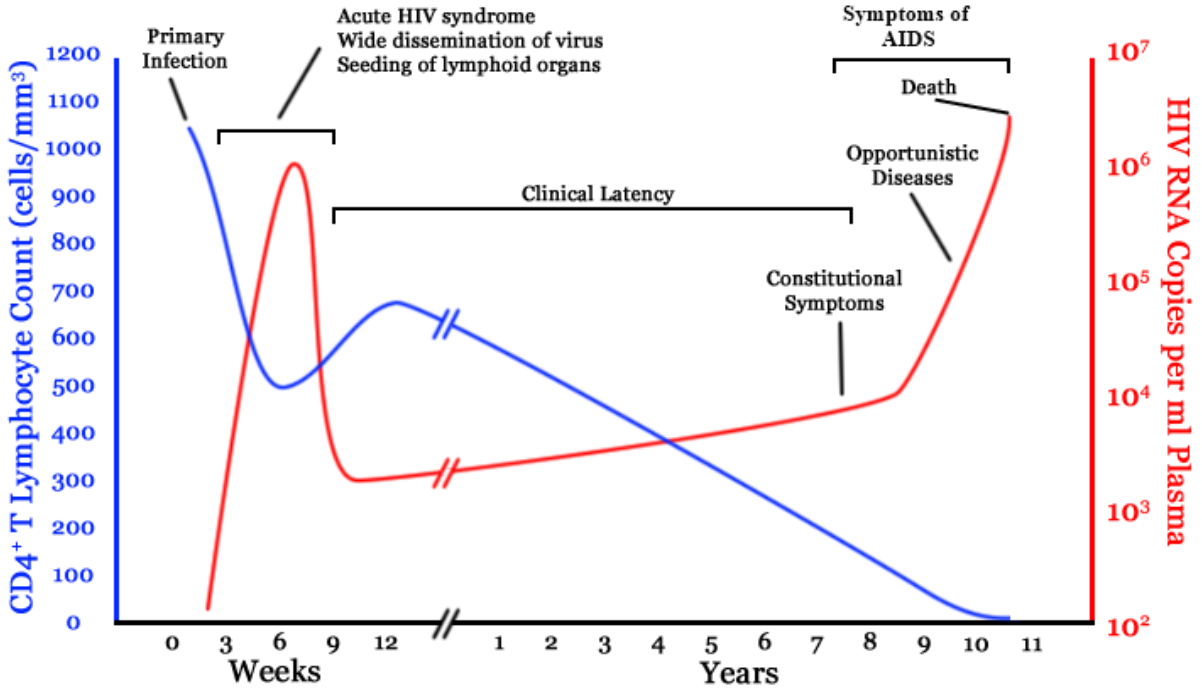


Figure 1.1: HIV course of infection. CD4 counts and HIV copies in an untreated patient. By Jurema Oliveira, and based on Figure 1 in Pantaleo *et al* [18]. Via Wikimedia Commons, GNU Free Documentation License.

The progression from HIV to AIDS can be divided into three stages (Figure 1.1). The initial infection stage lasting 2-4 weeks is characterized as the “acute” infection stage. During acute infection the virus

replicates rapidly and generic symptoms associated with immune activation may be experienced, such as a sore throat, rash and flu-like symptoms. Viral load increases dramatically (10^7 viral RNA copies/mL blood), likely contributing to high infectivity during this stage of infection [19]. Acute infection subsides once host adaptive immune response is initiated in the form of immunoglobulins of subtype G (IgG) specific against HIV proteins. Targeting viral particles for destruction, the appearance of IgG lowers viral loads. The following 8-10 years are known as the latent period of HIV infection, during which the patient shows no symptoms. At the start of latency the immune system is able to control the infection, but fails as infection persists. Concomitantly, viral load increases gradually, correlating with the decline of CD4 T cells and a progressive failure of the immune system. Eventually, the CD4 cell count drops below a critical level, around 200 cells/mm^3, whereupon effective cell-mediated immunity is lost. It is at this point that patients are considered to have developed AIDS and may contract opportunistic infections. Once the AIDS stage has been reached patients usually succumb to disease within a year [20].

1.2.2 Classification and life cycle

HIV is classified as a member of the retroviridae family in the lentiviral subfamily. The retroviral family is characterized by a single-stranded positive-sense RNA genome. The virus particle is spherical and 120nm in diameter and within each particle are two copies of the full genome. The genome is ~10,000bp and contains nine genes: *gag, pol, env* (structural), *tat, rev, nef, vif, vpr* and *vpu* (regulatory), which encode 19 proteins (Figure 1.2).

Target cells are those vital to the human immune system: CD4 T cells, macrophages and dendritic cells. The CD4 receptor and co-receptors CCR5 or CXCR4 are essential for HIV binding. Upon entry into cells, the viral RNA genome in converted to double stranded DNA (dsDNA) by the viral enzyme reverse transcriptase (RT). The dsDNA then integrates into the host cellular DNA using the viral protein integrase with a preference for active transcription units [21]. At this stage the virus can enter either the latent or the lytic pathway. During viral latency (not to be confused with clinical latency), the integrated virus lies dormant within the host DNA, and may be expressed at low levels, or not at all. The virus can reactivate at any time and enter the lytic pathway, characterized by high levels of viral gene expression and production of viral particles that are released from infected cells. Factors controlling entry into and out of latency remain poorly understood but include the site of integration into host DNA [22].

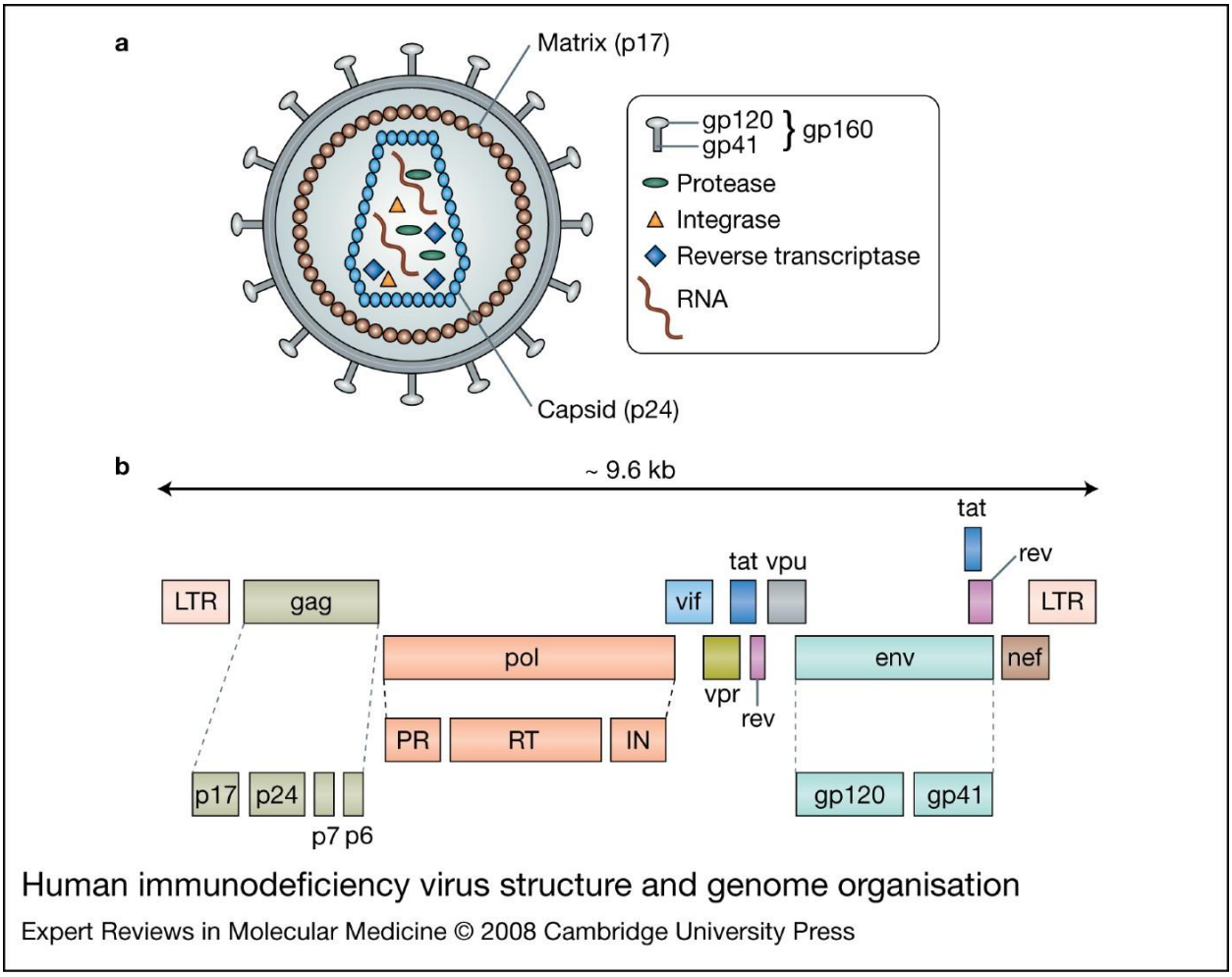


Figure 1.2: HIV structure (a) and genome organization (b). See text for details. Reproduced with permission from Roe *et al* [23]. ©2008 Cambridge University Press.

1.2.3 Viral replication and mutation

The reverse transcriptase (RT) of HIV, which transcribes the single-stranded RNA genome into dsDNA, is an error-prone enzyme. In contrast to the DNA polymerases RT has no proofreading mechanism, making one error every 1,000-10,000 bases [24]. As a consequence 5 to 10 nucleotide misincorporations are found per new HIV genome synthesized. In addition HIV has an extremely high replication rate. Based on measurements of the half-life of HIV virions (1-2 days) [25], it is estimated that in the absence of treatment infected individuals generate on average 10^{10} virions per day [26]. With its high replication and mutation rates combined the evolutionary rate of HIV is 1 million times that of humans [27].

As a result of its extreme mutation rate HIV displays high genetic variability both within a host and at the level of the population. While patients are usually infected with a single founder virus genetic diversity subsequently increases during the course of infection, an observation that I harness in Chapter 2. This ever-present variation within a host allows HIV to evolve rapidly, evading the human immune system and developing mutations that confer drug resistance. Furthermore, the high genetic variability at the level of the population is one reason why the search for a vaccine that would protect against a sizeable proportion of circulating isolates is extremely difficult.

1.3 Treatment and prevention

1.3.1 HIV treatment

Azidothymidine (AZT) became the first approved antiretroviral treatment (ART) for HIV in 1987. AZT competes with endogenous thymidine and terminates reverse transcription. As AZT is missing a 3'-hydroxyl group, the DNA chain cannot be extended beyond it. Although AZT also inhibits the human DNA polymerase it has a 100 fold higher affinity for RT, and has a more limited effect on normal human DNA replication.

However AZT does not halt viral replication; it only inhibits it and delays disease progression. More importantly, RT mutations allow the enzyme to develop resistance to the drug. AZT-resistant RT preferentially incorporates thymidine and not AZT. In order to prevent the emergence of drug resistant mutations within HIV by more efficiently suppressing viral replication, over 20 additional drugs have been developed. Drugs fall into each of seven classes: (i) nucleoside reverse transcriptase inhibitors (NRTIs), (ii) nucleotide reverse transcriptase inhibitors (NtRTIs), (iii) non-nucleoside reverse transcriptase inhibitors (NNRTIs), (iv) protease inhibitors, (v) fusion inhibitors (vi) co-receptor inhibitors, and (vii) integrase inhibitors [28]. Drug classes (i), (ii) and (iii) all target the reverse transcription of HIV, while other classes target different stages of the HIV life cycle. Since 1995, combination therapy is prescribed, containing three or more anti-HIV drugs from at least two classes. This treatment regimen is referred to as Highly Active Anti-Retroviral Treatment (HAART). HAART significantly decreases the morbidity of HIV and the number of AIDS cases. HAART does not cure HIV but has changed it to a chronic disease: viral loads are lowered, optimally to undetectable levels, and CD4 cell count increases.

Current WHO guidelines recommend that ART should be initiated in patients whose CD4 cell count has fallen below 350 cells/mm3 [29]. In clinical trials commencing drug regimens earlier demonstrates clear benefits [30], but this approach poses problems. First, worldwide, there are huge problems of access to ART because of its cost. In sub-Saharan Africa only 17% of eligible patients have access to ART. Secondly, even in regions where people have access there are many problems associated with adherence to HIV treatment. For numerous reasons including side effects and the complexity of drug regimens, patients do not adhere to their ART regimens, decreasing ART efficacy and promoting the development of drug resistant mutations. For these reasons it is not necessarily better to start ART as early as possible.

1.3.2 HIV prevention

As discussed previously, there are three routes for HIV transmission: sexual, MTCT, and through infected blood. Increasing evidence supports the hypothesis that HAART, by suppressing viral loads, decreases risk of all modes of HIV transmission [31, 32]. In addition each transmission route requires dedicated prevention tools.

Condoms are the most efficacious prevention tool against sexual HIV transmission [33] and have been available since the beginning of the epidemic but have not been successful in hindering the spread of HIV due to lack of compliance. As the presence of other sexually transmitted infections (STI) increases HIV transmission, STI treatment also decreases HIV transmission [9]. More recently male circumcision has been demonstrated to decrease HIV acquisition in HIV negative men from HIV positive women [34-37]. Male circumcision hence holds huge promise for the prevention of heterosexual HIV transmission. In MSM, circumcision does not offer any protection as HIV is more likely acquired during receptive anal sex [38].

In developed countries mother to child transmission has been practically eliminated by treating expectant mothers and newborn infants with HAART, delivering babies by caesarean and avoiding breastfeeding. Single-dose Nevirapine (an NNRTI) given to both mother and newborn has been demonstrated to significantly decrease HIV transmission during delivery in developing countries [39]. In many areas however breast milk is the only food available for the baby. Unavoidable breastfeeding can alone increase overall transmission to 18-24% [40].

The size of the HIV epidemic among IDUs in different countries is a direct consequence of that country's policy towards harm reduction and legislation surrounding drug use. Harm reduction is a strategy which

involves reducing health risks when eliminating them is not possible. For example, the benefit of needle exchange programs (NEP) in reducing HIV incidence among IDU has been demonstrated in many different cities around the world [41]. This leading example of harm reduction raises fears however that overall drug use will increase, or that needles will be discarded improperly. To the contrary, not only are NEPs an effective and inexpensive public health intervention, they also help IDUs access treatment for their drug use and for HIV if they are already infected by reducing risky behavior [42]. Such programs are illegal in many countries; in the USA, federal funding of NEP was prohibited for many years but has been allowed since December 2010.

Finally, new prevention options are currently being explored. In the last year, two successful clinical trials demonstrated that HIV acquisition could be decreased by the prophylactic use of antiretrovirals. In the iPrEx (Pre-Exposure Prophylaxis Initiative) trial, 2499 MSM from six countries received either a combination of two oral antiretroviral drugs, emtricitabine and tenofovir, or a placebo, resulting in a 44% reduction in HIV acquisition in the treated group [43]. In the CAPRISA004 (Centre for the AIDS Programme of Research in South Africa) trial, 445 South African women used an antiretroviral, tenofovir, as a topical microbicide before and after sex, resulting in 39% fewer infections among this group than among 444 women on a placebo [44]. An HIV vaccine trial published in 2009 was also able to demonstrate a modest efficacy of 31% [45].

1.4 Natural History of HIV

1.4.1 Discovery of HIV

Acquired Immunodeficiency Syndrome (AIDS) was first recognized in 1981, although the cause of the disease was unknown at the time. The AIDS-causing virus was first isolated in France, in 1983. The etiological agent was initially named lymphadenopathy-associated virus (LAV) in accordance with the studied patient's symptoms [46]. The retrovirus was subsequently re-isolated in the USA that same year, and a definite link between the virus and AIDS was established [47]. The virus was renamed Human T Lymphotrophic Virus III (HTLV-III) by the American group because it closely resembled the HTLV-I and II viruses studied in their laboratory. However, the newly isolated virus had serological and morphological properties distinct from the HLTV family members. In 1986 the appellation "HIV" became internationally accepted as the official nomenclature [48].

1.4.2 Origin of HIV

The lack of a close relative for HIV prompted the search for related lentiviruses in other species to better understand its origin. Simian immunodeficiency virus (SIV) was rapidly isolated from rhesus macaques [49], suggesting that HIV arose by zoonotic transfer from non-human primates. In 1986 a related virus, but with distinct antigenic components was isolated from West African patients with AIDS; this new virus was subsequently named HIV-2 [50]. Further investigations revealed different zoonotic origins of HIV-1 and HIV-2. HIV-1 is closely related to SIV in chimpanzees (SIVcpz), while HIV-2 more closely resembles SIV in sooty mangabeys (SIVsm) [51] (Figure 1.3). Although non-zoonotic theories on the origin of the virus have been put forward, such as the oral polio vaccine theory, they have been reliably discredited [52, 53].

Since the original isolation of SIV thirty-nine species of non-human primates have been shown to possess a strain of SIV [54]. However, in most natural hosts, the virus does not cause serious illness. This points to a long-term adaptive co-evolution in non-human primates which has resulted in low pathogenicity [54], in sharp contrast to the recent emergence and pathogenicity of HIV in humans. Although estimates of the date of HIV introduction into human populations have been as recent as the 1930s [52, 55], the discovery of ancient samples and the application of modern methods point to the beginning of the 20th century [56] as the most likely time of transmission. The current understanding is that the emergence of HIV in humans is the result of a series of independent cross-species transmission events from non-human primates to humans, some of which remained contained and others that led to the existing subtypes of HIV (section 1.4.3; Figure 1.3) [57, 58]. Most likely, these cross-species transmission events occurred in the Congo [59], as it is the country displaying the highest genetic diversity of HIV. While the route of transmission of HIV to humans is not known with certainty, in view of ongoing hunting for primate bush meat in central Africa, it is likely that transmission occurred though this practice. Exposure to infected blood may have occurred through animal bites or during butchering [58].

1.4.3 Sequence diversity

As more isolates of HIV were characterized and sequenced researchers became aware of its extraordinary genetic diversity. To manage this diversity HIV has been classified according to genetic similarities and sero-reactivity. HIV-1 contains three groups, M (main/most), O (outlier) and N (non-M, non-O), while HIV-2 separates into eight, A to H. HIV-1 group M is the pandemic-causing strain, and it

comprises a further 11 major subtypes (A1, A2, B, C, D, F1, F2, G, H, J, K) [60]. The HIV-1 groups M, N and O are each individually closer to SIV isolates than they are to each other and so are probably the results of different cross-species events (Figure 1.3).

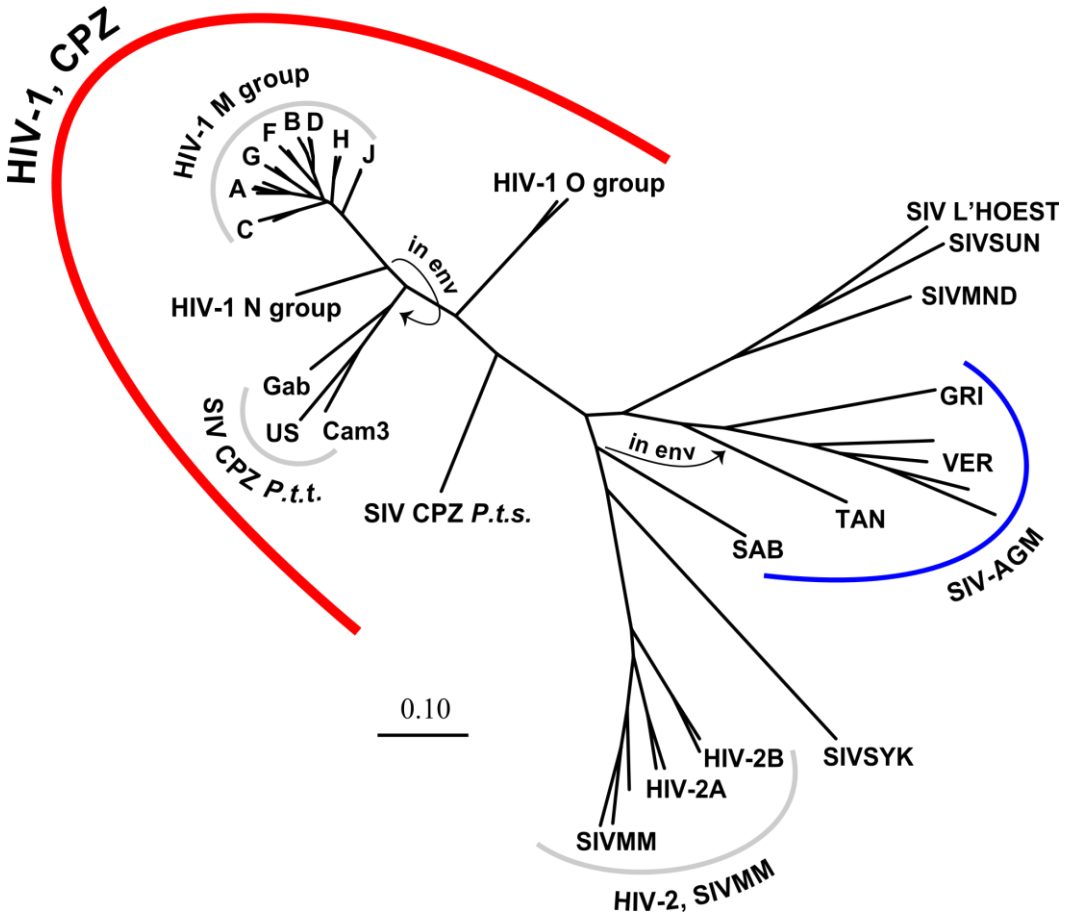


Figure 1.3: Primate lentivirus phylogenetic relationships based on the *pol* region. Scale in nucleotide substitutions per site. See text for details. By Kuiken *et al* [61], via Wikimedia.

HIV-1 group M subtypes are differentially distributed across the globe (Figure 1.4). Worldwide, subtype C is the most prevalent, accounting for 56% of infections, while subtype B predominates in Europe and North America accounting for 12% of infections globally [62].

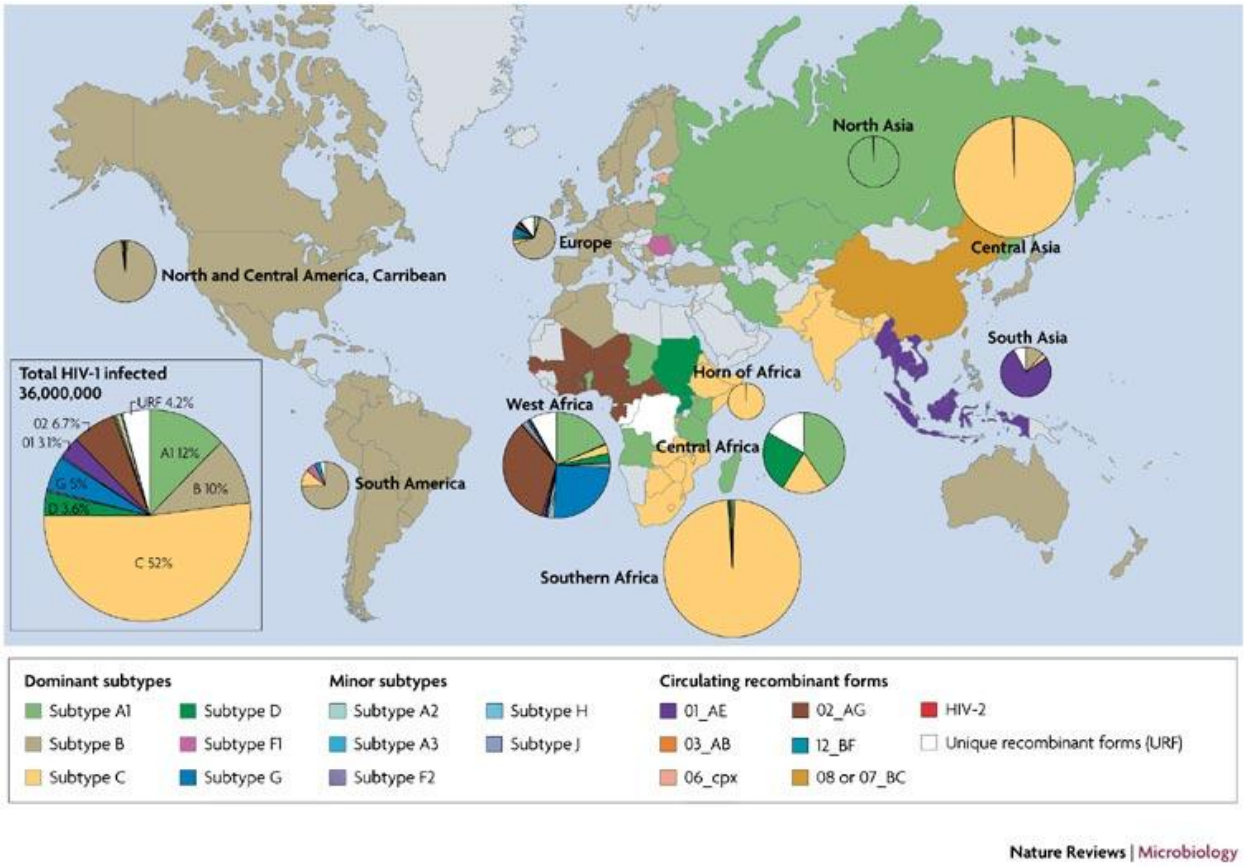


Figure 1.4: Worldwide HIV diversity. Countries are color-coded based on dominant HIV-1 subtype (data unavailable for countries in gray). The size of pie charts is proportional to the number of HIV infected individuals in each region. Reproduced from Ariën *et al* [63] with permission from Macmillan Publishers Ltd.

Subtypes are defined based on the high genetic distances found between them: 25-35%, as compared to 5-20% found within a subtype [64]. Their existence is believed to be due to a small number of original strains which expanded in number as a consequence of random founder effects [57, 58]. However, as the epidemic continues to spread and experimental sampling improves, subtype separation may become meaningless [26]. In particular, better sampling has uncovered numerous circulating recombinant forms (CRF) between previously sampled subtypes which in some geographical regions account for over 50% of infections [65]. CRF are officially recognized as independent strains once sequences have been isolated from three different patients with no epidemiological linkage.

Studies have further explored the pathological differences between the M group subtypes, suggesting variation in rates of MTCT [66] and of disease progression [67], but such results remain preliminary. The extensive spread of subtype C is more likely to be a consequence of its entering a population with high rates of partner exchange than because of increased transmissibility. However there may be a biological basis to the M group epidemic [26], as it appears to be more virulent and infective [68]. HIV-2 has a longer latent period and lower morbidity than HIV-1, with an epidemiologic distribution mainly confined to West Africa.

1.5 Molecular evolution of HIV

Due to the rapid replication and error-prone reverse-transcriptase of HIV, mutation rate is up to a million times higher than that of vertebrates [27]. Patients are usually infected with a single transmitted founder virus [69-71], which results in strong population bottlenecks. Subsequently, intra-patient viral genetic diversity increases over time [72], until HIV establishes itself as a genetically diverse population of viral genomes, called a quasispecies (Chapter 2). Although some mutations may accumulate due to genetic drift with no effect on viral fitness, others result from adaptation to the strong selective pressure exerted by the host immune response and are essential for viral success. Viral escape mutants that result in reduced recognition and destruction by the immune system are strongly selected [73-75]. Additionally, non-immune related selection is driven by ART which results in drug-resistant viral genomes. Together these forces drive HIV evolution both within hosts and at the population level.

1.5.1 Viral escape under host immune pressure

Two types of immune responses exert selective pressure on the virus. Antibodies target free-floating viruses by neutralizing epitopes on the surface of HIV; while the human leukocyte antigen (HLA) mediated cytotoxic T lymphocyte (CTL) response eliminates HIV-infected cells. Escape mutations for antibody responses are located in *env*, creating hypervariable regions where they accumulate incredibly rapidly. On average, one such adaptive mutation is generated every 2.5 months within a patient [76].

The CTL response is initiated when HLA cell-surface molecules bind and display short segments of HIV proteins (viral epitopes) to CD8+ T cells. There are over 5000 HLA class I alleles distributed across three loci, A, B and C, making them one of the most polymorphic alleles of the human genome. Different HLA proteins are able to display and bind a particular set of epitopes, as the pockets of HLA alleles will carry

different affinities for amino acid sequences. The replacement of amino acids within HIV epitopes can therefore avoid their being displayed by HLA proteins to the immune response by decreasing the efficiency of epitope binding. Other paths to immune escape include mutations that interfere with the CTL response either by reducing intracellular processing of epitopes or impairing recognition by T cells.

As any viral protein can be processed and displayed as an epitope, HLA-associated escape mutations are scattered across the entire genome [77]. Furthermore, HIV escape mutants are HLA allele-specific [78]. In large cohorts for which both HLA allele data and HIV sequences are available, links have been established between HLA alleles and the polymorphisms that they generate. HIV genome-wide maps of such escape mutants are now available [77]. As HLA alleles are differentially distributed across ethnic groups [79], ethnic-specific mutation patterns may be expected at HLA-associated sites in HIV. In Chapter 3, I specifically test this hypothesis in the ethnically diverse Canadian epidemic.

As a consequence of the founder effect of transmitted virus, HLA-associated polymorphisms can be transmitted. Within the new host there is evidence of either transmitted escape mutants reverting to wild-type or of mutation patterns becoming consistent with the new host HLA alleles [69]. With the potential of HLA-driven mutations becoming fixed in a population, frequencies of HLA-associated HIV polymorphisms have been shown to reflect the prevalence of HLA alleles in a population [80]. A recent study goes as far as suggesting that HLA-driven evolution may have been central to HIV subtype evolution [81].

Finally, HLA alleles are crucial to immune control of HIV and are the most important genetic determinant of disease progression identified so far [82, 83]. In particular some alleles such as HLA B57, offer a protective effect against HIV disease and are called ‘super-controller’ alleles [84].

1.5.2 Drug resistance

Since the 1990s an increasing number of antiretroviral drugs (ART) have become available for the treatment of HIV. However, because HIV is able to adapt to selective pressure at such an alarming rate, most drugs induce drug resistant mutations (DRM) which allow the virus to continue to replicate and increase the likelihood of treatment failure.

In developed countries, viral load testing is performed on patients who are being treated with ART to monitor their disease progression. Should viral loads increase, this is taken as evidence for treatment

failure. Most drugs available target protease (PR) and RT, so most DRM currently found occur in the *pol* region of HIV (Figure 1.2). To determine specific DRM, the HIV *pol* gene is sequenced, or *genotyped*. This allows for tailoring of drug regimens. Progression to second-line therapy may be more expensive.

Despite subtype C virus being the most prevalent globally, most HIV infections in the West are caused by subtype B (Figure 1.4). This has caused an imbalance in research efforts and most drugs have been developed against subtype B. Although no major differences have yet been observed in subtype C-infected patient responses to treatments developed for subtype B, some problems may arise due to drug resistance. Some DRM naturally occur more frequently in non-B subtypes [85]; furthermore, in the presence of drug selective pressure DRM are more rapidly acquired in non-B subtypes than in B [86].

1.5.3 Transmitted drug resistance

The use of HAART has led to the development of DRM in circulating strains. These DRM may be transmitted to drug naïve patients: this scenario is referred to as 'transmitted drug resistance' (TDR). TDR can considerably reduce the efficacy of first line therapies. Because of increasing prevalence of TDR [87, 88], and recent advances in our understanding of HIV drug resistance persistence [89, 90], *pol* genotyping is recommended for all newly diagnosed individuals and is routinely performed in a number of countries [91-93].

The WHO realized the importance of standardizing the list of mutations used to track TDR. The list facilitates comparisons between regional, national and international statistics and adjusts for non-drug related polymorphic DRM [94, 95]. These efforts have culminated with the publication of the Surveillance of Drug Resistance Mutations (SDRM) list [96]. Based on this list a software application called the Calibrated Population Resistance tool has been developed by Stanford University which uses an algorithm to determine whether sequences harbor mutations conferring drug resistance [97]. The tool has already been used in a number of WHO-sponsored studies [98, 99]. In Canada, TDR is tracked through the Strain and Drug Resistance Surveillance Program.

1.6 Availability of HIV sequences

1.6.1 The Strain and Drug Resistance surveillance program

The Strain and Drug Resistance surveillance program (SDR) was initiated in 1998 as an enhanced surveillance initiative to characterize the HIV epidemic in Canada [100]. It aims to (i) assess circulating

strains to monitor genetic diversity, (ii) evaluate TDR, (iii) assess patterns of transmission and (iv) ensure the safety of the blood supply. Through a collaboration between the National HIV and Retrovirology Laboratory (NHRL), the Surveillance and Risk Assessment Division and provincial health ministries the Public Health Agency of Canada (PHAC) publishes regular reports on the state of the HIV epidemic in Canada [101]. So far, seven provinces have taken part in the SDR program: British Columbia (BC), Alberta (AB), Saskatchewan (SK), Manitoba (MB), Nova Scotia (NS), Ontario (ON) and Newfoundland (NF).

Through the program, remnant diagnostic serum from newly diagnosed, treatment naive individuals is sent to the NHRL in Ottawa for analysis. Epidemiological data linked to each sample are also collected from each patient, including age, sex, ethnicity, year of positive test, city where the test was conducted, type of serological specimens collected and risk factors associated with infection (exposure category). Exposure categories are divided hierarchically (Appendix Figure 1) and patients are asked questions to determine their most likely route of transmission. Sequences and epidemiological data are irretrievably unlinked from patient identity before being sent to Ottawa, so that analyses at the federal level are non-nominal.

The National Laboratory for HIV Genetics (NLHG) generates *pol* genetic sequence (sequence position 2253-3549 in reference strain HXB2 [102]) using in house primers and assays (Appendix Figure 2). Sequence subtype is then determined using the Rega HIV-1 subtyping tool (http://dbpartners.stanford.edu/ RegaSubtyping/) and sequences are tested for the presence of DRM using the Stanford Calibrated Population Resistance Tool [97] (http://cpr.stanford.edu/cpr/).

In parallel, the National Laboratory for HIV References Services (NLHRS) determines timing of infection (< or > 155 days) using the Calypte BED-CEIA™ (capture enzyme immunoassay). The BED-CEIA measures the proportion of IgG antibodies which are HIV-specific in a sample [103], to detect the increased immune response to HIV in chronic infections. As explained previously, distinguishing between recent and chronic infections is essential for epidemiological analyses. Among these the estimation of the number of newly diagnosed infections which are recently acquired (HIV incidence) is of the upmost importance (Chapter 2).

1.6.2 Use of the *pol* region for HIV research

Population-based studies of evolutionary virology have been made possible by a growth in computational power and the increased availability of appropriate analytical tools but also by the

accumulation of HIV sequences. Indeed, HIV is one of the most sequenced organisms on earth [104]. In particular genotyping of HIV by sequencing the *pol* region for subtype determination and for the surveillance of TDR has resulted in an abundance of *pol* genetic information in databases.

Traditionally, genetic analyses of HIV have been carried out using the *gag* and *env* regions of the HIV genome [105-107] (Figure 1.2), while *pol* was criticized for being too genetically conserved [108]. However, the validity of *pol* for the phylogenetic reconstruction of transmission events (section 1.7.2.3) was demonstrated by Hue and colleagues [109], and this region of the genome is now extensively used in phylogenetic studies to address questions regarding HIV transmission dynamics [110-112]. *Pol* sequences have also been successfully used for the reconstruction of HIV epidemic history [113-116] (section 1.7.2, Chapter 4). Furthermore *pol* sequences contain extensive information on HLA-induced adaptation [77] (Chapter 3). Finally, within a patient, *pol* sequence diversity has been shown to gradually increase with time and to be a good measure of stage of infection [72] (Chapter 2).

Studies based on HIV genotype data can be carried out at little additional cost and have the added advantage of being non-biased (as opposed to cohort studies) and non-nominal (avoiding legal issues pertaining to transmission). As a consequence the SDR database has huge potential as a public health tool for research and epidemic management.

1.7 Recent advances in HIV molecular epidemiology

1.7.1 Methods in molecular evolution

While HIV evolves under strong selective pressure at the host level, its evolution is considered broadly neutral at the population level. Nevertheless HIV mutates so rapidly that it displays huge genetic variability across the population, making HIV sequences extraordinarily suitable for phylogenetic analysis and molecular epidemiology. In the case of HIV the accumulation of mutations is so rapid that it occurs on a similar time scale as ecological processes such as epidemic dynamics, transmission bottlenecks and immune pressure; thus, HIV evolution is *phylodynamic*, responding in regular ways during host infection and disease spread [117]. Consequently the reconstructed phylogenetic relationships between the genetic sequences of isolates can be used to infer the evolutionary and transmission history of the pathogen.

1.7.1.1 Phylogenetics

The evolutionary relationships between viral sequences can be visualized using a phylogenetic tree. Phylogenetic trees are reconstructed based on the differences found between sequences. However, the number of possible trees increases exponentially with the number of sequences, rendering the search for an optimal tree no trivial task. A number of methods exist to reconstruct trees, including maximum parsimony, distance-based methods, maximum likelihood, and Bayesian inference.

a) Maximum parsimony

Maximum parsimony is a character-based approach to phylogenetics, where the reconstructed tree is the one that requires the smallest number of evolutionary changes to explain the observed data [118].

However, the maximum parsimony approach is inconsistent, and is not guaranteed to produce the best tree as the number of sequences increases [119]. Moreover parsimony does not explicitly state which evolutionary model is used, and the premise that evolution is parsimonious is all the more incorrect as deeper phylogenetic depth is investigated. As a result, parsimony is particularly affected by a phylogenetic artifact named long-branch attraction, where highly divergent sequences appear more similar because multiple substitutions have hit the same site. Consequently, the number of mutations between them is dramatically underestimated. The maximum parsimony method is thus only appropriate for closely related sequences.

b) Distance-based methods

In order to construct a distance-based tree distances between each pair of sequences are calculated to construct a distance matrix. Distances are calculated based on a model of evolution which corrects for multiple substitutions and describes how frequently substitutions between nucleotides occur, based on both the frequency of those nucleotides, and the rates of change.

Once distances have been calculated, Neighbor-Joining [120], the most frequently used algorithm for tree reconstruction, joins the most closely related sequences in pairs according to the distance matrix to form subtrees, gradually forming the full tree. However information is lost by summarizing sequence information into distances and by looking at sequences only two at a time. Nevertheless, the method is fast and works well in practice [121].

c) Maximum likelihood

The maximum likelihood (ML) method is also based on a model of evolution. In addition, the ML method is character-based, using all the information contained in the sequences, and statistically robust and well-founded. The approach calculates the probability of the observed data (the sequence alignment), given the proposed model (tree topology, branch lengths and the parameters of the evolutionary model). For each site in the alignment, the probability of the observed nucleotides is calculated as the sum of probabilities of every possible reconstruction of ancestral states, given the substitution model. As sites in the sequence alignment are considered to evolve independently, the probability of the observed sequences is the product of the probabilities for each site. Under ML, the best tree is the one with maximizes the likelihood, or the probability of observing the data. Overall, this method produces accurate results [122] but can be computationally intensive.

d) Assessing the reliability of the tree

Bootstrap values are assigned to nodes evaluating our confidence in the observed branching patterns. They are calculated by resampling the data and rebuilding the tree. Specifically, columns in the original sequence alignment are re-sampled with replacement so that some columns may be represented more than once in the bootstrap replicates and others not at all. The phylogenetic tree is re-estimated from each bootstrap replicate, and the bootstrap value assigned to each node represents the frequency that each node is observed in the trees reconstructed from the bootstrap samples. For example if the data were bootstrapped 100 times and a group of sequences stayed together in 93 of the trees, the bootstrap value assigned to that grouping would be 93%. Bootstrapping can be used with MP, distance-based and ML tree reconstruction methods.

e) Bayesian inference

As previously explained, the ML approach calculates the likelihood of the data *D* (indicating the sequences), knowing the model and parameters θ (including tree topology, branch lengths, and parameters of the evolutionary model): p(*D*|θ). In contrast, Bayesian analysis calculates the posterior probability of parameters, given the data observed: p(θ|*D*). The two are related through Bayes' theorem equation:

$$p(\theta|D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)} \qquad (1)$$

where p(θ) is the prior probability distribution and p(*D*) is marginal probability of data [123].

ML searches for the tree that maximizes the probability of the data given θ, while the Bayesian approach searches for the distribution of trees that is most probable given the sequence alignment under a particular model of evolution. The Bayesian approach also makes use of likelihood, but incorporates prior knowledge, which enables the user to convey any expectations or uncertainty about some of the model parameters before looking at the data.

A difficulty exists however with the denominator p(*D*), as calculating the marginal probability of the data requires the complex integration of all parameters entering the model. This problem is solved by using a Markov chain Monte Carlo (MCMC) sampler [123, 124], a random walk algorithm that allows sampling directly from the target distribution, p(θ|*D*). The initial state is chosen either at random, or input by the user. The parameters are then modified by proposing a new state drawn from a proposal distribution. In the simplest case where proposal distributions are symmetrical, the probability of accepting the new state is calculated as a ratio of the posterior probabilities between the proposed and the current state of the chain. This ratio avoids calculating p(*D*), as it is common to both the proposed and current states. Therefore, the higher the probability of the new state is, the more likely its acceptance. However less probable proposed states can also be accepted. As a result the MCMC sampler samples from the target distribution, namely p(θ|*D*).

MCMC chains can nevertheless become entrapped in local optima. To check for this all chains must be run at least in duplicate and solutions validated only if replicate chains converge on the same distribution. Bayesian inference is computationally intensive and as a result running large number of sequences is a limiting factor (Chapter 4).

In the same spirit as the bootstrap values attributed to branching patterns in ML, clade posterior probabilities are estimated in Bayesian phylogenetic trees to measure confidence.

1.7.1.2 Phylogenetic dating under a strict molecular clock

The molecular clock is a technique used in molecular evolution to date events on phylogenetic trees. It was first observed, by Linus Pauling in 1962, that the number of differences in hemoglobin amino acid sequences sampled across species was proportional to the duration of time since the species diverged

[125]. The suggestion that mutations accumulate in sequences linearly with time, thus at a constant rate, constitutes the molecular clock hypothesis:

$$rate = \frac{genetic\ distance}{2 \times time} \quad (2)$$

In the case of species divergence, the clock is calibrated using fossil records to infer rate of evolution. In the case of viruses, which evolve much faster, sequentially sampled isolates have been used to calculate evolutionary rate [126]. Application of the molecular clock uses time-stamped sequences to date events in the evolutionary history of the virus which can be very useful to test different hypotheses, for example those concerning the origin of HIV [52].

1.7.1.3 Time-stamped Bayesian phylogenetics

a) The relaxed molecular clock

However, the strict molecular clock is rejected for most organisms, including HIV [127]. As suggested by equation (2), genetic distance, or branch length, is the product of the rate of molecular evolution and of the time duration of that branch. In the strict molecular clock, the rate is a constant, estimated by calibrating the clock with fossil information or time-stamped data, and branch lengths are directly proportional to time. In a relaxed clock model prior distributions are used to model rates of evolution on the one hand, and to model the speciation process on the other hand. In this way the relaxed molecular clock allows the rate of the clock to vary across the phylogenetic tree [128]. Earlier relaxed clock models were based on autocorrelated models of rate change, so that the rate of evolution at each branch was assumed to be inherited from the rate of the parental branch.

However because the topology of the tree is sometimes uncertain uncorrelated clocks dependent only upon the mean clock rate associated with the whole tree have been developed [129]. Uncorrelated clock rates are made to vary according to an underlying distribution, for example lognormal or exponential. Within a population, the prior placed on times is based on coalescence theory.

b) Coalescence theory

Coalescence theory is a retrospective model of population genetics used to trace the relationships between a set of sequences sampled from a population back to their most recent common ancestors (MRCA) [130]. Coalescent theory describes the distribution of the times at which each pair of sequences shares a common ancestor (or coalesces) on their phylogeny. Nevertheless the estimation of such a

distribution depends on the demographic model which describes the pattern of change of the viral population over time. Models of demographic change include constant population size, logistic growth, and exponential growth. More recently the Bayesian skyline plot (BSP) was introduced, which is more flexible [131].

c) Bayesian evolutionary analysis by sampling trees (BEAST)

These relaxed molecular clock models developed in a Bayesian framework are implemented in the program BEAST [132]. The program reconstructs the evolutionary history of time-stamped specimens, and provides non-parametric estimates of past population dynamics. In the case of a pathogen a number of parameters can be estimated from a subset of circulating viruses: the population size through time, the number of incident infections, the epidemic growth rate and doubling time, the time to the MRCA (or the date of introduction of the virus) and the evolutionary rate.

The best demographic model is selected in BEAST by means of the Bayes Factor (BF), which compares the marginal likelihoods (see equation (1)) of pairs of models (M):

$$BF = \frac{p(D \mid M_1)}{p(D \mid M_2)} \quad (3)$$

While the marginal likelihoods entering equation (1) can be estimated directly by the application of MCMC, their accurate estimation remains a problem. BEAST uses the harmonic means estimator, which, although widely used (and will be used for convenience in Chapter 4), is statistically unreliable [133].

1.7.1.4 Phylogeography

Demographic reconstructions can also been enhanced by the inclusion of geographic parameters into analyses [134]. This has made possible the analysis of viral spread with a spatial dimension, for example to understand the effect of migrations on the spread of HIV [65] or the connections between different national epidemics [113]. This field of study is called phylogeography.

1.7.1.5 Detecting Selection

Methods have been developed to determine whether the selective pressure driving the fixation of mutations in sequences is positive/adaptive (selecting for advantageous mutants) or negative/purifying (selecting against disadvantageous mutants). Because of the redundancy of the genetic code, nucleotide substitutions can lead to a change in amino acid sequence (non-synonymous change, N) or not

(synonymous change, S). At sites undergoing positive selection higher rates of N mutations are expected (ω = dN/dS > 1), while under negative selection, higher rates of S mutations are expected (ω = dN/dS < 1).

Several algorithms have been developed to estimate selective pressures. Counting methods compare the number of observed N and S to the maximum numbers possible [135], while ML fits a codon substitution model to the data that incorporates a parameter ω describing the proportion of N to S substitutions [136].

Nevertheless these methods suffer from several limitations: they are sensitive to the number of sequences used, and if S substitutions accumulate faster than N, it becomes harder to detective positive selection even if it exists. Finally, ω rate ratios make the false assumption that S mutations are selectively neutral.

1.7.2 Population histories of national and international HIV epidemics

BEAST has been used to investigate the epidemic histories of different HIV subtypes across geographical regions revealing subtype-specific and regional-specific patterns of growth, as well as to estimate dates of introduction into each region. However these types of analyses have not been broadly carried out for all geographical areas; therefore results presented here are limited to those currently available.

1.7.2.1 Demographic reconstruction of HIV

When studying HIV sequences from different regions or different subtypes using BEAST demographic models selected for the analysis can vary.

A worldwide analysis of subtypes B (most prevalent in the West) and C (most prevalent in sub-Saharan Africa) concluded that the growth of the former was logistic, and that of the latter, exponential [116]. Thus, while the subtype B epidemic initially exploded in the West (growing twice as fast as it had in Africa), it stabilized in the 1990s, whereas the subtype C epidemic in sub-Saharan Africa appears to be continuing to grow exponentially. In agreement a model of exponential growth was also the best fit for the Ethiopian subtype C epidemic [137]. In Eastern Europe however, an exponential model was selected for the subtype B epidemic [138]. In South America logistic models of growth have been selected for subtypes B, F [114] and C, as well as for the recombinant subtype CRF31_BC in Brazil [115] and the recombinant BF in Argentina [139].

By studying HIV samples within the same country it is possible to compare epidemic histories between different subtypes, or between CRF and non-CRF strains. In Albania such analysis revealed a growth rate five times higher for subtype A than for subtype B. In Brazil CRF31 appears to have spread at a faster rate than subtype C [115]. In the UK subtype B (the most prevalent) from MSM and non-B were analyzed separately [140, 141], under the assumption that most non-B sequences originated from heterosexuals. Transmission dynamics differed between the two groups with a smaller number of associated infections for each sequence in non-B, and more time between transmission events.

For the most part authors conclude that HIV epidemic growth patterns are likely a consequence of the underlying characteristics of the transmission networks. The most notable of these differences being that in sub-Saharan Africa, a large proportion of HIV transmission occurs though heterosexual sex and from mother to child, while in the rest of the world MSM and IDU are most affected. Within smaller populations and geographical areas other epidemiological factors are likely to affect the rate of spread, including times of introduction into different risk populations, sexual and drug behaviors, as well as laws, health policies and education relating to HIV. Nevertheless, some subtype-specific patterns have been observed which may be playing an additional role in the dynamics of these distinct epidemics.

1.7.2.2 Dating introductions of HIV across the globe

As well as being used to reconstruct past demographics of different subtypes as above and to date the zoonotic transmission of HIV from non-human primates to humans [53, 142], BEAST has also been used to date the introduction of HIV subtypes into different geographical regions and populations. While the strict molecular clock is consistently rejected for HIV sequences [65, 113, 115, 116, 137-139], dating is possible nonetheless using relaxed clock models. In the published literature dates of viral introductions into different parts of the worlds are consistent with our overall understanding of HIV spread across the world. Studies place the most recent common ancestor of subtypes B and C in the early 1960s, in Africa [116]. In the late 1960s, subtype B emerged in Haiti, and swiftly spread to the USA, then Europe [143]. Before the end of 1960s, subtype B was also circulating in South America [114], but was not introduced into Eastern Europe until the 1980s [138]. Other strains were not identified outside of Africa until later: F1 did not spread to South America until the late 1970s [114], although it had started diversifying in the Democratic Republic of Congo by the 1950s [113]; AE, now responsible for the epidemic in Thailand, also appeared there in the late 1970s [144], although the African MRCA dates back to the 1960s. CRF_BF seems to have appeared in Brazil only in the early 1990s [139].

Analyses of HIV epidemic history have a number of benefits. They allow for epidemics to be monitored for example to help guide intervention strategies, and it has even been suggested they may engender predictions about the future of the epidemic [145]. Finally demographic analyses have been applauded for their ability to give unbiased views of epidemics, in contrast to collecting epidemiological data. Similarly they may reveal important information in countries where epidemiological data are lacking. For example in Albania analysis of the HIV epidemic based on only 52 sequences suggested that the number of circulating infections was twice as high as that reported [138]. In Chapter 4, I use a similar approach to compare the epidemic history of HIV across Canadian provinces.

1.7.2.3 Transmission networks and Public Health

On a local scale phylogenetics is used to address questions related to transmission networks and events. Indeed while two sequences taken from the population are on average 10% different within a subtype [146], or 5% within a city [147], more closely related sequences can be deemed to represent epidemiologically linked infections, associated with the same transmission chain. These constitute transmission clusters and are defined in the literature based on low genetic distance and high bootstrap values.

a) Transmission patterns

Increasingly phylogenetics is being used as a public health tool for the analysis of transmission networks. Due to its intimate modes of transmission (sex and intravenous drug use) and the time-lapse between infection and diagnosis, HIV contact network data using traditional epidemiological tools are difficult to obtain and often unreliable [148]. Meanwhile as sequencing coverage increases and analyses become more fine-tuned, the evolutionary relationships between viral sequences reflect the transmission dynamics of the local population. Thus phylogenetic analyses can be used to supplement traditional epidemiological tools in order to elucidate characteristics of transmission networks.

In this way phylogenetics has been used to identify correlates of transmission including risk group [140], stage of infection [110, 112], cluster size [149], the presence or absence of co-infections, including other STI [150] as well as drug treatment and compliance. A recent study tried to determine the relative contribution of each of these variables to the risk of onward transmission [32], concluding for the first time that ART truly did decrease HIV transmission risk. A high resolution analysis in the UK investigated the pattern and timing of HIV transmissions within MSM transmission clusters, revealing that after the

first infection, other members became rapidly infected [141]. In San Diego, a phylogenetic approach was used to assist public health directly by identifying clustered individuals with high rates of onward transmission in order to target prevention intervention towards them [147].

It is generally agreed that a thorough understanding of transmission networks is essential for the proper targeting of prevention efforts [151], and may even allow for outbreak management, as implemented in San Diego. However although molecular epidemiological approaches of this type have been successfully used for epidemic management of infections such as syphilis [152] and tuberculosis [153, 154], severe limitations remain concerning establishing direct HIV transmission links based on molecular data [155], which will be addressed in the following section.

b) Transmission links

The first application of phylogenetic analyses in HIV was in fact in forensics, to confirm transmission pairs in contact investigations, the first of which established transmission of HIV from a dentist in Florida to his patients [106]. More recently in the news, phylogenetic analyses were used in a criminal trial to exonerate six foreign medical workers accused of transmitting HIV to over 400 children in a hospital in Libya [156].

However many researchers have insisted that phylogenetics alone cannot be used to prove a transmission link [157]. Indeed even if two isolates are very similar to each other, a missing partner linking the two could be involved, or both subjects may be part of a wider transmission network. Moreover the recipient could have been infected previously, and then re-infected [155]. Until recently, researchers were also adamant that directionality of transmission could not be inferred using phylogenetics [157]. However, a recent publication has claimed the contrary [158]; the authors assert that they can differentiate between transmitters and recipients based on tree structure. Nevertheless the article was based only on two cases and each time more than one recipient had become infected, facilitating the inference of directionality as compared to a transmission pair. Furthermore the authors admit that sequence divergence over time will decrease the efficacy of this approach if there is a delay between the transmission event and sampling. Finally, because of liabilities associated with transmitting HIV in a number of countries, caution should be used when establishing transmission links at the individual level.

1.8 Objectives of the thesis

As discussed in section 1.6, the SDR program has generated an abundance of HIV *pol* sequences that can be used to elucidate transmission networks, to reconstruct HIV epidemic history and to study HIV evolution at the individual and population level. In this thesis, I address three particularly pressing issues relating to HIV epidemiology in Canada.

a) The ability to distinguish between recent and established infections is of the upmost importance to Public Health, for the estimation of incidence rates, and hence for the evaluation of HIV prevention programs. Currently, PHAC uses the BED assay, but this test tends to over-classify infections as recent and hence inflate incidence rates. In Chapter 2, I validate and extend a mixed base classifier (MBC) to resolve infections as < or > 155 days based on the proportion of mixed bases in *pol* sequences generated during population-based Sanger sequencing.
b) As explained in section 1.5.1, a large part of HIV adaptation occurs in response to HLA immune pressure exerted by the host population. Understanding and potentially predicting evolution within HIV epitopes has important applications for vaccine design. The Canadian epidemic is ethnically diverse, presenting a complex genetic background for HIV evolution. In Chapter 3, I investigate the adaptive evolution of *pol* at HLA-associated sites in different ethnic groups in Canada.
c) Finally in Chapter 4, I draw conclusions on the applications of evolutionary analyses of HIV. I further present preliminary results on the population history of the Canadian HIV epidemic across different provinces.

Improvements in methods for the identification of recent infections will make it possible to provide relevant interventions to recently infected patients in order to reduce their chances of transmitting the virus. In addition, more accurate estimates of HIV incidence will help for the evaluation of prevention campaigns. Further characterizing the Canadian epidemic by investigating the different epidemic histories across provinces will help identify strategies that have already been successful and better target prevention efforts. Finally, an understanding of viral evolution at the population level will add to a body of work essential for the development of a vaccine against HIV.

Chapter 2

Genetic Diversity as a Marker for Timing Infection in HIV Patients: Validation and Comparison with Serological Methods

2.1 Abstract

It has been reported that the increase in HIV sequence diversity within a host, as measured by the number of mixed base calls in a *pol*-based genotype, may be used to determine stage of HIV infection. Here, we describe the optimization of a mixed base classifier (MBC) that was employed to distinguish infections < or > 6 months. We show that for best results, the area of the secondary peaks on the sequencing chromatogram should account for 15% or more of the area of primary peaks. Using a cutoff of 4.5 mixed bases per 1000 nucleotides (0.45%) in the *pol* region best distinguished between recent and established infections. We further evaluated the MBC for its performance in classification by enumerating mixed bases exclusively at sites that reflect evolutionary pressures: HLA selection sites, sites that increased in entropy over the course of infection, and position within a codon. An entropy-

based approach most significantly improved the MBC. The optimized classifier performed better in an extended dataset of specimens classified as recent or established by BED than the classifier based on full sequences. In conclusion, infections can accurately be classified as < or > than 6 months using the optimized MBC.

2.2 Contributions

This work was presented at the 18th Conference on Retroviruses and Opportunistic Infections in February 2011, for which I received a Young Investigator Award. It also received a prize at the 5th Public Health Agency of Canada Science and Research forum in March 2011.

Isabelle Joanisse, Harriet Merks, Dominic Vallée and Kyna Caminiti amplified and sequenced all the samples from the *NHRL training dataset* and *BED dataset*. Michael Rekart and Darryl Cook (BC Centre for Disease Control) identified the acute infections and shared specimens. John Kim and Laurie Malloch carried out all the BED assays. Additional sequences were obtained from Drs. Bluma Brenner (Jewish General Hospital, Montreal), Frank Maldarelli (National Institute of Health, Bethesda USA) and Richard Harrigan (BC Centre for Excellence, Vancouver). The initial idea to classify infections as recent or established based on the proportion of mixed bases was James Brooks'. Improvements to the method (setting the threshold, testing different categories of sites) were my suggestions and I conducted all the analyses. JB and Stéphane Aris-Brosou offered their guidance and support throughout the project. Xuhua Xia implemented the mixed base counter in DAMBE. I wrote the first draft of the manuscript and JB and SAB both revised and edited it. All analyses on the SDR database were conducted at PHAC.

2.3 Introduction

The early symptoms of HIV infection are non-specific, variable and rarely severe enough for patients to seek medical attention. Consequently, the majority of HIV infections are diagnosed among people who are chronically infected [3] with the true date of seroconversion remaining unknown. Nevertheless, correct identification of recent HIV infections (RHI), *i.e.* less than six months, is critical to public health for at least three reasons. First, reliable estimates of RHI are required to assess the population-based rate of HIV infection within a period of time (HIV incidence rate). This information is required to determine trends in the HIV epidemic as well as to evaluate the success of prevention strategies.

Second, contact tracing of RHI patients is critical as it is believed that high rates of onward HIV transmission are associated with the high viral loads during early infection [19]. Third, as viral strains in RHI patients are closely related to the virus population found in the transmitting partner, identification of RHI offers unique opportunities to improve our understanding of the biology of HIV transmission [159]. However, current methods used to tease apart recent from established infections have severe shortcomings [160].

Incident infections may be identified during the course of routine diagnostic testing. At the simplest level, discordant serology (negative test followed by a positive test) over a brief and defined time period accurately identifies a recently infected person. Alternatively, during the first 2-4 weeks of infection and prior to seroconversion, the detection of viral capsid (p24) and genetic material in a serology negative diagnostic specimen may be taken as reliable evidence of an acute infection [161, 162]. However, these methods will only identify incident infections in the minority of patients who are either frequently tested or present to health care during acute infection and are provided the opportunity for non-serologic testing.

Beyond the 2-4 week window, HIV-specific immunoglobulin G (IgG) appears as a consequence of the host adaptive immune response. Following this seroconversion event, the proportion of a patient's total IgG that is HIV-specific increases over the course of infection. As a result, the increase in HIV-specific antibody, relative to total antibody, can be used to follow the course of the infection within a patient [103], as implemented in the Calypte BED-CEIA™ test (capture enzyme immunoassay; originally developed for subtypes B, E and D) – or BED for short hereafter. Initially designed to classify infections as < or > 6 months (183 days), calibration using samples from patients with known duration of infection determined that the best cutoff for distinguishing `recent' from `established' infections was 155 days. Since its development, BED has been used in HIV surveillance programs to estimate the prevalence of recent infections and therefore, to assess HIV incidence. However, BED has a number of limitations. First, HIV-specific antibodies can drop during the course of an infection, either as patients progress to AIDS, or when viral load is naturally suppressed [163]. As a result of this decrease, BED misclassifies such infections as recent [164]. Although a corrective algorithm can be applied, incidence estimates may not always be accurate depending on the overall number of tests performed [165]. Second, the recommended 155-day cutoff varies dramatically between subtypes: from 155 for subtype B to 360 days

for subtype C [160, 166]. Thus, the application of BED to a population with a multitude of viral subtypes can be challenging.

An alternative approach to determining stage of infection has been to measure the genetic diversity of HIV within an individual [72]. Previous work showed that sexually transmitted HIV infection is established by only a limited number of viruses [70, 71], with viral genetic diversity increasing throughout the HIV infection [167, 168]. As a proxy for within-patient genetic diversity, the proportion of "mixtures" during population-based sequencing can be used to approximate diversity, reflecting polymorphisms within the viral population. In their pioneering work, Kouyos *et al*. determined that a mixed base cutoff of $\boldsymbol{\gamma}_{360}$ = 0.5% achieved a sensitivity of 86.8% and a specificity of 70% in predicting whether patients had been infected for more or less than a year [72]. Here, given the existing body of work utilizing BED-based incidence calculations, we wished to examine whether the MBC would provide a concordant classification of early and late infections using the 155-day BED cut-off. In addition, as the number of mixed bases in a nucleotide sequence implicitly depends on the threshold $\boldsymbol{\tau}$ for calling mixed bases, we varied this threshold based on the sequencing chromatograms to improve the accuracy of the classifier. Finally, because selective pressure on the *pol* gene is not evenly distributed, we evaluated whether mixed bases in specific subsections of *pol* would provide greater resolution of recent versus established infections.

2.4 Materials and Methods

2.4.1 Study populations

The Canadian HIV Strain and Drug Resistance Surveillance Program (SDR) monitors subtype and transmitted drug resistance among newly HIV-diagnosed antiretroviral-naive patients in Canada [169]. Remnant HIV diagnostic serum specimens are sent to the National HIV and Retrovirology Laboratories (NHRL) in Ottawa for genotyping. In the present analysis, 1450 samples received from Western Canada between 2002 and 2008 were analyzed. The SDR program is approved by the Health Canada research ethics board.

Additional external sequences originating from patients with known duration of infection were obtained from published work [77] or were kindly provided by other laboratories. All additional sequences were of subtype B.

2.4.2 Stage of infection

Three datasets were assembled, two with sequences of known infection period and one with sequences of unknown duration of infection. In the *NHRL training dataset* of 96 specimens, stage of infection was resolved as follows: 66 specimens were diagnosed as recent either because viral p24 antigen was detected in the sample in the absence of antibodies, or because the patient tested negative within the 155 days prior to diagnosis. Infections were classified as established for the remaining 30 specimens who tested positive with a second HIV test performed at least 155 days after initial HIV diagnosis.

An additional 237 sequences from other laboratories, or from the literature [77], that were clearly identified as originating from infections < or >155 days in duration were added to the *NHRL training dataset* to form the *full training dataset*. The aggregate, *full training dataset* contained 333 sequences in total, composed of 162 recent and 171 established infections.

NHRL sequences for which duration of infection was unknown were classified according to BED results as < or > 155 days (recent vs. established, respectively) [103]; this third dataset is denoted hereafter as the *BED dataset*.

2.4.3 Amplification and sequencing

Viral RNA was extracted from samples, and the *pol* region was reverse transcribed, amplified by two rounds of PCR and Sanger sequenced as previously described (Appendix 2) [149]. Contigs were assembled, edited, aligned to the NCBI HIV-1 reference genome (accession number: NC_001802) and trimmed using BioEdit Software, version 7. 0. 9 [170]. *Pol* sequences obtained covered 1305 bp, including the protease gene (PR, bases 1-297) and a portion of reverse transcriptase (RT, bases 1-1008). Sequences from samples collected before 2005 have previously been made available in GenBank [149] (accession numbers: HM468499-HM469374). All new sequences were deposited in GenBank (accession numbers: xxxxxx-xxxxxx). Subtype determination was based upon submission of the *pol* fragment to the REGA HIV-1 Subtyping Tool, version 2. 0 (http://dbpartners.stanford.edu/RegaSubtyping/) [171]. As subtype B predominates in the Canadian epidemic, we limited our dataset to this subtype in the present analysis.

2.4.4 Mixed base calls

Mixed bases were identified on a sequencing chromatogram at nucleotide positions where a second trace, representing a different base, was present above a threshold τ percent area of the dominant,

primary peak. For example, if a threshold τ = 15% is chosen, the secondary peak must represent at least 15% of the area of the primary peak in order for a mixed base to be called. Notation of mixed base calls follows the International Union of Pure and Applied Chemistry guidelines (R, Y, K, M, S, W, N). The threshold τ for calling mixed bases in SeqScape [172] was varied from 5% to 45% for all samples sequenced in our laboratory, creating a new sequence alignment each time. For each sequence, the number of mixed bases was then counted using DAMBE [173].

2.4.5 Receiver Operator Characteristic (ROC) curve analysis

ROC curve analysis was used to assess whether sequences could be classified as recent or established based on the number of mixed bases called during Sanger sequencing. We name this method the Mixed Base Classifier (MBC). Under the hypothesis that the number of mixed bases in sequences increases with duration of infection, we first determined with the *NHRL training dataset* the threshold τ that maximizes the area under the curve (AUC) in the ROC analysis. We then used the *full training dataset* to search for the mixed base cutoff γ_{155} that optimized classification of recent versus long-term infections at 155 days.

2.4.6 Sequence subsets

Multiple derivative datasets were created from the *full training dataset*, to contain only sites considered informative for evaluation with the MBC. Sites that were more variable in established than in recent infections were identified using two methods. First, the proportion of mixed bases at each site was calculated for recent and established infections separately. By calculating the difference at each site, those where the proportion of mixed bases increased in established infections were identified (True+). Second, an entropy-based approach was applied. Site-specific Shannon entropy in a sequence alignment measures the variability at each position. Site-specific difference in entropy between recent and established infections at the nucleotide and amino acid level was calculated and randomized using the Shannon Entropy-Two submission tool on the Los Alamos website (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy.html). Four alignments were created using entropy-based measures, containing sites that increased in entropy: (i) in nucleotide sequences (ΔE+Nuc), (ii) in amino acid sequences (ΔE+AA); and (iii) significantly after randomization (α = 0.05) in nucleotide (ΔE+Nuc.Sig) and amino acid sequences (ΔE+AA.Sig).

Human Leukocyte Antigen (HLA)-associated sites previously described [77] were mapped along the *pol* sequences (HLA+) and sites undergoing positive selection (Pos) were inferred using HyPhy under the

GTR substitution model and the SLAC method [174]. A significance level of 0.1 was chosen in order to increase the number of positively-selected sites. Based on this analysis, three additional alignments, HLA+, Pos and both together (HLA+Pos) were generated. Finally, three datasets were generated, for each codon position (CP1, CP2 and CP3). In total, 11 subdatasets containing select categories of sites were created for comparison against the full sequence. For each sequence in the subdatasets, we calculated the proportion of mixed bases. Using ROC analysis, we identified a mixed base cutoff $\boldsymbol{\gamma}_{155}$ to separate recent from established infections. Classification performance was compared between categories of sites. Finally, we evaluated the optimized MBC for its ability to distinguish infections classified as recent or established by the BED in the *BED dataset*.

All statistical analyses were carried out in SPSS [175].

2.5 Results

2.5.1 Composition of the datasets

Datasets were analyzed differently according to the information available for each sequence. First, sequencing chromatograms were used with the *NHRL training dataset* to identify the threshold $\boldsymbol{\tau}$ for calling mixed bases which most improved classification performance of the MBC. Then, the *full training dataset* was employed to identify the cutoff $\boldsymbol{\gamma}_{155}$ that best distinguished recent from established infections. The full training dataset was further employed to generate 11 derivative datasets (Figure 2.1, Table 2.1), each to be evaluated with the MBC.

2.5.2 A mixed base threshold τ of 15% most improves accuracy of the mixed base classifier

We wanted to test whether the MBC could classify infections as < or > 155 days as accurately as infections < or > 1 year [72]. Knowing that the number of mixed bases in a sequence implicitly depends on the threshold $\boldsymbol{\tau}$ chosen for calling for mixed bases, we first examined the effect of varying this threshold $\boldsymbol{\tau}$ on the performance of the MBC. Sequencing chromatograms were reinterpreted from the *NHRL training dataset* by adjusting the base caller threshold $\boldsymbol{\tau}$ to identify mixed bases at 5%, 15%, 25%, 35% and 45% of the area of the dominant peak. The agreement between the proportion of mixed bases and stage of infection was evaluated using ROC curve analysis, which plots the true positive rate of the new test (sensitivity) against its true negative rate (1-specificity). The threshold $\boldsymbol{\tau}$ that maximizes the

area under the curve (AUC) represents the best classifier setting. The AUC for different thresholds **τ** ranged from 0.702 (95% CI: 0.597-0.806) to a maximum of 0.802 (95% CI: 0.706-0.897**,** Figure 2.2) for a threshold **τ** of 15%, i.e. secondary peaks accounting for at least 15% of the dominant peak should be called as mixed bases. The MBC **τ** =15% yielded a sensitivity and specificity of 78.8% and 80%, respectively, in distinguishing recent from established infections. At this threshold **τ** =15%, only a small proportion of sequences (14.6%) continued to be misclassified.

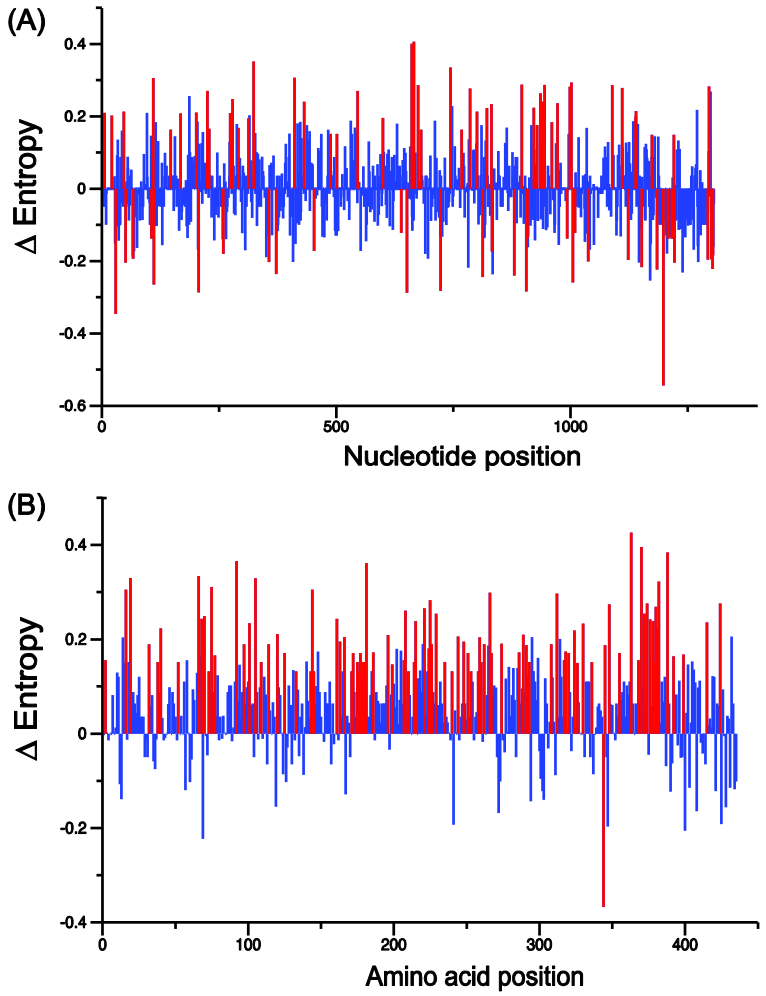


Figure 2.1: Site-specific differences in entropy (ΔEntropy) were calculated between recent and established infections in the *full training dataset*, both in nucleotide (A) and in amino acid (B) sequences. Site-specific ΔEntropy are color-coded: red represents significant differences; blue represents non-significant differences.

Table 2.1: Alignment sizes s for the categories of sites evaluated with the MBC at the amino acid (AA) and nucleotide (nuc) levels

Alignment name	Full Sequence	True Increase	ΔE+ Nuc	ΔE+ Nuc. Sig	ΔE+ AA	ΔE+ AA. Sig	HLA+	Pos	HLA+ Pos
s_{AA}	435				311	104	78	26	85
s_{nuc}	1305	551	356	50	933	312	234	78	255

NOTES–Four entropy-based measures are listed: sites which increased in entropy in nucleotide (ΔE+Nuc) and amino acid sequences (ΔE+AA), and sites which significantly increased in entropy in nucleotide (ΔE+Nuc.Sig) and amino acid sequences (ΔE+AA.Sig). Sites previously shown to be HLA-associated (HLA+); sites inferred under positive selection (Pos); and a combination of both (HLA+Pos+) are also shown.

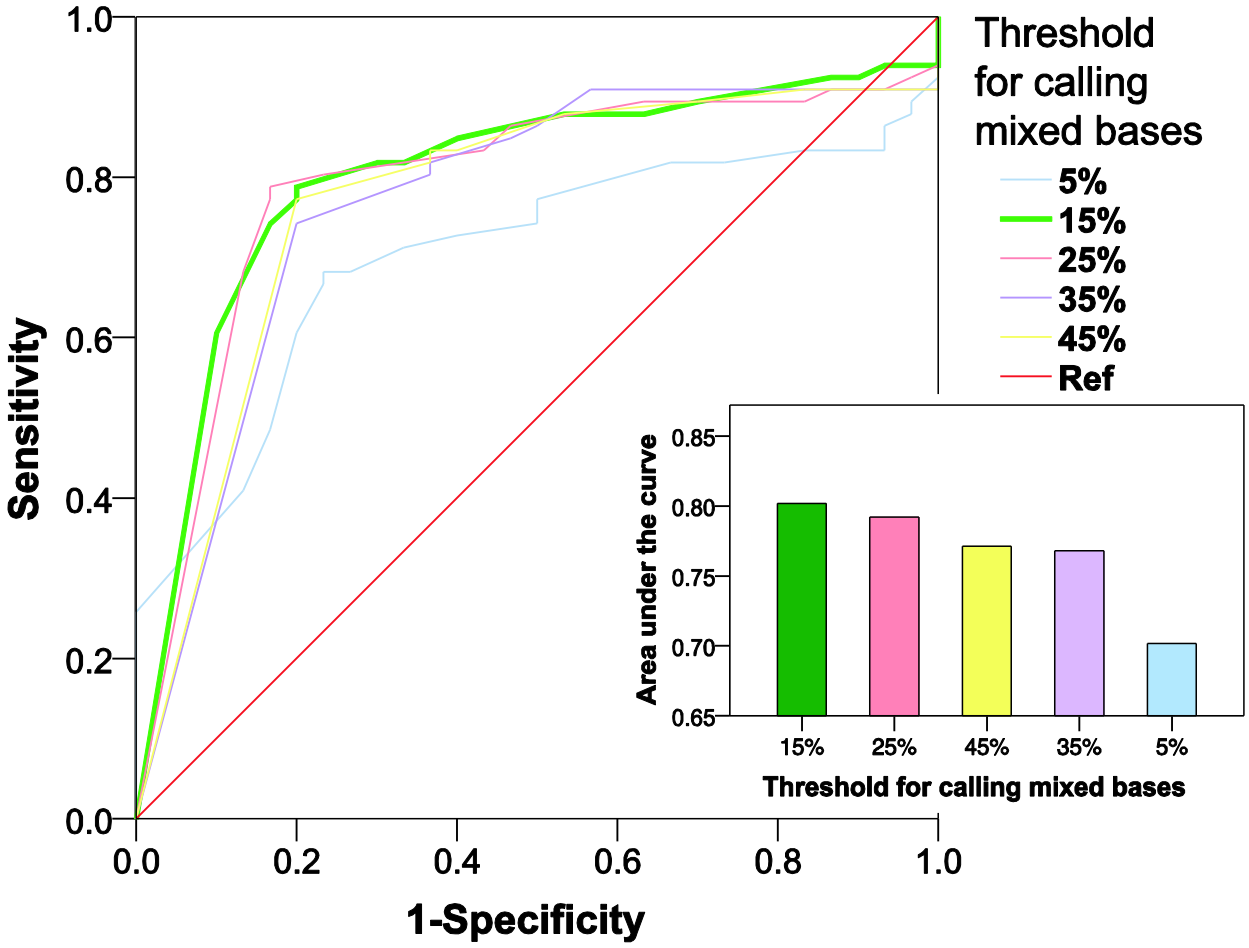


Figure 2.2: A mixed base threshold *τ* of 15% most improves prediction of recent infections. The threshold ***τ*** for calling mixed bases was varied from 5 to 45% in SeqScape to generate a new alignment each time. For each threshold ***τ***, the mixed base classifier (MBC) was evaluated through ROC analysis.

2.5.3 An entropy-based approach can increase sensitivity and specificity of the MBC

With this optimal threshold τ =15%, we explored whether the sensitivity and specificity of the MBC could still be improved. We used the *full training dataset* (*n*=333) to identify which sites were most important for predicting whether infections were recent or established by comparing the performance of the categories of sites in 11 derivative datasets (Table 2.1). The threshold τ could not be varied for all sequences in the *full training dataset*, because chromatograms were not available, but τ was known to have been set between 15% and 20%. When full sequences were analyzed, the frequency distribution of mixed bases differed significantly between recent sequences and established sequences (0.22% and 1.28%, on average, respectively; Kolmogorov-Smirnov test, $p < 0.001$). ROC analysis indicated that γ_{155} = 0.45% was the optimal cutoff, such that sequences would be classified as recent if they contained fewer than 0.45% mixed bases, and as established otherwise. At this γ_{155} cutoff, the MBC achieved a sensitivity of 77.8% and a specificity of 81.9% (AUC = 0.878) in distinguishing recent from established infections. Performance of the MBC was further increased in four of the derivate datasets tested: True+, ΔE+ AA and ΔE+ AA.Sig and ΔE+Nuc (Table 2.2). Among these, the ΔE+Nuc sites, which included only 25% of the sequenced nucleotide positions, provided the best improvement in the ability of the MBC to correctly classify recent from established infections. Concentrating on those nucleotide positions alone, we were able to increase the sensitivity to 85.2% and specificity to 83.5% (AUC = 0.906), at a mixed base cutoff γ_{155} = 0.82%. While the negative predictive value was increased from 0.82 to 0.86, the positive predictive value increased from 0.80 to 0.83 as compared to using full sequences.

2.5.4 HLA-associated sites and sites under positive selection are not sufficient to predict recent infections

HIV is known to diversify under intense selective pressure early in infection from both the innate and adaptive immune responses [75, 176]. We therefore investigated whether focusing on HLA-associated sites, or on other sites inferred to be under positive selection, increased the predictive power of the MBC. Neither category of sites alone improved the sensitivity and specificity of the MBC as compared to the full sequence (Table 2.2). Nevertheless, sites under positive selection (Pos) yielded a surprisingly high AUC of 0.802, given that these sites covered only 6.4% of the alignment (28/435 amino acid sites). A combination of HLA+ and Pos sites increased performance beyond either category of sites alone (AUC =0.858), but not beyond the MBC using full sequences. Although both HLA-associated and positively

selected sites associated weakly with sites increasing in entropy (Fisher's exact test, $p = 0.013$ and $p = 0.037$, respectively), these categories of sites were insufficient for optimal MBC performance.

Table 2.2: MBC performance of each category of sites in the *full training dataset.*

Alignment name	Full Sequence	True Increase*	ΔE+ Nuc*	ΔE+ Nuc. Sig	ΔE+ AA.*	ΔE+ AA. Sig*	HLA+	Pos	HLA+ Pos+
Best cutoff γ_{155} (%)	0.45	0.90	0.82	1.00	0.50	0.54	0.71	0. 64	1.10
AUC	0.878	0.899	0.906	0.843	0.890	0.899	0.849	0.802	0.858
Sensitivity	0.827	0.846	0.852	0.870	0.852	0.889	0.796	0.809	0.870
Specificity	0.788	0.747	0.835	0.759	0.776	0.782	0.788	0.747	0.712

Notes–Four entropy-based measures are listed: sites which increased in entropy in nucleotide (ΔE+Nuc) and amino acid sequences (ΔE+AA), and sites which significantly increased in entropy in nucleotide (ΔE+Nuc.Sig) and amino acid sequences (ΔE+AA.Sig). Sites previously shown to be HLA-associated (HLA+); sites inferred under positive selection (Pos); and a combination of both (HLA+Pos+) are also shown. * indicate categories of sites which outperformed the full sequence.

2.5.5 All three codon positions are informative for distinguishing recent from established infections

About 70% of the mutations at the third codon position are synonymous (do not lead to changes in the amino acid chain) while most mutations in first and all mutations at second positions are non-synonymous. We examined whether measuring mixed bases at each codon position, reflecting either synonymous or non-synonymous diversification, improved the discriminatory power of the MBC. Overall, restriction of the analysis to individual codon positions did not improve MBC performance (Figure 2.3). The best performance was achieved with 3rd codon positions, yielding an AUC of 0.872, but there was little difference between codon positions. It appears that mixed bases at all three positions contribute to diversification with stage of infection suggesting that mixed bases resulting from both synonymous and non-synonymous mutations can contribute information for timing of infection classification using the MBC.

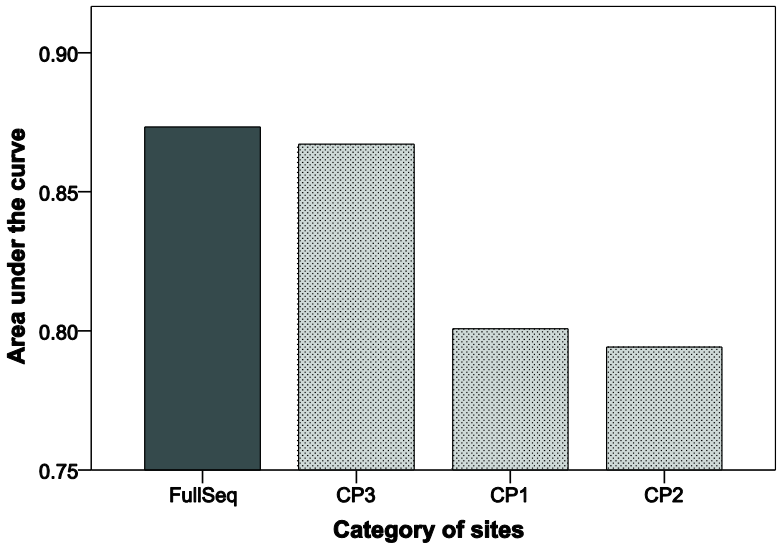


Figure 2.3: Mixed base classifier (MBC) AUC-performance at individual codon positions. In gray, based on the *full training dataset*, hatched gray, separate analysis of codon positions 1, 2 and 3 (CP1, CP2, CP3).

2.5.6 Evaluation of the concordance between MBC and BED results

Finally, the MBC was compared to BED using the *BED dataset*. 1354 serum specimens were classified by BED as 440 recent and 914 established. At a 15% threshold τ for calling mixed bases, the average proportion of mixed bases was 0.852% per site for established infections and 0.346% for recent infections. Mixed base frequency distribution differed significantly between the two groups (Kolmogorov-Smirnov test, $p < 0.001$). Using full sequences and $\gamma_{155} = 0.45\%$, the MBC and BED agreed on the classification of 909/1354 samples (67.13%). Of 445 discordant samples, three quarters were 'established' by BED but 'recent' according to the MBC. When the ΔE+Nuc sites determined in the *full training dataset* and its $\gamma_{155} = 0.82\%$ cutoff was used, agreement between the BED assay and the MBC increased to 68.98% (Figure 2.4).While using all sites or the ΔE+Nuc sites fared equally well in classifying BED recent samples, the ΔE+Nuc was better able to exclude BED established samples from being classified as recent. Therefore the frequencies of mixed bases at the ΔE+Nuc sites were higher in BED established sequences. Nevertheless, the highest proportion of discordant results was found among specimens classified as established by BED but recent by the MBC.

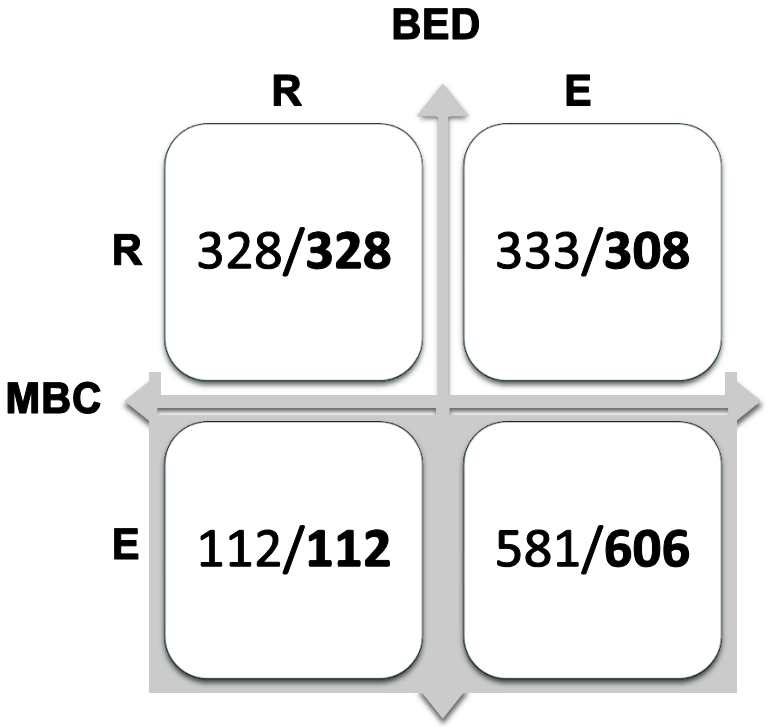


Figure 2.4: Performance of the mixed base classifier (MBC) in the BED dataset. Of 1354 samples, BED testing classified 440 as recent (R), and 914 as established (E). Using full sequences and the γ_{155} = 0.45% cutoff, the MBC correctly classified 328 of the recent sequences and 581 of the established sequences (numbers in plain). When the ΔE+Nuc sites determined in the full training dataset and its γ_{155} = 0.82% cutoff was applied, the MBC continued to correctly classify 328 of the recent sequences, but the number of correctly classified established sequences increased to 606 (numbers in bold).

2.6 Discussion

In this study, we found that among a precisely timed cohort, stage of HIV infection can reliably be inferred from the number of mixed bases identified during population-based sequencing of the *pol* region, consistent with the work of others [72, 177-179]. In addition, we were able to improve the resolution of the MBC and to reliably resolve infections < or > than 155 days, as opposed to the 1-year threshold used by Kouyos *et al.* [72]. Indeed, at a mixed base γ_{155} = 0.45% cutoff, the MBC was able to classify infections as <155 days with a sensitivity of 81.5% and a specificity of 78.9%. This re-calibration is crucial as both historical and current estimates of incidence in many Canadian [180] and international studies [181] classify only infections < 6 months as recent. As risk of HIV transmission is elevated early in

infection due to high viral load levels [19], being able to reliably classify infections as occurring within 155 days can offer greater value to understanding HIV transmission dynamics.

We were able to optimize the performance of the MBC by varying the threshold τ for calling mixed bases during Sanger sequencing. The universal application of MBC classification depends on the calibration of sequencing instruments and base-callers across laboratories [72]. There is no clear standard for the threshold at which mixed bases are called. Furthermore, both intra and inter laboratory variability is further exacerbated by the manual review of sequences that is accepted as a component of the genotyping process. A mixed base threshold τ of 15%, available in both SeqScape and Trugene, performed best in the MBC. Manual editing was not carried out on any sequences in this analysis, out of concern of introducing further bias. External validation of any MBC method, including ours, would benefit from the use of an expert base-calling system such as ReCall which has been shown to provide consistent base-calling that mimics expert human review [182].

We next explored whether certain categories of sites offered greater discriminatory power to tell recent from established infections, and if the elimination of irrelevant "genetic noise" could improve predictive power of the MBC. We addressed this question by analyzing the three codon positions separately, by examining sites that are under positive and immune selective pressure and by focusing on sites that increased in entropy. Careful analysis of codon positions revealed that, although the third codon position contained most of the information relevant to the MBC, the predictive power of the MBC was reduced when using individual codon positions compared to using the full sequence. We then tested whether HLA-associated sites, known to diversify rapidly after transmission [75, 176], were informative. Applying the MBC to the sites under HLA selection did not improve the resolution of the classification of the stage of infection. It is possible that although variability at these sites initially increases under adaptive immune pressure, continued selective pressure ultimately results in a reduction in the number of mixed bases. In contrast, the performance of the MBC was improved by focusing on those 25% of sites among which an increase in entropy was found between recent and established infections. By including only those sites that increased in entropy, the optimized classifier reached a sensitivity and specificity of 85.2% and 83.5%, respectively.

Our results are consistent with the idea that as HIV infection progresses, the combined effects of rapid viral demographic expansion and diversifying selection pressure [69] result in the accumulation of mixed

bases at both synonymous and non-synonymous positions. Diversification at all three codon positions, in the absence of preferential selection of the mostly synonymous 3rd positions, is quickly followed by an adaptive phase, mostly from immune-mediated selection pressure at HLA-associated sites. Given that changes in the viral genome in the early stage of HIV infection are the result of both neutral and adaptive evolution [69], it is not surprising that the best MBC optimization was achieved using our entropy-based approach that captures both forms of evolution. Future work should validate our list of entropy-identified, highly informative sites, and determine the functional significance of those that were not HLA-associated.

Focusing only on ΔEntropy sites was more consistent with the results from BED testing than an MBC analysis based on full sequences. However, the optimized MBC could only achieve a concordance of 70% with the BED. The high frequency of mixed bases observed in sequences classified as recent by BED is consistent with the assay's tendency to over classify infections as recent [183] and may partially explain the lack of agreement between the two methods. As estimates of BED success vary widely [184, 185], it is possible that the lack of observed concordance on a per specimen basis is reflective of the less than perfect performance of BED in predicting recent infections. Unfortunately, due to the lack of serology positive specimens, it was not possible to run the BED assay on the precisely timed test datasets used in the validation of the MBC. This being said, the widespread reporting of BED results in the scientific literature means that the reconciling of BED results with the MBC and other incidence measurements remains an important, albeit challenging, endeavor.

There is a global need for methods that allow for the correct classification of HIV infections as recent or established. In the absence of a clear gold standard, the sensitivity and specificity of the MBC on a precisely timed dataset is comparable to other assays used to classify recent infections. Our results add to a growing body of evidence that the MBC may be considered as a potential addition to the public health toolbox for the identification of recent infections. A significant advantage of the MBC method is that no additional laboratory work needs to be performed beyond the genotyping performed for clinical or surveillance purposes. A mixed base counter, freely available, has been implemented in DAMBE [173], and the cutoff $\boldsymbol{\gamma}_{155}$ identified in this study can be directly applied to existing *pol* sequences generated for subtype determination and the estimation of the emergence rate of drug resistance. Standardization of an MBC threshold for identification of mixed bases could lead to the re-evaluation of existing longitudinal sequencing databases to create new potential models of HIV transmission patterns.

Although our findings are compelling, there are several limitations to our analysis. Most importantly, even the optimized MBC misclassified 16% of validated recent infections. This may in part be due to an inability to optimize the threshold $\boldsymbol{\tau}$ for calling mixed bases in the *full training dataset*, as chromatograms were not available for 237 specimens. In addition, 18% of the validated established infections in the test dataset contained less than 1% mixed bases, resulting in misclassification. It is possible that the slight decrease in genetic diversity in advanced HIV infection [167] may compromise the classification of some established infections by the MBC. This last point is somewhat ironic in its convergence with the similar limitation of the BED assay. Indeed as patients progress to advanced stages of infection HIV-specific antibodies drop causing BED to misclassify such samples as recent [164]. A principal component analysis, that would incorporate both the MBC as well as results from laboratory assays, such as the BED, or clinical information, may further improve our ability to correctly classify infections.

It is notable that the cutoff $\boldsymbol{\gamma}_{155}$ of 0.45% for classifying recent and established infections at six months is close to that previously demonstrated for classification at one year ($\boldsymbol{\gamma}_{365}$ = 0.5%), suggesting that mutations accumulate much faster during the first six months of infection than later during infection. This is consistent with the notion that mutations accumulate at a decreasing rate over the course of infection [72]. Such a result implies that it may easier to classify recent and established samples for time points earlier than six months.

The sequences used in this study represent a fairly heterogeneous group of early and late infections. The late infection sequences were obtained from individuals who were infected anywhere from 156 days to more than three years prior to the sample being drawn. Furthermore, the early infection group consisted of both very early infections from patients identified as p24 positive prior to seroconversion as well as from individuals who were identified as recently infected based on discordant HIV tests separated by up to 155 days. Despite this heterogeneity, the optimized MBC performed extremely well on the precisely timed dataset. As more sequences collected over a continuous time frame become available, the MBC could be further tested and refined.

In conclusion, we have validated the utility of MBC for a 6-month threshold, and have optimized its predictive ability by focusing on sites that increase in entropy during the course of the infection. Sanger sequencing is slowly being replaced by next generation sequencing technologies which are better suited

to quantifying population genetic diversity. Our laboratory is currently investigating the ability of estimates of genetic diversity obtained through pyrosequencing of HIV samples to predict stage of infection.

Chapter 3

Adaptive Selection of HIV at HLA Epitopes is Associated with Ethnicity in Canada

3.1 Abstract

Host immune selection favors the development of mutations in HIV that allow immunologic escape. Specifically, the human leukocyte antigen (HLA) induces HLA-associated mutation patterns that are allele-specific. As ethnic groups have distinct and characteristic HLA allele frequencies, we can expect divergent viral evolution within ethnicities, but this pattern has never been documented before. Here, we sequenced and analyzed subtype B HIV *pol* genes from 1248 individuals from 5 ethnic groups in Canada. Phylogenetic analysis showed no separation between *pol* sequences from different ethnic groups, yet F_{ST} values showed significant divergence between ethnicities. A total of 17 amino acid sites showed an ethnic-specific fixation pattern ($0.015< F_{ST} <0.060$, $p <0.01$), and 27 codons were inferred to be under positive selection ($p <0.01$), with both strongly associated with HLA sites ($p =1.78\times10^{-6}$ and $p=1.91\times10^{-7}$, respectively). Within the *pol* gene, eight sites were HLA-associated and were correlated with ethnicity, indicating 'adaptive divergence' between the groups studied. Our results emphasize the

importance of collecting ethnicity data when carrying out HIV/ HLA genetic association studies in ethnically heterogeneous populations such as in Canada.

3.2 Contributions

This work was presented at the Robert Cedegren Bioinformatics Colloquium in November 2010, where it received a prize, and at the 18th International HIV Dynamics and Evolution conference in May 2011.

Isabelle Joanisse, Harriet Merks, Dominic Vallée and Kyna Caminiti amplified and sequenced all the samples. Nicolas Petronella provided a Perl script for conversion of amino acid frequency data to Arlequin format. The idea to look for mutation patterns in the sequences was mine and I conducted all the analyses. Stéphane Aris-Brosou and James Brooks offered their guidance and support throughout the project. I wrote the first draft of the manuscript and they both revised and edited it.

3.3 Introduction

Immune-mediated selection pressure is one of the strongest forces driving HIV evolution. Specifically, the human leukocyte antigen (HLA) genes encode cell-surface proteins that bind and display antigenic epitopes cleaved from viral proteins. Epitope display initiates the cytotoxic T lymphocyte (CTL) response and the destruction of HIV infected cells. However, HLA proteins are able to display only epitopes to which they bind tightly. Consequently, amino acid replacements within HIV epitopes can interfere with the CTL response by decreasing the efficiency of epitope binding, disrupting the intracellular processing of epitopes or impairing recognition by T cells. Thus, the incredibly high evolutionary rate of HIV [186], combined with strong selective pressure, permits immune escape through the mutation of sequences that are targeted by the CTL response [73-75].

HLA alleles are extremely diverse, each allele allowing binding to a specific set of viral epitopes. Therefore, HIV escape mutations are HLA-allele specific and, among hosts who share HLA alleles, HIV can evolve convergently at HLA-associated sites [77, 187, 188]. Concordantly, the frequency of HLA-associated polymorphisms in circulating HIV isolates has been shown to reflect the prevalence of HLA alleles in different populations [80]. Links between HLA alleles and the HIV mutation patterns they generate have been established by multiple large scale association studies on cohorts for which both HLA allele data and viral sequences are available [77, 78, 187, 189-191]. In the largest study of its type,

Brumme *et al*. mapped polymorphisms due to HLA immune escape across HIV genome sequences within a multi-center cohort of over 1500 HIV patients (International HIV Adaptation Collaborative, or IHAC) from the USA, Canada and Australia [77]. In a subsequent study, John *et al*. noted that, in view of HLA alleles being differentially distributed across ethnic groups, ethnicity data should be included as a variable for analysis, as well as HLA type and viral polymorphisms. Among the ethnically diverse USA population, ethnic-specific selection patterns were observed even between HLA variants that bind similar epitopes (HLA supertypes) [188].

Like the American HIV epidemic, the Canadian HIV epidemic is ethnically heterogeneous. According to surveillance data reported in 2008 and for which ethnicity data was available, 44.3% of HIV cases were Caucasian, 33.3% Aboriginal, 11.6% African-Caribbean, 4.5% Asian, and 4.1% Latin-American [3]. Of particular note is the over-representation of Aboriginals in the Canadian HIV epidemic, estimated to account for 8% of prevalent infections [192] but only 4% of the population [193]. Populations studies in the USA have shown that HLA allele frequencies differ significantly between the five major 'outbred' ethnic groups: African-Caribbeans, Asians, Caucasians, Native Americans and Latin-Americans [79, 188]. To gain insight into the forces driving the evolution of the HIV epidemic, we sought to investigate whether HIV sequences coming from different ethnic groups in Canada exhibited characteristic mutation patterns resulting from shared host-driven selective pressures.

In the studies cited above, HLA allele frequency data were used to examine HIV evolution, but such data are currently not available for the Canadian population. However, Perez-Sweeney *et al*. recently developed a method to compare host selection pressure between populations in the absence of HLA allele frequency data [194]. In order to examine host selective pressure exerted by patients from different ethnic groups, we compared site-specific frequencies of amino acids in HIV *pol* sequences. Using this method, we show that HIV sequence patterns are divergent between ethnic groups at 8 sites under positive selection, shown in other studies to be under HLA-associated immune pressure. We therefore conclude that HIV is evolving convergently within ethnic groups at these sites in response to shared HLA alleles.

3.4 Results

3.4.1 Epidemiological characteristics of the study population

In order to maximize the probability that observed mutation patterns were associated with the genetic background of the patient currently infected, and were not polymorphisms acquired upon transmission, we chose to include only samples from long-term infections (older than 155 days), as determined by the capture enzyme immunoassay or BED-CEIA test [103]. In the present analysis, sequences from 1248 ethnicity-typed subtype B samples, from established infections, were included. Sequences were separated into five ethnic groups (Table 3.1): Caucasian (907, 72.68%), Aboriginal (179, 14.34%), African-Caribbean (23, 1.84%), Asian (81, 6.49%) and Latin-American (58, 4.65%). The 1239 bp (413 amino acids) sequenced *pol* fragment encompasses the entire protease region (PR, 297 bp, 99aa) and the first 942 nucleotides of the reverse transcriptase gene (RT, 314 aa).

Table 3.1: Epidemiological characteristics of the population

		Aboriginal	African-Caribbean	Asian	Caucasian	Latin-American	All ethnicities
Sex	Female	173	9	6	105	3	1049
	Male	105	14	75	801	54	196
	Other/ unknown	1	0	0	1	1	3
Age	<20	3	0	0	4	0	7
	20-29	33	3	18	129	23	206
	30-39	67	12	34	255	17	385
	40-49	62	4	16	326	12	420
	>50	14	4	13	193	4	228
	Other/ unknown	0	0	0	0	2	2
	Mean	37.39	39.74	37.99	41.57	34.77	40.40
Exposure category	MSM	21	6	47	448	35	557
	MSM/IDU	7	1	1	33	0	42
	IDU	95	2	6	226	5	334
	HET	46	10	23	172	14	265
	Other/ unknown	19	6	5	56	6	92
Total		**179**	**23**	**81**	**907**	**58**	**1248**

Notes–**MSM** men who have sex with men, **IDU** intravenous drug users, **HET** heterosexual.

3.4.2 Viral divergence cannot be explained by phylogenetic history

Because observed sequence patterns could be due to founder effects and phylogenetic clustering among ethnic groups, we first wanted to exclude sequences showing an association between phylogenetic tree topology and ethnicity. A maximum likelihood (ML) tree containing 1272 sequences was first constructed with FastTree 2.1 [195] under the most appropriate model selected by jModelTest [196], which was the general time reversible model with among-site rate heterogeneity (GTR+Γ). Two apparent within-ethnicity clusters of sequences were removed, and ML tree reconstruction was repeated on 1248 sequences. The APE package [197] implemented in R [198] was then used to sample subtrees randomly across the tree for further Bayesian phylogenetic analysis in BEAST [132]. Seven subtrees containing 88-161 sequences were sampled, covering 66.4% of the tree (829/1248 sequences). Subtrees were then tested for clustering of ethnicities in BaTS (Figure 3.1, Supplementary Figure 3.1 for all other subtrees) [199].

Within each subtree, terminal nodes were annotated with our character of interest (ethnicity) and we tested whether the distribution of ethnicity on the phylogeny was non-random. Association indices and parsimony scores both indicated an absence of phylogenetic clustering by ethnicity groups in any of the subtrees (Figure 3.2). Therefore, sequences from ethnic groups were no more clustered on the final tree than expected by chance, and we could proceed with testing whether *pol* sequences nevertheless showed evidence of ethnic-specific mutation patterns.

3.4.3 Seventeen sites in *pol* are divergent between ethnicities

In spite of this lack of shared viral ancestry within ethnic groups, an analysis of molecular variance (AMOVA) [200] of 413 amino acid sites in the *pol* region demonstrated significant divergence in amino acid composition between ethnicities at 17 sites ($p < 0.01$; Supplementary Figure 3.2). Divergent sites were highly polymorphic, as compared to the rest of the *pol* sequence, and their average entropy was seven times higher (0.28 and 0.04; respectively, $p < 5\times10^{-20}$). The extent of population differentiation at these sites was measured through the fixation index (F_{ST} [201]). F_{ST} quantifies the proportion of the observed variation that is contained within the subpopulation ('S', here, ethnic group) as compared to the total population ('T'). Measured in this way, population differentiation could either be indicative of divergence in amino acid composition between all groups, or could point to a deviation within a single ethnic group. Significant F_{ST} values (at the 1% level) ranged from 1.5% to 6.0%, signifying that at sites

that are divergent between groups, ethnic subgrouping accounted for up to 6% of the observed variation. At all but two of these 17 sites the most common amino acid was conserved across ethnic groups; only the frequencies of amino acids differed between groups. The highest F_{ST} values were noted at sites where the most prevalent amino acid differed between ethnicities, PR93 (F_{ST} =0.060) and RT277 (F_{ST} =0.052, Figure 3.2; Supplementary Figure 3.2).

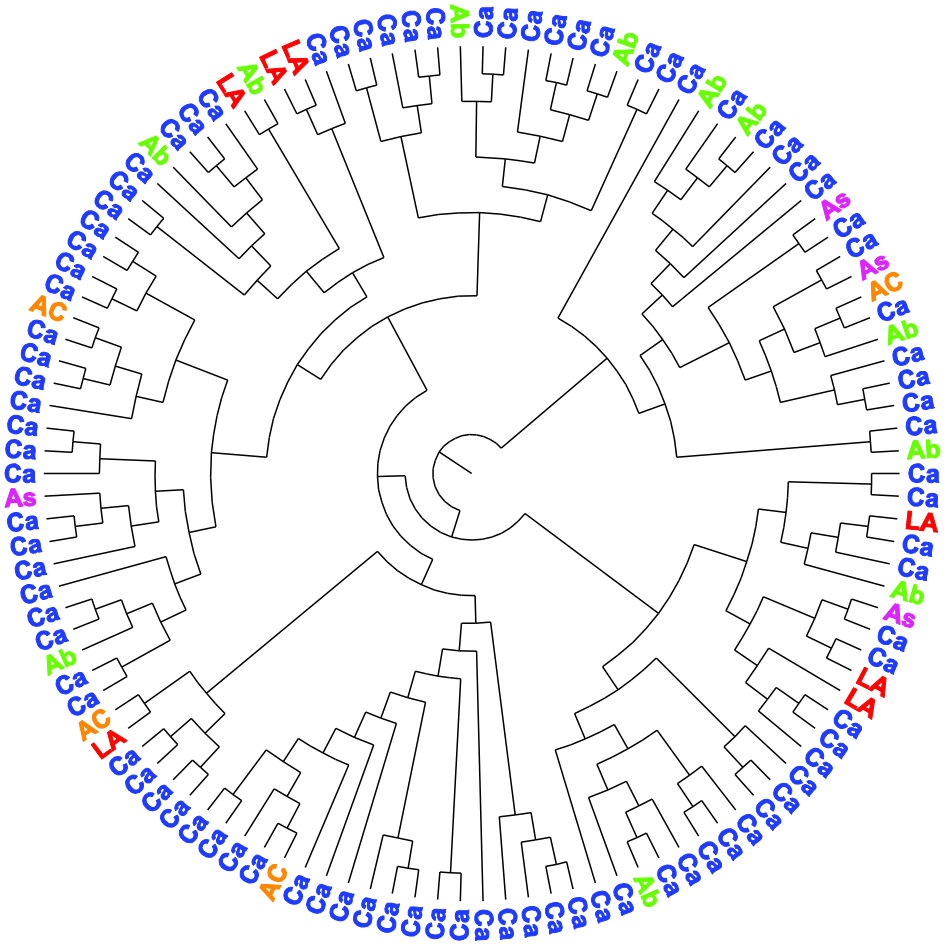


Figure 3.1: Phylogenetic subtree 2. The phylogenetic relationships between taxa in the subtree were reconstructed in BEAST for subsequent analysis in BaTS. For graphic visualization, a maximum clade credibility tree was generated in TreeAnnotator and annotated in FigTree.

Table 3.2: Results from the analysis of seven subtrees in BaTS

	n	AI	*p*	PS	*p*
Subtree 1	161	8.43	0.09	50.33	0.25
Subtree 2	118	5.06	0.50	24.74	0.32
Subtree 3	152	7.70	0.15	44.53	0.12
Subtree 4	99	3.92	0.25	21.05	0.07
Subtree 5	88	3.72	0.70	15.98	1.00
Subtree 6	91	2.52	0.22	12.95	0.08
Subtree 7	121	4.71	0.85	22.89	0.36

Notes–***n*** is the number of sequences in each subtree. The results of two statistics are shown: **AI** association index and **PS** parsimony score. *p* values > 0.05 indicate that sequences are not clustered by ethnicity in the subtrees.

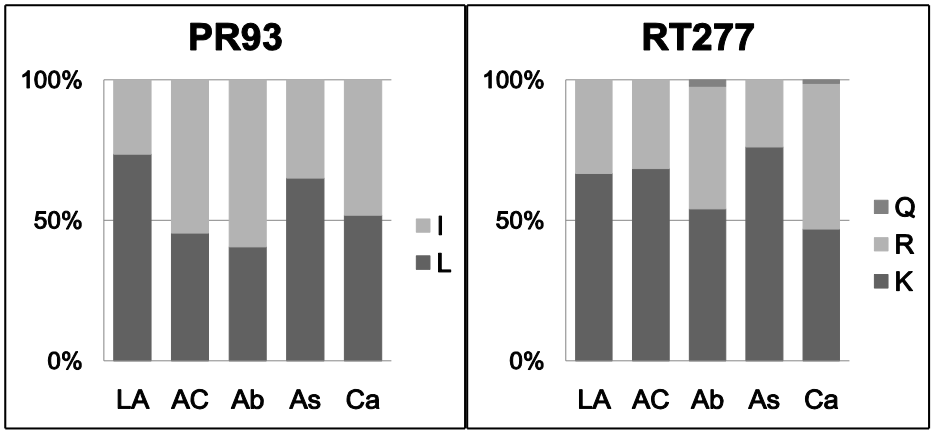


Figure 3.2: Amino acid frequency distributions at sites PR93 and RT277. For each ethnic group, the frequencies of alternative amino acids are shown for one site in protease (PR) and one in reverse transcriptase (RT). Amino acids represented are: leucine (L), isoleucine (I), glutamine (Q), arginine (R) and lysine (K). F_{ST} values at PR93 and RT277 are 0.06 and 0.0521, respectively. Notation of ethnicities is as follows: Aboriginal (Ab), African Caribbean (AC), Asian (As), Caucasian (Ca), and Latin American (LA).

3.4.4 Divergent sites are strongly associated with HLA sites and sites under positive selection

In order to understand the origin of this differentiation at viral sites according to ethnicity, we further investigated the role of HLA-driven convergent selective pressures. Based on the literature, 78 sites known to be HLA-associated (HLA+) were mapped across the sequenced *pol* region [77]. In parallel, using the Single Ancestor Counting algorithm in HyPhy [174], 27 codons were inferred to be under positive selection (Pos+, p <0.01). As most selective pressure in HIV is immune-mediated [78], we expected HLA+ and Pos+ sites to be strongly associated, which was indeed the case (p =2.93×10^{-13}, Fisher's exact test, Figure 3.3) [174].

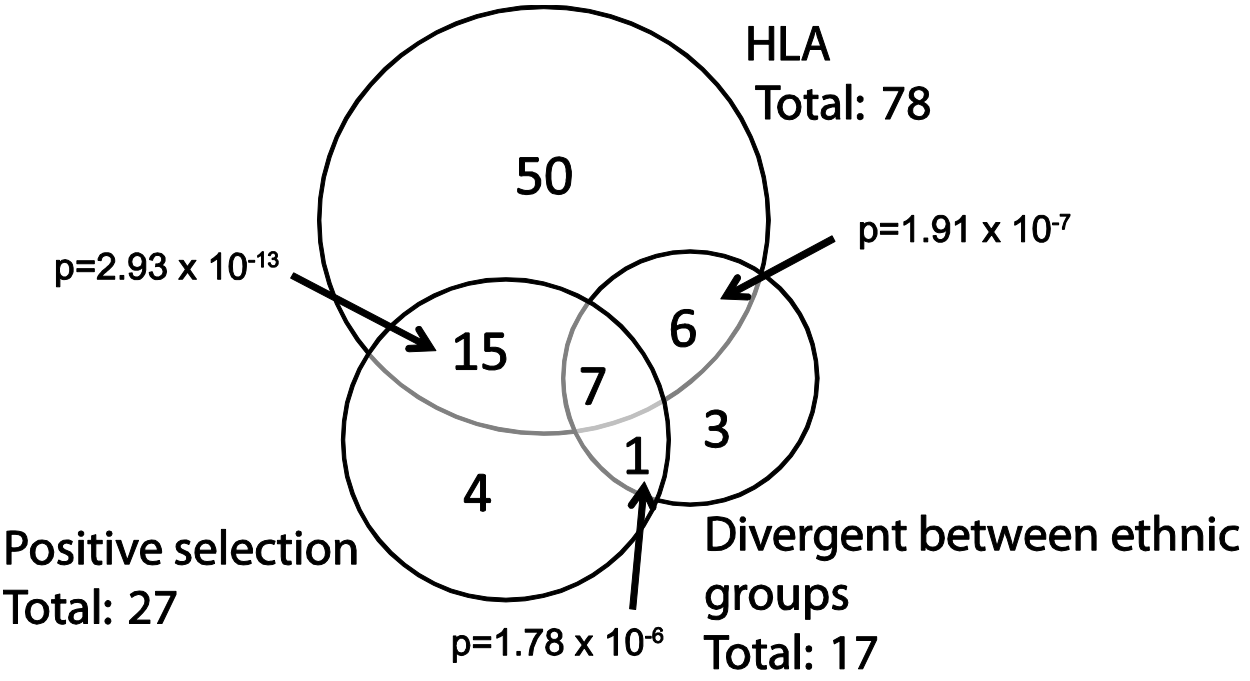


Figure 3.3: Overlap between sites of interest. Of a total of 413 amino acid sites, considerable overlap is observed between HLA-associated sites, sites under positive selection and sites divergent between ethnicities. Associations were tested using Fisher's exact test. Note that all three tests take into account the 7 sites in the center of the diagram.

Sites divergent between ethnicities were also strongly associated with both HLA+ (p=1.91×10^{-7}, Fisher's exact test) and Pos+ sites (p=1.78×10^{-6}, Fisher's exact test, Figure 3.3). Eight codons were both divergent between groups, and inferred to be under positive selection: PR35, PR93, RT135, RT162, RT245, RT277,

RT293 and RT297 (Supplementary Figure 3.2). When sequences from each ethnic group were analyzed separately, sites under positive selection differed between groups (Supplementary Figure 3.2). According to the definition put forward by Perez-Sweeney *et al*. [194], these sites can be considered "adaptively divergent", *i.e.* selective pressure differs at these codons between the groups studied. Of these, all except RT293 were also on the list of HLA-associated sites produced by Brumme *et al*., suggesting that the observed selection pressure is the result of HLA alleles differentially distributed across ethnicities. Moreover, fifteen additional sites were HLA+ and Pos+, although not divergent between ethnicities. These sites appear to be under strong HLA induced selection pressure within the Canadian population, in response to HLA alleles shared among the different ethnic groups examined. Finally, six sites were divergent between groups, and HLA+, but not Pos+ (Figure 3.3). These sites could either be false positives due to low frequency haplotypes, or false negatives not detected by the codon model.

3.5 Discussion

The main aim of this study was to characterize HIV sequence diversity and evolution in the ethnically heterogeneous Canadian epidemic, in the absence of HLA data. We first demonstrated that HIV sequences from patients of diverse ethnic origin showed distinctive mutation patterns, but that these patterns were not associated with viral ancestry. Further analysis confirmed that, although small in number, divergent sites were strongly associated with sites known to be under HLA selective pressure. The five ethnic groups studied had previously been demonstrated to have characteristic HLA allele frequencies [79]; we therefore suggest that HIV has evolved convergently at these sites within ethnic groups because of shared HLA alleles.

Although not included in the initial list of HLA sites used, our data showed that RT293 was divergent between ethnicities and under positive selection [77]. This site has, however, been identified to be under HLA immune pressure in a subsequent HIV/HLA genetic association study performed in Mexico [189]. Because amino acid frequencies at this site differed between ethnicities and positive selection was detected, we suggest that selection pressure at this site is indeed HLA-driven in Canada.

Understanding the forces shaping HIV genetic diversity is crucial to surveillance efforts and for the control of the HIV epidemic. It is clear that HLA alleles, and thus ethnicity, are strong forces shaping the

evolution of HIV [78, 81]. The role of ethnicity is particularly important in Canada where a single ethnic group, Aboriginals, is disproportionately affected by the HIV epidemic. Although most risk factors are behavioral [202, 203], recent evidence suggests there may be a genetic basis to HIV susceptibility among Aboriginals due to HLA alleles. Specifically Keynan *et al.* identified high rates of HLA allele B35 among Aboriginals [204]. B35 is known to cause faster disease progression [205], which could explain the higher proportion of AIDS diagnoses made in Aboriginals [180]. Meanwhile, in newborns, the B35 allele has been shown to increase risk of HIV acquisition from their mothers [206], so in this way B35 may also confer susceptibility to HIV through sex or shared needles. In addition, the B57 allele, which is associated with low viral loads [207] and thus reduced transmission risk [10], was found by Keynan *et al.* to be at a lower frequency among Aboriginals . Consequently, there is ongoing research in Manitoba investigating the role of HLA alleles in disease progression and their contribution to risk of transmission among Aboriginals (Canadian HIV Trials Network CTNPT 004).

Our results reconfirm the recommendation of Carrington *et al*., that studies examining the effects on HLA alleles on HIV evolution require large populations, viral phylogenies, HLA types and information on patients' ethnic background [208]. In John *et al*., high-resolution HLA typing (to four digits) was used to decipher the effect of HLA alleles on HIV evolution. At this level of resolution, different allele frequencies were found between ethnic groups studied, resulting in diverse HLA-associated mutations in HIV even within HLA supertypes. In our study, different HLA-associated mutation patterns were identified between ethnic groups, even in the absence of HLA data.

In HIV research, phylogenetics is used not only for the analysis of global historical evolution patterns, but more recently for the molecular epidemiological analysis of patterns of transmission [209]. In this context, part of the analysis consists in identifying "clusters" of closely related sequences, which represent transmission chains. Clusters are strictly defined based on low intra-cluster distance (usually 0.015 nucleotide substitutions/site [109]) and high bootstrap values. The presence of HLA-driven evolution across ethnicities has implications for the use of *pol* sequences for the reconstruction of transmission events, due to artifactual HLA-driven clustering. Indeed, a recent publication highlighted the role of HLA in driving the evolution of HIV clades and the resulting clustering identified by phylogenetic trees [81]. From our study and others examining HLA induced mutations, it appears that hosts who share HLA alleles, or more recent transmission events where HLA-associated sites have not yet adapted to a new host's genetic background, might be more likely to conform to this strict

definition, and therefore more likely to be considered clusters. In contrast, sequences from related infections where patients have different HLA types may more rapidly diverge. We are currently investigating whether removing such convergent sites alters phylogenetic clustering at the population level. In view of the present findings, we will also be evaluating whether such a cluster definition tends to promote clustering among individuals of the same ethnicity.

Although our findings are compelling, there are several limitations to our analysis. Most importantly, the HLA alleles expressed across our ethnic groups are not population-specific. In particular the Caucasian group is likely to be of admixed origin, and to contain subpopulations with different HLA allele frequencies. Concordantly, a larger number of sites under positive selection were observed among Caucasians than in any other group. Moreover, distinct HLA alleles can drive the selection of the same escape mutant [80], potentially obscuring a correlation between the two. Because of these limitations, it was beyond the scope of this paper to establish links between the mutation patterns observed in HIV and frequencies of HLA alleles in the different ethnic groups studied. However, despite the potentially imperfect grouping in our analysis, statistical associations were found between mutation patterns and ethnic groups, confirming that our analysis is robust.

Another limitation of this type of analysis is that HLA-associated mutation patterns in HIV may not necessarily reflect the HLA background of the patient currently infected. Although HIV evolves rapidly within a new patient to match his or her HLA type [210], usually during early infection [75], there is also evidence that HIV may revert to wild type [211] or, in the absence of a fitness cost, maintain escape mutants associated with the transmitter's HLA type. In the acute stage in particular, an adaptation profile specific to the newly infected individual may not yet have been generated. Instead, HLA-associated sites may reflect adaptation to a previous HLA type. To avoid this issue, we used HIV sequences only from patients classified as having established infections according to the BED assay, as with time HIV is more likely to adapt to the new patient's HLA type.

We have shown that HLA-associated mutation patterns differ across the ethnicities studied. These observations support ongoing research investigating whether differences in HLA allele frequencies could explain the susceptibility of Aboriginals to HIV. Furthermore, our study illustrates the importance of incorporating ethnicity information to future HLA association studies in Canada. It may also be important to consider HLA background and ethnicity when identifying clusters of related infections,

because of convergence at HLA-associated sites. Finally, ethnic specific host influence on viral evolution may prove important in vaccine development, as an effective vaccine will require targeting regions that are reactive across groups.

3.6 Materials and Methods

3.6.1 Study Population

The Canadian HIV Strain and Drug Resistance Surveillance Program (SDR), receives serum samples from all newly diagnosed, treatment-naïve HIV patients in Canada. During sample collection, basic epidemiological data including age, sex, ethnicity and exposure category are recorded. Between 2002 and 2009, 3648 samples were sent to the National HIV and Retrovirology Laboratory in Ottawa. The SDR program has institutional research ethics board approval.

3.6.2 Laboratory analysis

For each sample, viral RNA was extracted and the HIV *pol* region amplified by RT-PCR for sequencing as previously described [149]. *Pol* sequence fragments from different sets of primers were assembled in BioEdit v7.0.4.1 [170]. Assembled sequences were translated for in-frame alignment to the NCBI HIV-1 subtype B reference genome (accession number: NC_001802) using TranslatorX [212]. Nucleotide sequences were trimmed to identical length (1239 bp) and deposited in GenBank (accession numbers: HM468499-HM469374). All sequence manipulations, such as nucleotide to amino acid translations, were carried out in BioEdit. For each sequence, the subtype was determined by submission of *pol* to the REGA HIV-1 Subtyping Tool v2.0 (http://dbpartners.stanford.edu/RegaSubtyping/). In addition, stage of infection was determined for each sample as < or > 155 days (recent vs. established, respectively) using the Calypte BED-CEIA™ (capture enzyme immunoassay). The BED-CEIA (or BED for short) measures the proportion of IgG antibodies which are HIV-specific in a sample [103]. Only subtype B samples, which account for the majority of infections circulating in Canada [101], and originating from established infections, as determined by BED, were included in the present analysis. Sequences were divided into five datasets based on ethnicity: Caucasian, Aboriginal, African-Caribbean, Latin-American and Asian, for analysis.

3.6.3 Phylogenetic analysis and character association

Phylogenetic interrelationships between 1272 sequences were reconstructed in FastTree 2.1 under a GTR+Γ model, as selected by jModelTest. Two clusters of sequences for which there was an association between ethnicity and phylogeny were removed from the dataset and the ML reconstruction was repeated on 1248 sequences. Using the APE package in R, subtrees were randomly selected for Bayesian phylogenetic analysis in BEAST. Sequence subsets were run in duplicate under a Bayesian Skyline Plot model with 10 breakpoints and linear splines. Convergence was assessed in Tracer after 100 million generations. After elimination of a burn-in period (10-20% of run), a posterior distribution of trees was generated for analysis in BaTS. Terminal nodes were annotated with our character of interest, ethnicity, and the non-random distribution of ethnicity was tested in each subtree.

3.6.4 Positive selection and population divergence

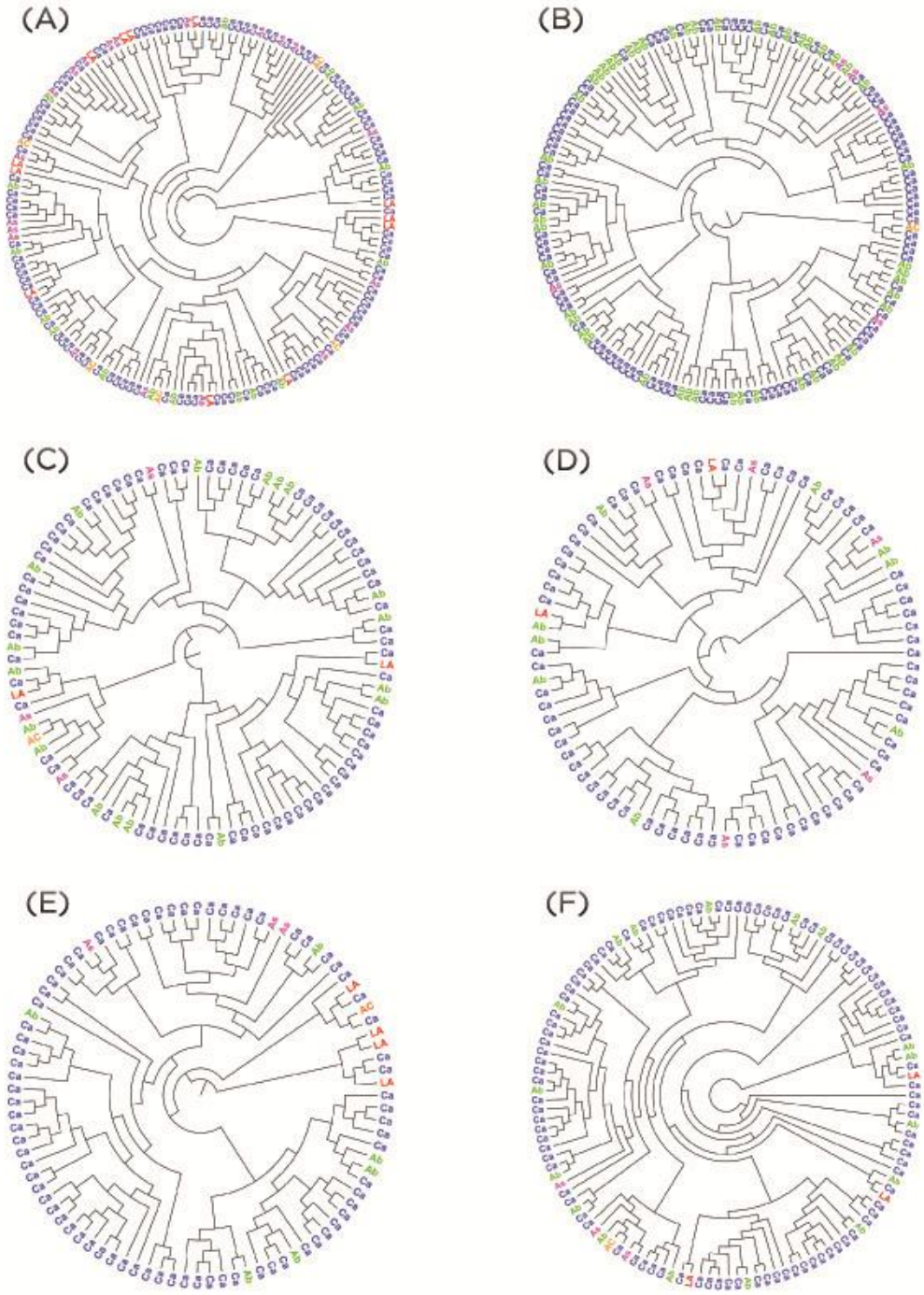
Sites previously identified to be under HLA-mediated selection [77] were mapped along sequences in the alignment. Sites inferred to be under positive selection ($p<0.01$) were determined using HyPhy in each ethnic group dataset. In datasets exceeding 50 sequences, the SLAC algorithm was employed [213]. The SLAC algorithm calculates observed numbers of non-synonymous (N) and synonymous (S) mutations at each codon in an alignment, and compares these to expected numbers (E[N] and E[S]) in order to estimate selection pressure (dN=N/E[N], dS=S/E[S]). A low dN/dS ratio (<1) indicates purifying selection, while a high dN/dS (>1) suggests diversifying positive selection pressure. In datasets <50 sequences, both the FEL and REL algorithms were used, and only sites appearing in both lists were considered to be under positive selection (as recommended by the HyPhy user manual).

Site-specific entropy values for each amino acid within the alignment were calculated using the Los Alamos Entropy Tool (http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html). Entropy is a measure of the variability at each position in an alignment; a site with high entropy is highly variable.

Finally, population divergence between ethnic groups was measured at the amino acid level by calculating the Fixation Index (F_{ST}) using an analysis of molecular variance, as implemented in Arlequin v3.5.1.2 [214]. The significance of the input genetic structure (here, ethnic groups) is tested by permuting haplotypes between populations and recalculating all statistics to estimate their null distribution. In order to calculate F_{ST}, positional amino acid frequency data were generated in BioEdit

and transformed into a format readable by Arlequin using an in-house Perl script (available upon request).

In order to determine whether sites that were divergent between ethnic groups were associated with sites under positive selection and HLA-associated sites, we used Fisher's exact test. All statistical analyses were carried out in SPSS [175].



Supplementary Figure 3.1: All the subtrees. Maximum clade credibility reconstructions for all six remaining subtrees evaluated in BaTS. Taxa are colored by ethnicity: Aboriginal (Ab, green), African Caribbean (AC, orange), Asian (As, pink), Caucasian (Ca, blue), Latin-American (LA, red). (A) subtree 1, (B) subtree 3, (C) subtree 4, (D) subtree 5, (E) subtree 6, (F) subtree 7.

Supplementary Figure 3.2: Sites of interest in the pol sequence. Only amino acid sites that are HLA-associated (in blue, 78 sites), positively selected (pink, 27 sites) or divergent between ethnicities (green, 17 sites) are shown. The top line indicates position in protease (PR) and reverse transcriptase (RT) in the reference sequence HXB2. The second line is the amino acid in HXB2. For sites divergent between ethnicities, amino acids are displayed in order of frequency. HLA alleles driving selection at each HLA-associated site are displayed in the bottom line.

Chapter 4

Conclusions and Future Directions

4.1 Can evolutionary approaches solve the HIV epidemic?

The year 2011 marks 30 years since the first AIDS diagnosis. After HIV was identified as the etiological agent of AIDS, researchers and politicians promised the world a vaccine and a cure would be available within a year. It turned out, however, to be much more complicated than anticipated, in part because of the extraordinary adaptability of HIV.

However, this great ability of HIV may also be the key to its demise. In recent years, evolutionary analyses have yielded an incredible amount of information about HIV which can be used to prevent new infections. Resolving the origin of HIV has made it possible to monitor regions of the world where zoonotic transmissions occurred to prevent new cross-species events [54]. Applying phylogenetics to transmission networks has isolated the factors most strongly associated with transmission, so that prevention campaigns can be better targeted [215]. In treated patients, drug regimens tailored to HIV sequences prevent treatment failure due to drug resistance. In fact, no study on new antiretrovirals or on a vaccine is complete without an evolutionary component.

The fastest moving field of scientific research is the interface between biology and engineering. HIV and other rapidly evolving viruses such as influenza are particularly suited to this type of analysis. Combined with technologies such as GPS and 3D imaging, real-time monitoring and tracking systems have been

developed based on genetic sequences. Thus, the spread of epidemics can be visualized; potentially highlighting new possibilities for prevention [216, 217].

In this thesis I used a bioinformatic approach to address two problems. I first developed a tool to distinguish recent from established infections based on genetic sequence. Then, focusing solely on established infections, I examined mutation patterns in HIV that were shared by hosts of the same ethnicity. Both studies used minimal information in addition to sequences, as the SDR collects only limited epidemiological data and no clinical data. Nevertheless, both studies yielded important results.

In Chapter 2, I developed a classifier that distinguished between infections < or > 155 days based on the proportion of mixed bases in *pol* sequences generated during population-based Sanger sequencing. The correct identification of recent infections is crucial because of their contribution to onward transmissions and for the calculating of incidence rates. The mixed base classifier (MBC) put forth by Kouyos *et al* in their groundbreaking publication earlier this year distinguished infections < or > 1 year [72]. Our re-calibration is in better concordance with the more standard 6-month threshold. We were further able to improve the performance of the MBC. Using an entropy-based approach, we identified sites more likely to contain mixed bases in established infections, and using these sites alone in the MBC radically improved its specificity. The MBC performs as well as laboratory-based assays to identify recent infections, and can be carried out at no additional expense on *pol* sequences. However, just like the other tools available, the MBC does not achieve a sensitivity and specificity of 100%. Indeed, a problem may remain for *pol* sequences from patients in advanced stages of infection. The MBC would hypothetically misclassify these infections as recent because genetic diversity is known to decrease. However, stage of infection in this case would be evident to a trained physician, or by measuring CD4 levels. As discussed in Chapter 2, further studies should therefore test the MBC in combination with other tools, or by incorporating clinical data. Moreover, if the sites detected here are shown to improve MBC performance in other datasets, future work should also focus on identifying the function of those that were not HLA-associated, and on elucidating the pressures resulting in their diversification during the course of infection.

In Chapter 3, I identified differences in *pol* sequence patterns across five major ethnic groups in Canada. The characterization of HIV epitopes in circulating isolates is essential for the correct selection of vaccine epitopes appropriate for the epidemic at hand. Furthermore, an understanding of HLA allele diversity

across the country is crucial to ensure any vaccine developed will be broadly reactive across hosts. Future work in this area would hopefully take into account the results presented here, and incorporate ethnicity information to HLA/HIV genetic association studies, as discussed in Chapter 3. This is of particular importance in Canada, where a single ethnic group, Aboriginals, is so disproportionately affected by the HIV epidemic. In addition, characterizing HLA alleles and the escape mutants associated with them may also prove important if drug regimens become tailored to host genetic background. For example, personalized HAART regimens in the future might include supplements targeting HLA escape mutants predicted from patients' HLA alleles. Already, screening is conducted for the allele HLA B*5701 because of its known association with hypersensitivity to the antiretroviral abacavir.

As well as the specific applications of the results in each chapter, this thesis as a whole highlights the wealth of information contained within viral sequences. Thus evolutionary approaches have huge potential for tackling aspects of the HIV epidemic. Nevertheless, the importance of epidemiological, clinical and sociological research cannot be overstated.

As discussed in Chapter 1, HIV sequences such as those generated through the SDR program can also be used to investigate the history of HIV in Canada. Ongoing work is addressing whether the HIV epidemic is under control in Canada, and whether there are differences in the epidemic dynamics between provinces.

4.2 Ethical considerations

Chapter 3 investigates whether common mutational patterns are observed in HIV in patients of the same ethnicity. This piece of research raises certain ethical considerations because of the use of ethnicity as a variable. Indeed, the merit of using of ethnicity as a classifier in biomedical and epidemiological research has been questioned, and the term in itself does not have a generally agreed upon definition.

Studies have repeatedly demonstrated associations between self-identified ethnicity and susceptibility to illness and disease outcome. For example, African Americans suffer higher rates of cardiovascular events than Caucasians in the USA, while rates of HIV in Canada are higher amongst Aboriginals than any other group. However, these observed patterns are for the most part not biological, but sociological [218]. Ethnicity is not the cause of the observed differences, but instead is linked to disparities in the

social determinants of health: culture, diet, socioeconomic status, access to health care and education [219]. Although ethnicity associates with the observed outcome, it is not causal, and so serves as a surrogate variable concealing a far more complex truth. As such, the use of ethnicity as a surrogate variable has been severely criticised [220]. Nevertheless, epidemiologists counter that they are limited to the data available when estimating associations between variables and outcome. For example, in the case of HIV only four variables are collected: age, sex, exposure category and ethnicity. Furthermore, its role as a proxy variable is in fact the reason ethnicity data is often collected by public health agencies. Its inclusion highlights the vulnerabilities of disadvantaged populations and can help target prevention campaigns. Finally, inclusion of ethnicity data ensures that ethnicities that are traditionally underrepresented in biomedical research and often overrepresented in diseases are included in studies.

In the study presented here, ethnicity is not used as a proxy for sociological determinants of health, but rather for its genetic basis, in reference to the geographical ancestry of patients. Based on multi-locus genetic data, humans cluster into the five major geographic regions: Africa, Europe and Middle East, East Asia, Oceania and the Americas [221]. Even more surprisingly, Europeans cluster into their native countries [222]. Here, we sorted patients into five groups on the basis of HLA genes: HLA alleles are differentially distributed across our five ethnic groups [79, 188], and are one of the strongest predictors of HIV mutation patterns [78, 187]. In the paper, we reject the hypothesis that the patterns shared within ethnic groups are a consequence of shared viral ancestry, and demonstrate that they are more likely linked to HLA alleles shared among hosts of the same ethnicity. Heterogeneous human populations such as are found in North America thus act as a microcosm in which to study the impact of human genetic diversity on the evolution of HIV.

There are several limitations to our classification, however, only some of which are addressed in the discussion of chapter 3. First, the genetic basis of ethnicity is imperfect. In fact, 90% of genetic variation is found within ethnic groups rather than between them. In addition, based on multi-locus genetic data, Hispanics do not cluster alone but rather are distributed across the other ethnic groups; however, they do differ from other groups based on HLA allele frequencies alone. Caucasians, meanwhile, are a large group of likely admixed origin that are not very well defined genetically. Secondly, ethnicity is self-identified, and may not reflect true 'genetic' ethnicity. Because the ethnicities studied have cohabited in North America for some time, it is likely that many individuals are of mixed ancestry, even without

necessarily being aware of it. Individuals of self-declared mixed ancestry were excluded from the analysis, but this remains a limitation nevertheless, which will worsen as ethnic groups continue to mix.

Because of all the limitations addressed above, implications and conclusions should be drawn from this paper with caution. Most importantly, it is not at all the wish of the authors that this paper be used to justify research or policies of a racist nature. To the contrary, the authors wish to highlight the importance of including Aboriginal patients in research on HIV evolution and vaccine development in Canada, in particular in view of their overrepresentation among HIV cases. In other parts of the world, similar efforts are needed to ensure that studied populations are of mixed ethnic origin.

In parallel, further research on the social determinants of health, as they pertain to HIV, is desperately required, and efforts must be made to ensure that basic scientists take an integrated approach to their research. Both worldwide and in Canada, HIV is concentrating in the most vulnerable of populations. The highest risk predictors for HIV are country of birth and poverty. Risk is further exacerbated by gender inequality, racism and other types of discrimination; abuse during childhood, access to healthcare and to education. As well as increasing risk of acquiring HIV, the social determinants of health influence whether patients will have access to ART, and thus how well they can manage their infection and how fast they progress to AIDS. An integrated approach to healthcare is so important that the WHO established in 2005 a Commission on Social Determinants of Health, which aims to improve living conditions and reduce inequalities. This has also been the approach of the Bill and Melinda Gates foundation, which focuses as much on improving access to clean water and sanitation, decreasing homelessness and developing agriculture, as on basic scientific research. The Gates foundation further emphasizes the importance of knowledge translation: research carried out in the lab should be followed through until it is applicable in the field. Social determinants of health act both at the individual level, by increasing the likelihood that an individual will engage in high-risk behaviours, and at the societal level, for example by undermining the sustainability of national health care systems [218]. As such, reducing inequities should be a focus of all health-related research and policy.

4.3 Ongoing work

The delivery of healthcare is a provincial responsibility in Canada, and the provinces have each developed different strategies to control the HIV epidemic. British Columbia (BC), in particular has been

applauded for the vast extent of its HIV programs targeted towards intravenous drug users (IDU), men who have sex with men (MSM), and Aboriginals [223]. BC was the first province to instigate a needle exchange program, in Vancouver in 1989; now over 100 exist across the country. BC still runs the only safe injection site in Canada. Furthermore, differences in the epidemic dynamics are expected between provinces, because 81% of HIV case reports are distributed across only three provinces: Ontario (ON), Quebec (QC) and BC. Finally, 6000kms separate Halifax from Vancouver, making Canada the second largest country in the world, yet travel within the nation's borders is common. Thus it is unclear whether the epidemics in each province are distinct from each other.

To address these questions, we sought to reconstruct the Canadian HIV epidemic for each of five provinces for which time-stamped *pol* sequences were available: BC, Saskatchewan (SK), Alberta (AB), Manitoba (MB) and Nova Scotia (NS). QC does not currently participate in the program, while samples from ON were available only from a single year. Preliminary investigations in BEAST v1.6.1 [132] indicated that MCMC chains for large datasets (>250 sequences) would not converge in a reasonable amount of time. For the provinces of MB and NS, fewer than 250 sequences were available; we therefore used all sequences from these two provinces. For the provinces of BC, SK and AB, we subsampled two replicates of 100 sequences, without replacement, for analysis.

For each sequence dataset, three parametric and one non-parametric demographic models were compared (constant population size, logistic increase, exponential increase and Bayesian skyline plot, or BSP) and three clock models (strict, lognormal, and random) under a general time reversible substitution model with rate heterogeneity (GTR+Γ), as selected by jModelTest [196]. Hue *et al.* demonstrated that HIV collection dates spanning ≤5 years were insufficient for the accurate calculation of evolutionary rates [224]. As the span of collection dates for each province ranged from three to six years, we input the authors' calculated substitution rate for *pol* of 2.55×10^{-3} [224]. In total, six models were run on each dataset for 100 million generations and convergence was assessed using Tracer (tree.bio.ed.ac.uk/software/tracer). After elimination of a burn-in period (10-20% of chains), the marginal likelihood for each model was estimated in Tracer, and the model with the highest marginal likelihood was selected for subsequent analysis.

Table 4.1: Marginal likelihood for each demographic and clock model tested

Models	BC	SK	AB	MB	NS
cst_strict	-16,122.045 (+/- 0.167)	-8,553.247 (+/- 0.109)	-15,893.806 (+/- 0.144)	-31,112.376 (+/- 0.187)	-18,468.713 (+/- 0.172)
cst_logN	-16,101.098 (+/- 0.176)	-8,546.711 (+/- 0.139)	-15,822.834 (+/- 0.198)	-30,954.818 (+/- 1.104)	-18,369.513 (+/- 0.207)
cst_random	-16,117.286 (+/- 0.159)	-8,547.981 (+/- 0.140)	-15,868.762 (+/- 0.149)	-30,996.723 (+/- 0.381)	-18,451.010 (+/- 0.298)
log_logN	**-16,095.044* (+/- 0.179)**	-9,170.917 (+/- 0.120)	**-15,819.622* (+/-0.167)**	-30,976.725 (+/- 0.371)	-18,368.529 (+/- 0.220)
exp_logN	-16,095.268 (+/- 0.178)	-9,166.319 (+/- 0.147)	-15,819.680 (+/-0.133)	**-30,930.002* (+/- 0.458)**	**-18,367.746* (+/- 0.221)**
BSP_logN	-16,104.214 (+/- 0.147)	**-8,545.442* (+/- 0.121)**	-15,821.180 (+/- 0.179)	-30,980.394 (+/- 0.394)	-18,387.250 (+/- 0.171)

NOTES–British Columbia (BC), Saskatchewan (SK), Alberta (AB), Manitoba (MB), Nova Scotia (NS). Demographic models: Constant (cst), logistic (log), exponential (exp), Bayesian skyline plot (BSP); rate models: strict clock (strict), random local clocks (random), uncorrelated lognormal (logN). * Indicate models with the highest marginal likelihood, used for analysis. SEs are indicated in brackets.

Table 4.2: Basic statistics and parameter estimates of HIV dynamics in different Canadian provinces

	BC	SK	AB	MB	NS
Total *n*	1673	557	469	242	135
n	100	100	100	242	135
Root for subtype B	1974	1986	1985	1988	1978
Incidence	44094.547	NA	1653.063	278.171	3100.121
Pop size	4500000	1000000	3725000	1233000	1000000
Incidence rate	0.98%	NA	0.04%	0.02%	0.31%
Growth rate	0.575	NA	0.183	0.149	0.142
Substitution rate	2.531×10^{-3}	2.550×10^{-3}	2.448×10^{-3}	8.266×10^{-3}	2.211×10^{-3}

NOTES– British Columbia (BC), Saskatchewan (SK), Alberta (AB), Manitoba (MB), Nova Scotia (NS). Total *n* is the total number of samples received from the province since the SDR program was instigated. *n* is the number of sequences used in the present analysis. The roots of subtype B clusters were estimated from phylogenetic trees. Pop size refers to the population size of the province (from Wikipedia), thus incidence rate was calculated by dividing the incidence by the population size. NA: not applicable (different parameters are used by the BSP prior).

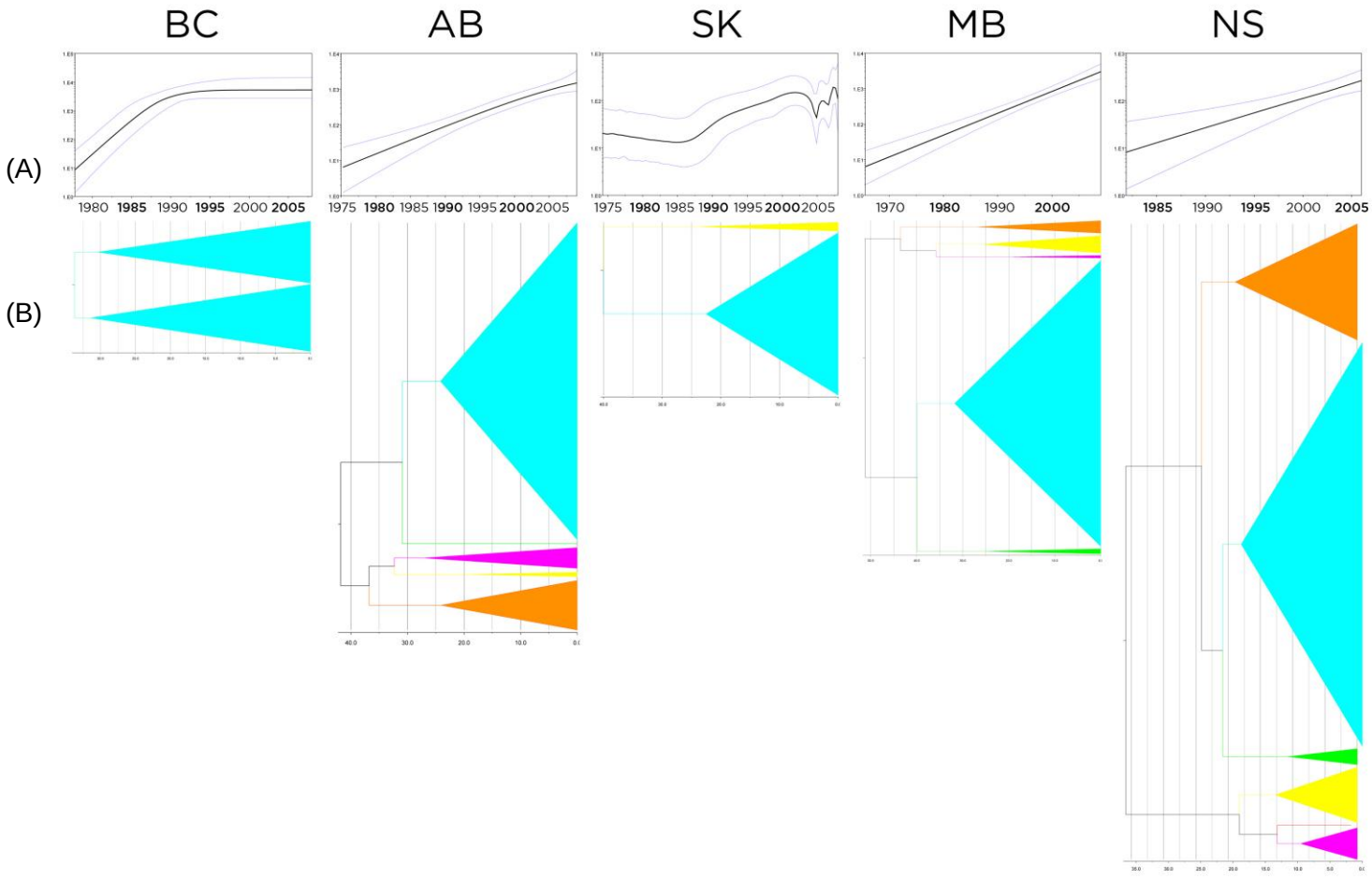


Figure 4.1: Demographic reconstructions (A) and dated phylogenies (B). British Columbia (BC), Saskatchewan (SK), Alberta (AB), Manitoba (MB), Nova Scotia (NS). Different subtypes are shown in different colors: subtype B blue, C orange, A and AG pink, other A (mostly A1, some AE) yellow, D green, K red.

Only in BC does the epidemic appear to be under control. For all provincial datasets, the strict molecular clock was rejected, consistent with the work of others [65, 113, 115, 116, 137-139]. The most appropriate demographic model differed between provinces. Most importantly, HIV epidemic growth was found to be exponential in MB and NS, but logistic in BC and AB (Figure 4.1). Meanwhile in SK, BSP was selected as the best demographic model. Incidence in SK appeared to have increased rapidly since the 1990s. In AB, MB, and NS, the incidence increased each year since the introduction of HIV, while it has remained stable in BC since 1995. Nonetheless, the growth rate in BC was 3 times higher than in any of the other provinces (Table 4.2).

HIV subtypes are differentially distributed across the country. For each province, a consensus time-scaled phylogeny was reconstructed under the best model (Figure 4.1). In BC, two clusters were observed, both of which exclusively contained subtype B sequences. In the other provinces, however, B and non-B clusters of infection were apparent. In SK, 94/100 sequences were subtype B, with a small cluster of 6 subtype A1 sequences. In AB, 22/100 sequences were non-B, including clusters of subtype C, A and G/AG. In MB, 87/242 sequences were non-B (34%), belonging to subtypes A, C, D and G/AG. Meanwhile in NS, 19/135 sequences were non-B (14%), including subtypes A, C, D and G. Thus, although subtype B accounts for 88.3% of infections in Canada this number hides important geographic variation in the distribution of subtypes, as noted previously [225]. BC is dominated by subtype B and accounts for the majority of new diagnoses in Canada (Figure 4.2). However, in MB, non-B subtypes account for 33.7% of sampled infections.

HIV subtype B spread through Canada during the 1970s and 1980s. Based on the dated phylogenies, we estimated the age of the root of the B subtype cluster in each province (Table 4.2). The HIV epidemic seems to have diversified in the 1970s in BC and NS, on each coast of Canada, but not to have spread inland until the 1980s. Overall, these dates are consistent with studies placing the most recent common ancestor of the North American subtype around 1969 [143, 226].

Non-B subtypes are over-represented among heterosexuals, women and African-Caribbeans. Previous studies have demonstrated that in the UK the subtype B epidemic circulates mainly in MSM populations [141], while non-Bs circulate mostly in heterosexuals [140]. We tested this hypothesis with our Canadian data. The distribution of subtypes across the two exposure categories was highly biased, with non-Bs

much more frequent among heterosexuals (X^2=253.93, 1df, p <0.0001). A higher proportion of non-Bs was also apparent among females than among males (X^2=185.84, 1df, p <0.0001), and among African-Caribbeans than among Caucasians (X^2=2693.37, 1df, p <0.0000) (Table 4.3). Such observations highlight that comparisons of the population dynamics of each province and subtype will reflect the underlying characteristics of the transmission networks.

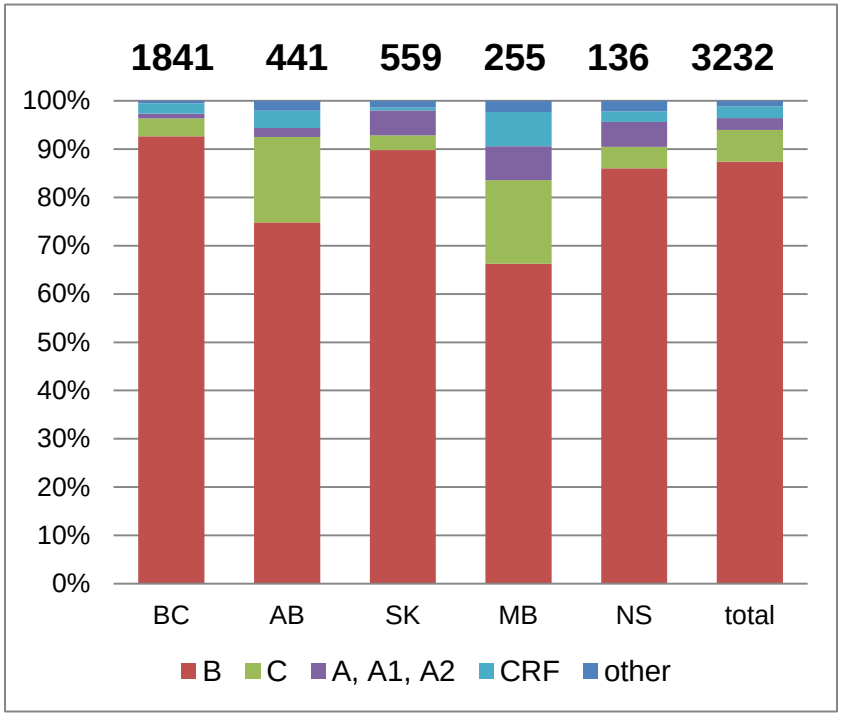


Figure 4.2: Subtype distribution of samples received through the SDR program 2003-2010. British Columbia (BC), Saskatchewan (SK), Alberta (AB), Manitoba (MB), Nova Scotia (NS). Circulating Recombinant Forms (CRF) include AE, AG, AB, AD and BC. Numbers above the columns indicate the number of samples received from each province and in total.

Demographic patterns and past population dynamics may be shaped by the uneven geographical distribution of subtypes. For example, the evolutionary rate of HIV in MB was three times higher than in any other province (Table 4.2). This could potentially be due to MB having the highest proportion of circulating non-B subtypes. Alternatively, an inverse correlation between epidemic size and evolutionary rate has been demonstrated [227]. Thus, this high rate may be a consequence of the small size of MB's epidemic. Similarly, the logistic growth model was selected only for BC, where only subtype B sequences

were analyzed. In spite of the unduly (and as yet unexplained) high growth rate in BC (Figure 4.2), observed growth patterns could therefore suggest that the epidemic is under control in BC or that the subtype B epidemic is under control across Canada but this is masked by the co-circulation of non-B subtypes in the other provinces.

Table 4.3: Differences in epidemiological characteristics between B and non-B patients

	B	non-B	X^2	*p*
Exposure category				
MSM	1026	34	253.9325	<0.0001*
HET	538	227		
Sex				
Female	661	231	185.8481	<0.0001*
Male	2183	189		
Ethnicity				
Caucasian	3643	60	2693.3790	<0.0001*
African-Caribbean	35	240		

NOTES– MSM Men who have sex with men, HET heterosexual. * Indicate a significant difference between B and non-B subtypes. The degree of freedom is 1 for each test.

4.4 Future directions

Subtypes should be analyzed separately to better characterize and compare HIV epidemics between provinces and estimate subtype-specific growth rate, incidence and evolutionary rate. Our next step will consist in analyzing subtypes B and C separately for each province, the number of sequences for other subtypes being insufficient for provincial analysis. This approach will also allow us to make comparisons between subtypes. As noted previously, subtype B appears to be better controlled worldwide than subtype C [116]. We will test whether this is also true in Canada.

Furthermore, it has been suggested that the high local frequency of non-Bs in MB and AB is related to travel and immigration from countries where non-B subtypes predominate [225]. We will conduct a maximum likelihood phylogenetic analysis on non-B sequences from across the country to identify groups of closely related infections. Such clusters are defined based on low intra-cluster distances and high bootstrap values, as explained in Chapter 1. As the SDR includes samples only from infections acquired in Canada, this will allow us to determine whether non-B transmissions are taking place within the country. Moreover, if such clusters are indeed identified, their history could be reconstructed as has been done in the UK [140, 141] to resolve more accurately their origin and rate of spread.

If our results indicate that non-B transmission clusters are established in Canada and that non-B infections are spreading more rapidly than subtype B infections, this could indicate that non-Bs are an increasing threat to public health. As we saw in the previous section, non-B infections are circulating among specific epidemiological groups, namely heterosexuals, women and African Caribbeans. Numerous studies have demonstrated the importance of specifically targeting groups with prevention campaigns [228]. Until now, the most successful campaigns have targeted MSM. Thus public health should redirect messaging towards a different audience.

Furthermore, if subtype diversity is increasing in Canada, this would have important implications for vaccine design. Current vaccine strategies are likely to protect only against a limited range of subtypes. For example, the vaccine tested in the successful trial in Thailand protected only against subtypes B and E [45]. Thus vaccines developed for the Canadian epidemic might have to cover a broader repertoire of subtypes.

Finally, future work would benefit from incorporating geographical information available with each sample to reconstruct the spatial dispersal patterns of HIV across Canada as well as temporal. A number of phylogeographic studies have already been performed, and a user-friendly tool has only just been made available [217].

References

1. UNAIDS/ WHO Working Group on Global HIV/ AIDS and STI Surveillance. AIDS Epidemic Update 2009. In: 2009.
2. UNAIDS/ WHO Working Group on Global HIV/ AIDS and STI Surveillance. Epidemiological Fact Sheet on HIV and AIDS, Canada, 2008 Update. In: 2009.
3. Public Health Agency of Canada. HIV and AIDS in Canada. Surveillance Report to December 31, 2008. In. Surveillance and Risk Assessment Division CfCDaIC (editor); 2009.
4. Dosekun O, Fox J. An overview of the relative risks of different sexual behaviours on HIV transmission. *Curr Opin HIV AIDS* 2010; **5(4)**:291-297.
5. European Study Group on Heterosexual Transmission of HIV. Comparison of female to male and male to female transmission of HIV in 563 stable couples. *BMJ* 1992; **304(6830)**:809-813.
6. Gray RH, Wawer MJ, Brookmeyer R, Sewankambo NK, Serwadda D, Wabwire-Mangen F*, et al.* Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *Lancet* 2001; **357(9263)**:1149-1153.
7. Boily MC, Baggaley RF, Wang L, Masse B, White RG, Hayes RJ*, et al.* Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *Lancet Infect Dis* 2009; **9(2)**:118-129.
8. Schwarcz S, Scheer S, McFarland W, Katz M, Valleroy L, Chen S*, et al.* Prevalence of HIV infection and predictors of high-transmission sexual risk behaviors among men who have sex with men. *Am J Public Health* 2007; **97(6)**:1067-1075.
9. Rothenberg RB, Wasserheit JN, St Louis ME, Douglas JM. The effect of treating sexually transmitted diseases on the transmission of HIV in dually infected persons: a clinic-based estimate. Ad Hoc STD/HIV Transmission Group. *Sex Transm Dis* 2000; **27(7)**:411-416.
10. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F*, et al.* Viral load and heterosexual transmission of Human Immunodeficiency Virus type 1. *N Engl J Med* 2000; **342(13)**:921-929.
11. Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, Laeyendecker O*, et al.* Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* 2005; **191(9)**:1403-1409.
12. Newell ML. Mechanisms and timing of mother-to-child transmission of HIV-1. *AIDS* 1998; **12(8)**:831-837.

13. Donegan E, Stuart M, Niland JC, Sacks HS, Azen SP, Dietrich SL*, et al.* Infection with human immunodeficiency virus type 1 (HIV-1) among recipients of antibody-positive blood donations. *Ann Intern Med* 1990; **113(10)**:733-739.
14. Franceschi S, Dal ML, La VC. Trends in incidence of AIDS associated with transfusion of blood and blood products in Europe and the United States, 1985-93. *BMJ* 1995; **311(7019)**:1534-1536.
15. Rosenberg PS, Goedert JJ. Estimating the cumulative incidence of HIV infection among persons with haemophilia in the United States of America. *Stat Med* 1998; **17(2)**:155-168.
16. Savarit D, De Cock KM, Schutz R, Konate S, Lackritz E, Bondurand A. Risk of HIV infection from transfusion with blood negative for HIV antibody in a west African city. *BMJ* 1992; **305(6852)**:498-502.
17. Centre for Disease Control. HIV/AIDS Surveillance Report-cases reported through December 2001. In: 2002.
18. Pantaleo G, Graziosi C, Fauci AS. New concepts in the immunopathogenesis of human immunodeficiency virus infection. *N Engl J Med* 1993; **328(5)**:327-335.
19. Pilcher CD, Joaki G, Hoffman IF, Martinson FE, Mapanje C, Stewart PW*, et al.* Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. *AIDS* 2007; **21(13)**:1723-1730.
20. Morgan D, Mahe C, Mayanja B, Okongo JM, Lubega R, Whitworth JA. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS* 2002; **16(4)**:597-603.
21. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* 2007; **17(8)**:1186-1194.
22. Winslow BJ, Pomerantz RJ, Bagasra O, Trono D. HIV-1 latency due to the site of proviral integration. *Virology* 1993; **196(2)**:849-854.
23. Roe B, Hall WW. Cellular and molecular interactions in coinfection with hepatitis C virus and human immunodeficiency virus. *Expert Rev Mol Med* 2008; **10**:e30.
24. Preston BD, Poiesz BJ, Loeb LA. Fidelity of HIV-1 reverse transcriptase. *Science* 1988; **242(4882)**:1168-1171.
25. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P*, et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 1995; **373(6510)**:117-122.
26. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nat Rev Genet* 2004; **5(1)**:52-61.
27. Gojobori T, Moriyama EN, Ina Y, Ikeo K, Miura T, Tsujimoto H*, et al.* Evolutionary origin of human and simian immunodeficiency viruses. *Proc Natl Acad Sci U S A* 1990; **87(11)**:4108-4111.
28. De Clercq E. The design of drugs for HIV and HCV. *Nat Rev Drug Discov* 2007; **6(12)**:1001-1018.
29. WHO HIV/AIDS Programme. Antiretroviral Therapy for HIV Infection in Adults and Adolescents - Recommendations for a public health approach. In: 2010.
30. Kitahata MM, Gange SJ, Abraham AG, Merriman B, Saag MS, Justice AC*, et al.* Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med* 2009; **360(18)**:1815-1826.
31. Jin F, Jansson J, Law M, Prestage GP, Zablotska I, Imrie JC*, et al.* Per-contact probability of HIV transmission in homosexual men in Sydney in the era of HAART. *AIDS* 2010; **24(6)**:907-913.

32. Fisher M, Pao D, Brown AE, Sudarshi D, Gill ON, Cane P*, et al.* Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *AIDS* 2010; **24(11)**:1739-1747.
33. Weller S, Davis K. Condom effectiveness in reducing heterosexual HIV transmission. *Cochrane Database Syst Rev* 2001**(3)**:CD003255.
34. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, Puren A. Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: the ANRS 1265 Trial. *PLoS Med* 2005; **2(11)**:e298.
35. Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, Nalugoda F*, et al.* Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. *Lancet* 2007; **369(9562)**:657-666.
36. Williams BG, Lloyd-Smith JO, Gouws E, Hankins C, Getz WM, Hargrove J*, et al.* The potential impact of male circumcision on HIV in Sub-Saharan Africa. *PLoS Med* 2006; **3(7)**:e262.
37. Bailey RC, Moses S, Parker CB, Agot K, Maclean I, Krieger JN*, et al.* Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. *Lancet* 2007; **369(9562)**:643-656.
38. Sanchez J, Sal y Rosas V, Hughes J, Baeten J, Fuchs JD, Buchbinder SP*, et al.* Male circumcision and risk of HIV acquisition among MSM. *AIDS* 2011; **25(4)**:519-523.
39. Guay LA, Musoke P, Fleming T, Bagenda D, Allen M, Nakabiito C*, et al.* Intrapartum and neonatal single-dose nevirapine compared with zidovudine for prevention of mother-to-child transmission of HIV-1 in Kampala, Uganda: HIVNET 012 randomised trial. *Lancet* 1999; **354(9181)**:795-802.
40. Coovadia H. Antiretroviral agents--how best to protect infants from HIV and save their mothers from AIDS. *N Engl J Med* 2004; **351(3)**:289-292.
41. Hurley SF, Jolley DJ, Kaldor JM. Effectiveness of needle-exchange programmes for prevention of HIV infection. *Lancet* 1997; **349(9068)**:1797-1800.
42. Gibson DR, Flynn NM, Perales D. Effectiveness of syringe exchange programs in reducing HIV risk behavior and HIV seroconversion among injecting drug users. *AIDS* 2001; **15(11)**:1329-1341.
43. 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; **363(27)**:2587-2599.
44. Abdool KQ, Abdool Karim SS, Frohlich JA, Grobler AC, Baxter C, Mansoor LE*, et al.* Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science* 2010; **329(5996)**:1168-1174.
45. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R*, et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N Engl J Med* 2009; **361(23)**:2209-2220.
46. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J*, et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 1983; **220(4599)**:868-871.
47. Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF*, et al.* Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* 1984; **224(4648)**:500-503.
48. Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N*, et al.* Human immunodeficiency viruses. *Science* 1986; **232(4751)**:697.

49. Daniel MD, Letvin NL, King NW, Kannagi M, Sehgal PK, Hunt RD*, et al.* Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. *Science* 1985; **228(4704)**:1201-1204.
50. Clavel F, Guetard D, Brun-Vezinet F, Chamaret S, Rey MA, Santos-Ferreira MO*, et al.* Isolation of a new human retrovirus from West African patients with AIDS. *Science* 1986; **233(4761)**:343-346.
51. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* 1989; **339(6223)**:389-392.
52. 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; **288(5472)**:1789-1796.
53. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. Phylogeny and the origin of HIV-1. *Nature* 2001; **410(6832)**:1047-1048.
54. Peeters M, Chaix ML, Delaporte E. Genetic diversity and phylogeographic distribution of SIV: how to understand the origin of HIV. *Med Sci (Paris)* 2008; **24(6-7)**:621-628.
55. Zhu T, 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; **391(6667)**:594-597.
56. 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)**:661-664.
57. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF*, et al.* Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 1999; **397(6718)**:436-441.
58. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. *Science* 2000; **287(5453)**:607-614.
59. Keele BF, Van HF, Li Y, Bailes E, Takehisa J, Santiago ML*, et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006; **313(5786)**:523-526.
60. Leitner T, Korber B, Daniels M, Calef C, Foley B. Subtype and Circulating Recombinant Form (CRF) Reference Sequences. In. Hahn BH, Marx P, McCutchan FE, Mellors JW, Wolinsky S (editors). Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos Laboratory; 2005.
61. Kuiken C, Foley B, Hahn BH, Marx P, McCutchan FE, Mellors JW*, et al.* A compilation and analysis of nucleic acid and amino acid sequences. In. Theoretical Biology and Biophysics Group LANL (editor). Los Alamos, New Mexico; 1999.
62. Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *J Acquir Immune Defic Syndr* 2002; **29(2)**:184-190.
63. Arien KK, Vanham G, Arts EJ. Is HIV-1 evolving to a less virulent form in humans? *Nat Rev Microbiol* 2007; **5(2)**:141-151.
64. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. *N Engl J Med* 2008; **358(15)**:1590-1602.
65. Esbjornsson J, Mild M, Mansson F, Norrgren H, Medstrand P. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. *PLoS One* 2011; **6(2)**:e17025.
66. Renjifo B, Fawzi W, Mwakagile D, Hunter D, Msamanga G, Spiegelman D*, et al.* Differences in perinatal transmission among human immunodeficiency virus type 1 genotypes. *J Hum Virol* 2001; **4(1)**:16-25.
67. Vasan A, Renjifo B, Hertzmark E, Chaplin B, Msamanga G, Essex M*, et al.* Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. *Clin Infect Dis* 2006; **42(6)**:843-852.

68. Gilbert PB, McKeague IW, Eisen G, Mullins C, Gueye-NDiaye A, Mboup S*, et al.* Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. *Stat Med* 2003; **22(4)**:573-593.
69. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H*, et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 2011.
70. 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 U S A* 2008; **105(21)**:7552-7557.
71. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA*, et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993; **261(5125)**:1179-1181.
72. Kouyos RD, von W, V, Yerly S, Boni J, Rieder P, Joos B*, et al.* Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis* 2011; **52(4)**:532-539.
73. Meier UC, Klenerman P, Griffin P, James W, Koppe B, Larder B*, et al.* Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants. *Science* 1995; **270(5240)**:1360-1362.
74. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO*, et al.* Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* 1991; **354(6353)**:453-459.
75. Price DA, Goulder PJ, Klenerman P, Sewell AK, Easterbrook PJ, Troop M*, et al.* Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc Natl Acad Sci U S A* 1997; **94(5)**:1890-1895.
76. Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol* 2003; **20(8)**:1318-1325.
77. Brumme ZL, John M, Carlson JM, Brumme CJ, Chan D, Brockman MA*, et al.* HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins. *PLoS One* 2009; **4(8)**:e6687.
78. Moore CB, John M, James IR, Christiansen FT, Witt CS, Mallal SA. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science* 2002; **296(5572)**:1439-1443.
79. Cao K, Hollenbach J, Shi X, Shi W, Chopek M, Fernandez-Vina MA. Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum Immunol* 2001; **62(9)**:1009-1030.
80. Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, Addo M*, et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature* 2009; **458(7238)**:641-645.
81. Matthews PC, Leslie AJ, Katzourakis A, Crawford H, Payne R, Prendergast A*, et al.* HLA footprints on human immunodeficiency virus type 1 are associated with interclade polymorphisms and intraclade phylogenetic clustering. *J Virol* 2009; **83(9)**:4605-4615.
82. Kaslow RA, Carrington M, Apple R, Park L, Munoz A, Saah AJ*, et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* 1996; **2(4)**:405-411.
83. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI, Walker BD*, et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 2010; **330(6010)**:1551-1557.

84. Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E*, et al.* Advantage of rare HLA supertype in HIV disease progression. *Nat Med* 2003; **9(7)**:928-935.
85. Frater AJ, Beardall A, Ariyoshi K, Churchill D, Galpin S, Clarke JR*, et al.* Impact of baseline polymorphisms in RT and protease on outcome of highly active antiretroviral therapy in HIV-1-infected African patients. *AIDS* 2001; **15(12)**:1493-1502.
86. 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.
87. Grant RM, Hecht FM, Warmerdam M, Liu L, Liegler T, Petropoulos CJ*, et al.* Time trends in primary HIV-1 drug resistance among recently infected persons. *JAMA* 2002; **288(2)**:181-188.
88. Little SJ. Is transmitted drug resistance in HIV on the rise? It seems so. *BMJ* 2001; **322(7294)**:1074-1075.
89. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C*, et al.* Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. *PLoS Med* 2008; **5(7)**:e158.
90. Little SJ, Frost SD, Wong JK, Smith DM, Pond SL, Ignacio CC*, et al.* Persistence of transmitted drug resistance among subjects with primary human immunodeficiency virus infection. *J Virol* 2008; **82(11)**:5510-5518.
91. Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. The EuroGuidelines Group for HIV resistance. *AIDS* 2001; **15(3)**:309-320.
92. Gazzard BG. British HIV Association guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008. *HIV Med* 2008; **9(8)**:563-608.
93. Hirsch MS, Gunthard HF, Schapiro JM, Brun-Vezinet F, Clotet B, Hammer SM*, et al.* Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Clin Infect Dis* 2008; **47(2)**:266-285.
94. Pillay D. Current patterns in the epidemiology of primary HIV drug resistance in North America and Europe. *Antivir Ther* 2004; **9(5)**:695-702.
95. Shafer RW, Rhee SY, Pillay D, Miller V, Sandstrom P, Schapiro JM*, et al.* HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS* 2007; **21(2)**:215-223.
96. 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. *PLoS ONE* 2009; **4(3)**:e4724.
97. Gifford RJ, Liu TF, Rhee SY, 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)**:1197-1198.
98. Abegaz WE, Grossman Z, Wolday D, Ram D, Kaplan J, Sibide K*, et al.* Threshold survey evaluating transmitted HIV drug resistance among public antenatal clinic clients in Addis Ababa, Ethiopia. *Antivir Ther* 2008; **13 Suppl 2**:89-94.
99. Nguyen HT, Duc NB, Shrivastava R, Tran TH, Nguyen TA, Thang PH*, et al.* HIV drug resistance threshold survey using specimens from voluntary counselling and testing sites in Hanoi, Vietnam. *Antivir Ther* 2008; **13 Suppl 2**:115-121.
100. Jayaraman GC, Gleeson T, Rekart ML, Cook D, Preiksaitis J, Sidaway F*, et al.* Prevalence and determinants of HIV-1 subtypes in Canada: enhancing routinely collected information

through the Canadian HIV Strain and Drug Resistance Surveillance Program. *Can Commun Dis Rep* 2003; **29(4)**:29-36.
101. Public Health Agency of Canada. HIV and AIDS in Canada. Surveillance Report to December 31, 2009. In. Surveillance and Risk Assessment Division CfCDaIC (editor); 2010.
102. Korber BT, Foley BT, Kuiken CL, Pillai SK, Sodroski JG. Numbering positions in HIV Relative to HXB2CG. In. Hahn BH, McCutchan FE, Mellors JW, Sodroski JG (editors): Theoretical Biology and Biophysics Group, Los Alamos Laboratory, Los Alamos, NM; 1998. pp. 102-111.
103. Parekh BS, Kennedy MS, Dobbs T, Pau CP, Byers R, Green T*, et al.* Quantitative detection of increasing HIV type 1 antibodies after seroconversion: a simple assay for detecting recent HIV infection and estimating incidence. *AIDS Res Hum Retroviruses* 2002; **18(4)**:295-307.
104. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, Kosakovsky Pond SL. HIV-specific probabilistic models of protein evolution. *PLoS One* 2007; **2(6)**:e503.
105. Leitner T, Albert J. Reconstruction of HIV-1 transmission chains for forensic purposes. *AIDS Rev* 2000; **2**:241-251.
106. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT*, et al.* Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992; **256(5060)**:1165-1171.
107. Pistello M, Del SB, Butto S, Bargagna M, Domenici R, Bendinelli M. Genetic and phylogenetic analyses of HIV-1 corroborate the transmission link hypothesis. *J Clin Virol* 2004; **30(1)**:11-18.
108. Sturmer M, Preiser W, Gute P, Nisius G, Doerr HW. Phylogenetic analysis of HIV-1 transmission: *pol* gene sequences are insufficient to clarify true relationships between patient isolates. *AIDS* 2004; **18(16)**:2109-2113.
109. Hue 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)**:719-728.
110. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C*, et al.* High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis* 2007; **195(7)**:951-959.
111. Gifford RJ, de OT, 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; **81(23)**:13050-13056.
112. Pao D, Fisher M, Hue S, Dean G, Murphy G, Cane PA*, et al.* Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS* 2005; **19(1)**:85-90.
113. Guimaraes ML, Vicente AC, Otsuki K, da Silva RF, Francisco M, da Silva FG*, et al.* Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. *Retrovirology* 2009; **6**:39.
114. Bello G, Eyer-Silva WA, Couto-Fernandez JC, Guimaraes ML, Chequer-Fernandez SL, Teixeira SL*, et al.* Demographic history of HIV-1 subtypes B and F in Brazil. *Infect Genet Evol* 2007; **7(2)**:263-270.
115. Bello G, Guimaraes ML, Passaes CP, Matos Almeida SE, Veloso VG, Morgado MG. Short communication: Evidences of recent decline in the expansion rate of the HIV type 1 subtype C and CRF31_BC epidemics in southern Brazil. *AIDS Res Hum Retroviruses* 2009; **25(11)**:1065-1069.

116. Walker PR, Pybus OG, Rambaut A, Holmes EC. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol* 2005; **5(3)**:199-208.
117. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA*, et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004; **303(5656)**:327-332.
118. Fitch WM. Toward defining the course of evolution: minimum change of specified tree topology. *Syst Zool* 1971; **20**:406-416.
119. Felsenstein J. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Syst Biol* 1978; **27(4)**:401-410.
120. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4(4)**:406-425.
121. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc Natl Acad Sci U S A* 2004; **101(30)**:11030-11035.
122. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci U S A* 1996; **93(20)**:10864-10869.
123. Rannala B, Yang Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol* 1996; **43(3)**:304-311.
124. Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* 1997; **14(7)**:717-724.
125. Zuckerkandl E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in Biochemistry*. Kasha M, Pullman B (editors). New York: Academic Press; 1962. pp. 189-225.
126. Rambaut A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 2000; **16(4)**:395-399.
127. Posada D, Crandall KA. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* 2001; **18(6)**:897-906.
128. Sanderson MJ. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* 1997; **14**:1218-1231.
129. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 2006; **4(5)**:e88.
130. Kingman JFC. The coalescent. *Stochastic Processes and their Applications* 1982; **13**:235-248.
131. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005; **22(5)**:1185-1192.
132. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007; **7**:214.
133. Lartillot N, Philippe H. Computing Bayes factors using thermodynamic integration. *Syst Biol* 2006; **55(2)**:195-207.
134. Avise JC, Arnold J, Ball JrE, Bermingham T, Lamb JE, Neigel JE. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* 1987; **18**:489-522.
135. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 1986; **3(5)**:418-426.
136. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 1994; **11(5)**:725-736.

137. Tully DC, Wood C. Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *AIDS* 2010; **24(10)**:1577-1582.
138. Salemi M, de OT, Ciccozzi M, Rezza G, Goodenow MM. High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. *PLoS One* 2008; **3(1)**:e1390.
139. 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)**:427-429.
140. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog* 2009; **5(9)**:e1000590.
141. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med* 2008; **5(3)**:e50.
142. Archer J, Robertson DL. Understanding the diversification of HIV-1 groups M and O. *AIDS* 2007; **21(13)**:1693-1700.
143. Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. The emergence of HIV/AIDS in the Americas and beyond. *Proc Natl Acad Sci U S A* 2007; **104(47)**:18566-18570.
144. Liao H, Tee KK, Hase S, Uenishi R, Li XJ, Kusagawa S*, et al.* Phylodynamic analysis of the dissemination of HIV-1 CRF01_AE in Vietnam. *Virology* 2009; **391(1)**:51-56.
145. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. *Science* 2001; **292(5525)**:2323-2325.
146. Robertson DL, Anderson JP, Bradac JK, Carr JK, Foley B, Gao F. HIV-1 Nomenclature Poposal. In: *Human Retroviruses and AIDS*. Kuiken C, McCutchan FE, Marx P, Mellors JW, Mullins JI, Wolinsky S (editors). Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory; 1999. pp. 492-505.
147. Smith DM, May SJ, Tweeten S, Drumright L, Pacold ME, Kosakovsky Pond SL*, et al.* A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS* 2009; **23(2)**:225-232.
148. Resik S, Lemey P, Ping LH, 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; **23(3)**:347-356.
149. 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. *J Acquir Immune Defic Syndr* 2010; **55(1)**:102-108.
150. 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.
151. DeGruttola V, Smith DM, Little SJ, Miller V. Developing and evaluating comprehensive HIV infection control strategies: issues and challenges. *Clin Infect Dis* 2010; **50 Suppl 3**:S102-S107.
152. Sutton MY, Liu H, Steiner B, Pillay A, Mickey T, Finelli L*, et al.* Molecular subtyping of Treponema pallidum in an Arizona County with increasing syphilis morbidity: use of specimens from ulcers and blood. *J Infect Dis* 2001; **183(11)**:1601-1606.
153. van Deutekom H., Gerritsen JJ, van SD, van Ameijden EJ, van Embden JD, Coutinho RA. A molecular epidemiological approach to studying the transmission of tuberculosis in Amsterdam. *Clin Infect Dis* 1997; **25(5)**:1071-1077.

154. Salamon H, Behr MA, Rhee JT, Small PM. Genetic distances for the study of infectious disease epidemiology. *Am J Epidemiol* 2000; **151(3)**:324-334.
155. Hecht FM, Wolf LE, Lo B. Lessons from an HIV transmission pair. *J Infect Dis* 2007; **195(9)**:1239-1241.
156. de Oliveira T., Pybus OG, Rambaut A, Salemi M, Cassol S, Ciccozzi M*, et al.* Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* 2006; **444(7121)**:836-837.
157. Bernard EJ, Azad Y, Vandamme AM, 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)**:382-387.
158. 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 U S A* 2010; **107(50)**:21242-21247.
159. Cohen MS, Gay CL, Busch MP, Hecht FM. The detection of acute HIV infection. *J Infect Dis* 2010; **202 Suppl 2**:S270-S277.
160. Murphy G, Parry JV. Assays for the detection of recent infections with human immunodeficiency virus type 1. *Euro Surveill* 2008; **13(36)**.
161. Fiscus SA, Pilcher CD, Miller WC, Powers KA, Hoffman IF, Price M*, et al.* Rapid, real-time detection of acute HIV infection in patients in Africa. *J Infect Dis* 2007; **195(3)**:416-424.
162. Pilcher CD, Fiscus SA, Nguyen TQ, Foust E, Wolf L, Williams D*, et al.* Detection of acute infections during HIV testing in North Carolina. *N Engl J Med* 2005; **352(18)**:1873-1883.
163. Fisher M, Pao D, Murphy G, Dean G, McElborough D, Homer G*, et al.* Serological testing algorithm shows rising HIV incidence in a UK cohort of men who have sex with men: 10 years application. *AIDS* 2007; **21(17)**:2309-2314.
164. UNAIDS. UNAIDS Reference Group on estimates, modelling and projections-statement on the use of the BED assay for the estimation of HIV-1 incidence for surveillance or epidemic monitoring. *Wkly Epidemiol Rec* 2006; **81(4)**:40.
165. Hargrove JW, Humphrey JH, Mutasa K, Parekh BS, McDougal JS, Ntozini R*, et al.* Improved HIV-1 incidence estimates using the BED capture enzyme immunoassay. *AIDS* 2008; **22(4)**:511-518.
166. Parekh BS, Hanson DL, Hargrove J, Branson B, Green T, Dobbs T*, et al.* Determination of mean recency period for estimation of HIV type 1 incidence with the BED-capture EIA in persons infected with diverse subtypes. *AIDS Res Hum Retroviruses* 2011; **27(3)**:265-273.
167. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H*, et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 1999; **73(12)**:10489-10502.
168. Troyer RM, Collins KR, Abraha A, Fraundorf E, Moore DM, Krizan RW*, et al.* Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression. *J Virol* 2005; **79(14)**:9006-9018.
169. Jayaraman GC, Archibald CP, Lior L, Sutherland D. Integrating laboratory and epidemiological techniques for population-based surveillance of HIV strains and drug resistance in Canada. *Can J Infect Dis* 2000; **11(2)**:74-80.
170. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 1999; **41**:95-98.

171. 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; **21(19)**:3797-3800.
172. Applied Biosystems. SeqScape. In: 2009.
173. Xia X, Xie Z. DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* 2001; **92(4)**:371-373.
174. Kosakovsky Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 2005; **21(5)**:676-679.
175. IBM. SPSS for Windows. In. Chicago: SPSS Inc.; 2009.
176. Liu Y, McNevin JP, Holte S, McElrath MJ, Mullins JI. Dynamics of viral evolution and CTL responses in HIV-1 infection. *PLoS One* 2011; **6(1)**:e15639.
177. Ambrose J, Foster G, Chaytor S, Booth C, Geretti AM. Population sequence nucleotide ambiguity as a measure of HIV-1 infection length [abstract 1058]. In: *18th Conference on Retroviruses and Opportunistic Infections*; 2011. p. 503.
178. Andersson E, Shao W, Bontell I, Bertagnolio S, Cham F, Cuong D*, et al.* Evaluation of a new subtype independent bioinformatics algorithm to detect recent HIV-1 infection [abstract 1056]. In: *18th Conference on Retroviruses and Opportunistic Infections*; 2011. p. 502.
179. Wilson E, Shao W, Brooks J, Dewar R, Kearney M, Rehm C*, et al.* New Bioinformatic Algorithm to Identify Recent HIV-1 Infection [abstract 1057]. In: *18th Conference on Retroviruses and Opportunistic Infections*; 2011. p. 503.
180. Boulos D, Yan P, Schanzer D, Remis R, Archibald C. Estimates of HIV Prevalence and Incidence in Canada, 2005. In: 2006.
181. 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; **199(3)**:427-431.
182. Brooks J, Woods C, Merks H, Wynhoven B, Hall TA, Sandstrom P. Evaluation of an automated sequence analysis tool to standardize HIV genotyping results. In: 2009.
183. Niccolai LM, Verevochkin SV, Toussova OV, White E, Barbour R, Kozlov AP*, et al.* Estimates of HIV incidence among drug users in St. Petersburg, Russia: continued growth of a rapidly expanding epidemic. *Eur J Public Health* 2010.
184. Le Vu S, Meyer L, Cazein F, Pillonel J, Semaille C, Barin F*, et al.* Performance of an immunoassay at detecting recent infection among reported HIV diagnoses. *AIDS* 2009; **23(13)**:1773-1779.
185. Truong HM, Kellogg T, Louie B, Klausner J, Dilley J, McFarland W. Recent HIV-1 infection detection: comparison of incidence estimates derived by laboratory assays and repeat testing data. *J Acquir Immune Defic Syndr* 2009; **51(4)**:502-505.
186. Hahn BH, Shaw GM, Taylor ME, Redfield RR, Markham PD, Salahuddin SZ*, et al.* Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. *Science* 1986; **232(4757)**:1548-1553.
187. Brumme ZL, Brumme CJ, Heckerman D, Korber BT, Daniels M, Carlson J*, et al.* Evidence of differential HLA class I-mediated viral evolution in functional and accessory/regulatory genes of HIV-1. *PLoS Pathog* 2007; **3(7)**:e94.
188. John M, Heckerman D, James I, Park LP, Carlson JM, Chopra A*, et al.* Adaptive interactions between HLA and HIV-1: highly divergent selection imposed by HLA class I molecules with common supertype motifs. *J Immunol* 2010; **184(8)**:4368-4377.

189. Avila-Rios S, Ormsby CE, Carlson JM, Valenzuela-Ponce H, Blanco-Heredia J, Garrido-Rodriguez D*, et al.* Unique features of HLA-mediated HIV evolution in a Mexican cohort: a comparative study. *Retrovirology* 2009; **6**:72.
190. Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, Kadie C*, et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* 2007; **315(5818)**:1583-1586.
191. Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, Kadie C*, et al.* Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput Biol* 2008; **4(11)**:e1000225.
192. Public Health Agency Canada. HIV/ AIDS Epi Upates - July 2010. In: 2010.
193. Statscan. Canada Census 2006. In: 2007.
194. Perez-Sweeney B, DeSalle R, Ho JL. An introduction to a novel population genetic approach for HIV characterization. *Infect Genet Evol* 2010; **10(8)**:1155-1164.
195. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5(3)**:e9490.
196. Posada D. jModelTest: phylogenetic model averaging. *Mol Biol Evol* 2008; **25(7)**:1253-1256.
197. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004; **20(2)**:289-290.
198. R Development Core Team. R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing; 2008.
199. Parker J, Rambaut A, Pybus O. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect Genet Evol* 2008; **8**:239-246.
200. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131(2)**:479-491.
201. Hartl D, Clark A. Inbreeding, Population Subdivision and Migration. In: *Principles of Population Genetics*. Sinauers Associates (editor). Sunderland; 2007. pp. 288-293.
202. Craib KJ, Spittal PM, Wood E, Laliberte N, Hogg RS, Li K*, et al.* Risk factors for elevated HIV incidence among Aboriginal injection drug users in Vancouver. *CMAJ* 2003; **168(1)**:19-24.
203. Spittal PM, Craib KJ, Teegee M, Baylis C, Christian WM, Moniruzzaman AK*, et al.* The Cedar project: prevalence and correlates of HIV infection among young Aboriginal people who use drugs in two Canadian cities. *Int J Circumpolar Health* 2007; **66(3)**:226-240.
204. Keynan Y, Chee-Loong S, Bresler K, Pindera C, Becker M, Kasper K. The influence of HLA haplotype frequency on disease progression among HIV infected individuals in the province of Manitoba. In: 2010.
205. Itescu S, Mathur-Wagh U, Skovron ML, Brancato LJ, Marmor M, Zeleniuch-Jacquotte A*, et al.* HLA-B35 is associated with accelerated progression to AIDS. *J Acquir Immune Defic Syndr* 1992; **5(1)**:37-45.
206. Arnaiz-Villena A, Martin-Villa JM, Amador JT, Cendoya-Matamoros A, Tome MI, Rivera JM*, et al.* Risk of vertical HIV transmission combines the 'B35-Cw4 disadvantage' and the 'pattern of inheritance' theories of progression. *Curr HIV Res* 2009; **7(3)**:314-319.
207. Navis M, Schellens I, van BD, Borghans J, van SP, Miedema F*, et al.* Viral replication capacity as a correlate of HLA B57/B5801-associated nonprogressive HIV-1 infection. *J Immunol* 2007; **179(5)**:3133-3143.

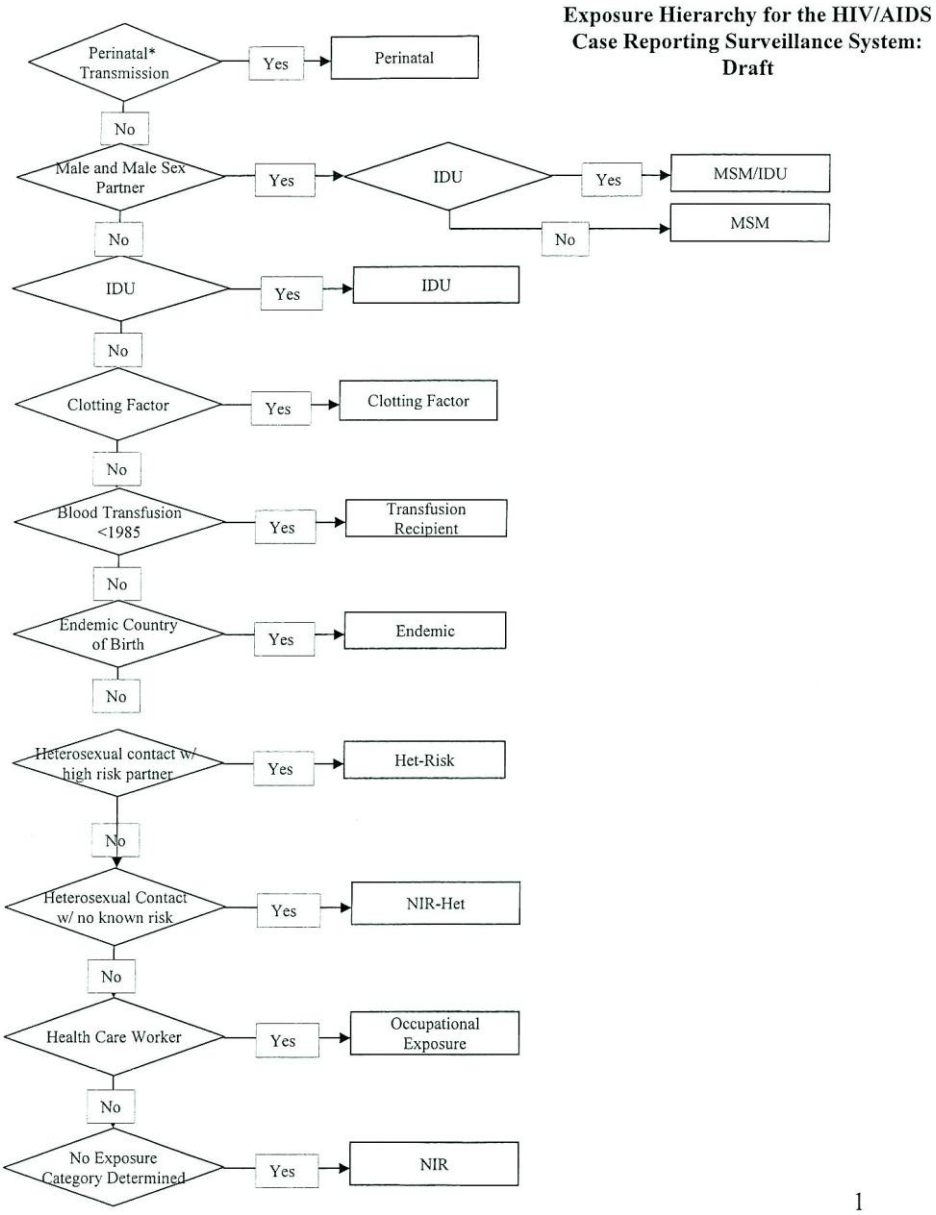
208. Carrington M, O'Brien SJ. The influence of HLA genotype on AIDS. *Annu Rev Med* 2003; **54**:535-551.
209. Pilcher CD, Wong JK, Pillai SK. Inferring HIV transmission dynamics from phylogenetic sequence relationships. *PLoS Med* 2008; **5(3)**:e69.
210. Allen TM, Yu XG, Kalife ET, Reyor LL, Lichterfeld M, John M*, et al.* De novo generation of escape variant-specific CD8+ T-cell responses following cytotoxic T-lymphocyte escape in chronic human immunodeficiency virus type 1 infection. *J Virol* 2005; **79(20)**:12952-12960.
211. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M*, et al.* HIV evolution: CTL escape mutation and reversion after transmission. *Nat Med* 2004; **10(3)**:282-289.
212. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 2010; **38(Web Server issue)**:W7-13.
213. Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* 2005; **22(5)**:1208-1222.
214. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online* 2005; **1**:47-50.
215. 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; **199(3)**:427-431.
216. Lemey P, Rambaut A, Drummond AJ, Suchard M. Bayesian phylogeography finds its roots. *PLoS Comput Biol* 2009; **5(9)**:e1000520.
217. Bielejec F, Rambaut A, Suchard M, Lemey P. SPREAD: Spatial phylogenetic reconstruction of evolutionary dynamics. *(submitted)* 2011.
218. Raphael D. When Social Policy is Health Policy. *Critical Public Health* 2000; **10(2)**.
219. Pearce N, Foliaki S, Sporle A, Cunningham C. Genetics, race, ethnicity, and health. *BMJ* 2004; **328(7447)**:1070-1072.
220. Hankivsky O, Christoffersen A. Intersectionality and the determinants of health: a Canadian perspective. *Critical Public Health* 2008; **18(3)**:271-283.
221. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA*, et al.* Genetic structure of human populations. *Science* 2002; **298(5602)**:2381-2385.
222. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A*, et al.* Genes mirror geography within Europe. *Nature* 2008; **456(7218)**:98-101.
223. Public Health Agency Canada. Canada's Report on HIV/AIDS 2005. In: Surveillance and Risk Assessment Division, Centre for Infectious Disease Prevention and Control; 2005.
224. Hue 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)**:4425-4429.
225. Public Health Agency Canada. HIV-1 Strain and Primary Drug Resistance in Canada: Surveillance Report to March 31, 2005. In: Surveillance and Risk Assessment Division, Centre for Infectious Disease Prevention and Control; 2005.
226. Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM*, et al.* U.S. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J Virol* 2003; **77(11)**:6359-6366.
227. Maljkovic B, I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY*, et al.* Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary

rate of HIV-1 slows down when the epidemic rate increases. *J Virol* 2007; **81(19)**:10625-10635.

228. Millet G, Marks G, Ding H, Jeffries W, Flores S, Murrill C*, et al.* Predictors of being HIV+ unaware among Black and Latino MSM. In: *Conference on Retroviruses and Opportunistic Infections*; 2011.

Appendices

Appendix 1: Hierarchical exposure category



Appendix Figure 1: Hierarchical exposure category. Scanned from working document drafted by PHAC, 2003. IDU intravenous drug user, MSM Man who has sex with men.

Appendix 2: Sequencing of the *pol* region from HIV samples

Nucleic acids are extracted from 200-400µl of serum or plasma using NucliSens easyMAG (bioMerieux, St Laurent, QC) according to manufacturer's instructions. Combinations of primers are used sequentially to amplify and sequence *pol*. Primer combinations are named 'algorithms'.

Algorithm 1: Reverse transcription of the *pol* region and a first round of PCR (RT-PCR) are performed on 10µl of extracted RNA (OneStep RT-PCR kit, Qiagen, Mississauga,ON). RT-PCR is followed by nested PCR amplification (AmpliTaq Gold, Applied Biosystems, California, USA). Samples successfully amplified are sequenced using four primers, generating two overlapping regions covering the PR and RT genes (Appendix Table 1, Appendix Figure 2).

Algorithm 3: If no product is obtained from algorithm 1, the PR and RT genes are amplified separately in two rounds from extracted RNA then sequenced with the nested PCR primers.

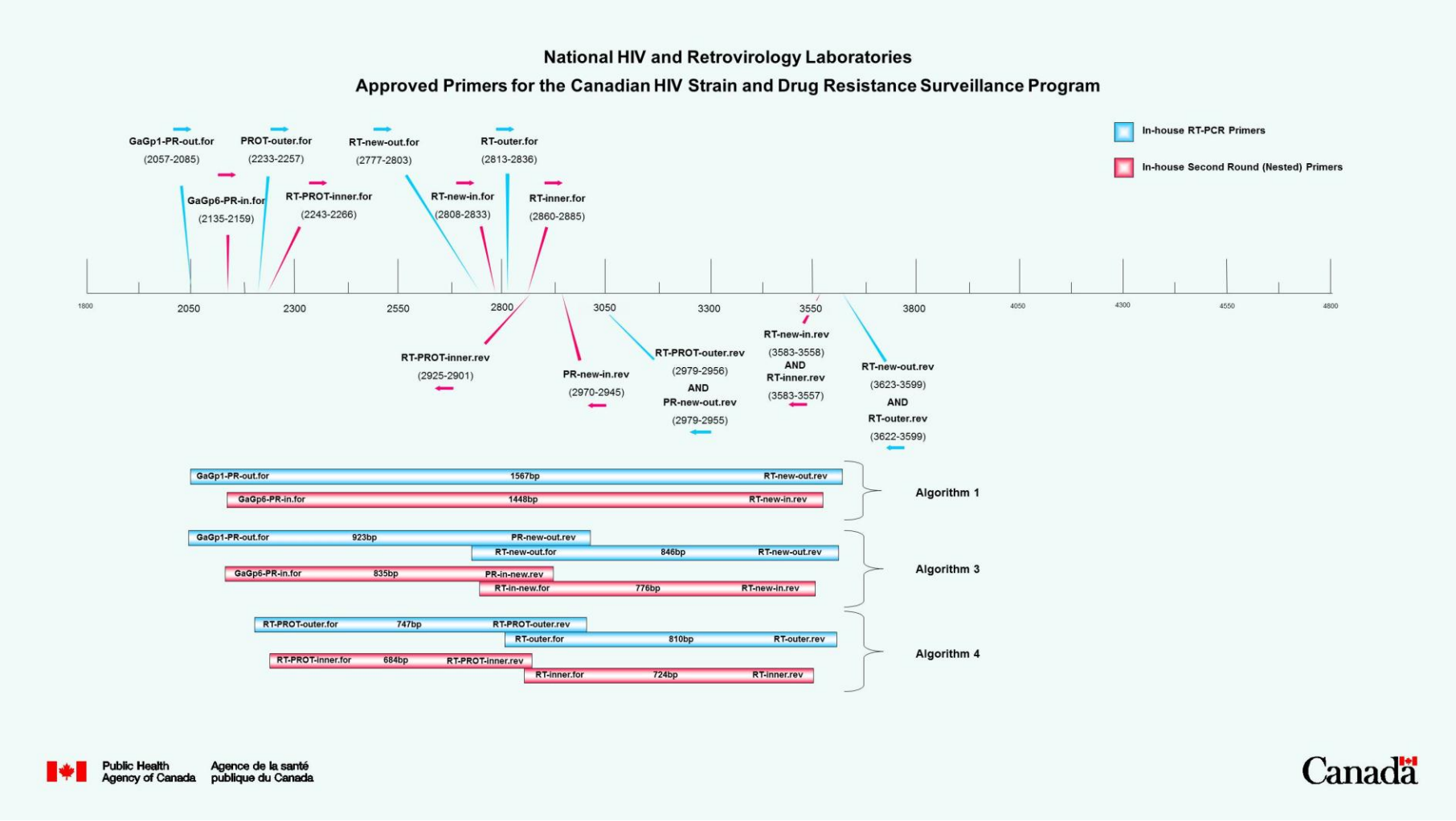
Algorithm 4: If no product is obtained from algorithm 3, algorithm 4 is performed.

All amplifications are carried out at 53°C. PCR products are sequenced using BigDye Terminator V3 on an ABI 3130xl Genetic Analyzer.

Appendix Table 1: Primer sequences and sizes of products for sequencing algorithms.

Algorithm	Region amplified	Sequencing step	Primer	Size of product	Sequences (5' → 3')
Algorithm 1	*pol*	RT-PCR	GaG1-PR-out.for RT-new-out.rev	1567bp	TGA ARG AIT GYA CTG ARA GRC AGG CTA AT CCT CIT TYT TGC ATA YTT YCC TGT T
		Nested PCR	GaGp6-PR-in.for RT-new-in.rev	1448bp	YTC AGA RCA GRC CRG ARC CAA CAG C GGY TCT TGR TAA ATT TGR TAT GTC CA
		Sequencing	GaGp6-PR-in.for PR-new-in.rev RT-new-in.for RT-new-in.rev	1448bp	YTC AGA RCA GRC CRG ARC CAA CAG C CTG GTG TYT CAT TRT TKR TAC TAG GT TTY TGG GAR GTY CAR YTA GGR ATA CC GGY TCT TGR TAA ATT TGR TAT GTC CA
Algorithm 3	*protease*	RT-PCR	GaGp1-PR-out.for PR-new-out.rev	923bp	TGA ARG AIT GYA CTG ARA GRC AGG CTA AT AYC TIA TYC CTG GTG TYT CAT TRT T
	RT		RT-new-out.for RT-new-out.rev	846bp	TTT YAG RGA RCT YAA TAA RAG AAC TCA CCT CIT TYT TGC ATA YTT YCC TGT T
	protease	Nested PCR Sequencing	GaGp6-PR-in.for PR-new-in.rev	835bp	YTC AGA RCA GRC CRG ARC CAA CAG C CTG GTG TYT CAT TRT TKR TAC TAG GT
	RT		RT-new-in.for RT-new-in.rev	776bp	TTY TGG GAR GTY CAR YTA GGR ATA CC GGY TCT TGR TAA ATT TGR TAT GTC CA
Algorithm 4	*protease*	RT-PCR	RT-PROT-outer.for RT-PROT-outer.in	747bp	GAA CTG TAT CCT TTA RCT TCC CTC A ATC TAA TCC CTG GTG TCT CAT TGT
	RT		RT-outer.for RT-outer.in	810bp	GGA AGT TCA ATT AGG AAT ACC ACA CTC ATT CTT GCA TAY TTT CCT GTT
	protease	Nested PCR Sequencing	RT-PROT-inner.for RT-PROT-inner.rev	684bp	CTT TAR CTT CCC TCA GAT CAC TCT TCC TGA AGT CTT YAT CTA AGG GAA C
	RT		RT-inner.for RT-inner.rev	724bp	AAT CAG TAA CAG TAC TGG ATG TGG GT GGC TCT TGA TAA ATT TGA TAT GTC CAT

NOTES– Nucleotide symbols are those designated by IUPAC. Degenerate bases (R, Y) indicate that primers are used as mixtures corresponding to all permutations.



Appendix Figure 2: Annealing sites of primers on the *pol* gene and expected products. Positions are those of the reference strain HXB2 (Genbank #K03455). Forward primers are listed above the map, and reverse primers below. Some primers may be used for amplification and for sequencing. By Natalie Masse and Harriet Merks (PHAC), modified with permission.