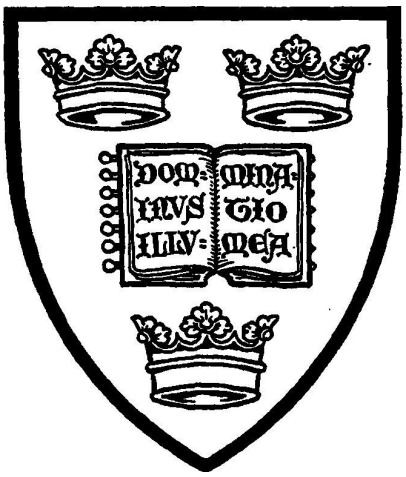
USING MOLECULAR EPIDEMIOLOGY TO INVESTIGATE HIV TRANSMISSION AND PREVENTION EFFORTS



Tetyana Vasylyeva

Hertford College

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

Abstract

Using molecular epidemiology to investigate HIV transmission and prevention efforts

Tetyana Vasylyeva, Hertford College, University of Oxford

Submitted for the degree of Doctor of Philosophy in Zoology, Trinity 2018

The Human Immunodeficiency Virus (HIV) is a retrovirus that causes HIV infection and, if untreated, leads to Acquired Immunodeficiency Syndrome (AIDS) and death. An estimated 36.7 million people in the world are HIV-infected; around 10% of them live in Europe. In this thesis I use molecular epidemiology in combination with mathematical modelling and traditional epidemiological analysis to describe HIV transmission dynamics and to assess various HIV prevention strategies, which might help to inform public health decisions.

I start by characterizing the explosive HIV growth in people who inject drugs (PWID) in Russia and Ukraine in the 1990s in Chapter 2. I chose those settings as a model system in which to study HIV transmission in PWID in the absence of prevention. I then use my estimates to build a mathematical model that can be used to evaluate preventive measures during HIV outbreaks in PWID.

In Chapter 3 I analyse newly-generated HIV-1 subtype A genetic sequences from Ukraine and apply a novel method to show how HIV spread among Ukrainian regions shifted after the initiation of war on the country's eastern borders. I show that, since 2014, HIV is moving westwards, carried by large-scale human migration.

In Chapter 4 I created the largest Ukrainian subtype B (the second most prevalent HIV-1 subtype in Ukraine) dataset to date, and show how the virus distribution changed in the country over the last 15 years. I then simulated phylogenies under scenarios similar to the Ukrainian subtype A and subtype B epidemics, to study the performance of various phylodynamic models that are used to estimate changes in HIV population size and transmission dynamics from virus gene sequences.

Finally, Chapter 5 describes the molecular epidemiological profiles of the two most prevalent HIV-1 subtypes in Portugal, subtypes B and G. I use novel phylodynamic approaches to trace and compare the effect of various public health interventions on the transmission dynamics of these subtypes, given their different transmission group composition.

Acknowledgements

First, I would like to express my deepest gratitude to my supervisors Dr Nuno Faria and Prof Oliver Pybus for their guidance, encouragement, continuous support, and invaluable advice. Dr Faria inspired and motivated me, was always supportive in my endeavours, and helpful in shaping my research ideas. Prof Pybus provided me with ever so wise advice in all the aspects of my academic career, he was always there for discussions, always patient, encouraging, and also challenging.

I know that this work would not have been possible without the support from the Clarendon Scholarship and Hertford College of the University of Oxford, which I greatly appreciate. Prof Martin Maiden, Mr Clive Hambler, and Prof Emma Smith of Hertford College, provided me with guidance and encouragement during some challenging times.

I am very thankful to all the colleagues that I have worked with in these four years. Particularly, to Dr Mariia Liulchuk and Mr Pavlo Smyrnov, who trusted me with their data and made it possible for me to explore the topic I am deeply interested in; to Dr Samuel Friedman, who inspired and motivated me to start my DPhil training; to Dr Louis du Plessis, who patiently helped me with technical aspects of my research and who was always available for discussions; to Prof Sunetra Gupta, Dr Jayna Ragwani, and Dr Jose Lourenco, who have given me practical ideas and valuable feedback.

I want to express a very special thank you to my colleagues Miss Cindy Santander, Miss Audrey Lin, and Mr Charlie Ince, for always motivating me to believe in myself, for being my patient tutors in the field of HIV biology and my faithful proofreaders,

and in general for helping me to improve the standard of my spoken English language (mostly through discussions of international cuisine).

I cannot say enough how grateful I am to my family, my parents and my brother. Even if they never get to read this thesis, it is their support and sacrifice that allowed me to even start writing it. I also thank my friends, my basketball teammates, my flatmates, and other people who were always there to share a desert and to listen to my complaints.

Finally, the last, but in no way the least, I want to state here my sincere gratitude to my partner, Ludovico Lazzeretti, for his enormous moral support and practical advice with some of my work. He was always making me feel better about my work and myself, which enabled me to write this thesis.

Table of Contents

i List of frequently used abbreviations……..9

1 Introduction……10

1.1 HIV nomenclature and geographic distribution…….…..11

1.2 HIV biology……...…13

1.3 HIV epidemiology……....16

1.3.1 Transmission routes and transmission groups…….……17

1.3.2 Stages of infection…….……20

1.3.3 Epidemiological parameters of HIV spread…….24

1.3.4 HIV prevention……..26

1.3.5 The HIV epidemic in Eastern Europe……29

1.4 Molecular epidemiology…….……...32

1.4.1 Phylogeography…….33

1.4.2 Population dynamics models…….34

1.4.2.1 Coalescent-based models……..……..…...35

1.4.2.2 Birth-death models……..…….....36

1.4.3 Evolutionary models……...…38

1.4.4 Applications of molecular epidemiology to HIV……39

1.5 Thesis outline……..40

1.6 References…….....43

2 Reducing HIV infection in people who inject drugs is impossible without targeting recently-infected subjects……..52

2.1 Abstract……...53

2.2 Introduction…….54

2.3 Methods……...55

2.3.1 Molecular evolutionary analysis…….55

2.3.1.1 Nucleotide sequences and initial phylogenetic analysis……..…….....55

2.3.1.2 Phylodynamic analysis……..…..56

2.3.2 Estimates of epidemiological parameters……...58

2.3.3 Compartmental Modelling……...62

2.4 Results……..68

2.4.1 Phylogenetic analysis……...68

2.4.2 Phylodynamics and epidemiologic parameters estimates……70
2.4.3 Modeling the establishment of HIV epidemics in PWID……..72
2.4.4 Sensitivity analyses……...76
2.4.5 Limitations-Assumptions on network effects…….……...78
2.5 Discussion……79
2.6 References……82
3 Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine……86
3.1 Abstract……87
3.2 Significance statement……..88
3.3 Introduction……..89
3.4 Methods……...…….91
3.4.1 Data……..91
3.4.1.1 Sequences sampling and phylogenetic analysis……...91
3.4.1.2 Human migration and behaviours……93
3.4.2 Phylogeographic analysis……94
3.4.2.1 Dating phylogenetics……94
3.4.2.2 Discrete trait analysis ……..95
3.4.2.3 Phylogeographic transition……...96
3.4.3 Spearman correlation analysis……..……98
3.5 Results……..101
3.5.1 HIV-1 dissemination among locations in Ukraine……...101
3.5.2 Temporal change in virus lineage movement after the beginning of the conflict……...107
3.5.3 Correlation between observed virus flow and PWID behaviours…108
3.5.4 Time to failure and drug resistance prevalence……...110
3.5.5 Limitations……....113
3.6 Discussion……114
3.7 Conclusions……...…119
3.8 Ethical approval and acknowledgements……..120
3.9 References……121
4 The changing epidemiological profile of HIV-1 subtype B epidemic in Ukraine……..……..125
4.1 Abstract…….....126

4.2 Introduction……...127
4.3 Methods……128
4.3.1 Data……..128
4.3.1.1 Genetic data……128
4.3.1.2 Alignment……130
4.3.1.3 Epidemiological data……...130
4.3.1.3 Statistical analysis……...130
4.3.2 Phylogenetic analysis……..131
4.3.2.1 Maximum likelihood phylogenies……131
4.3.2.2 Time-calibrated phylogenies……...131
4.3.2.3 Estimates of the effective reproductive number……..132
4.3.3 Simulation study……..133
4.3.3.1 Simulation study rationale and aims……...133
4.3.3.2 Simulation scenarios……...136
4.3.3.3 BEAST analyses……..138
4.3.3.4 Interpreting the simulation study results……139
4.4 Results……142
4.4.1 Data……142
4.4.2 Molecular clock analysis……147
4.4.3 Simulation study……..151
4.4.3.1 Performance of the coalescent models……151
4.4.3.2 Performance of the BDSKY model……..160
4.5 Discussion……...161
4.5.1. Transmission dynamics, and geographical and transmission group composition of HIV-1 subtype B clades circulating in Ukraine…….161
4.5.1. Discussion of the simulation study results……..166
4.6 Ethical approval and acknowledgements……..……..169
4.8 References……...170
5 Tracing the impact of public health interventions on HIV-1 epidemic in Portugal using genetic epidemiology……...174
5.1 Abstract……175
5.2 Introduction……..176
5.3 Methods……177
5.3.1 Genetic data……..177

5.3.2 Initial phylogenetic and molecular clock analyses……...179
5.3.3 Reproductive number estimates……181
5.3.4 Discrete trait and association index analyses……...185
5.4 Results……...186
5.4.1 Genetic sequences……186
5.4.2 Phylogenetic cluster analysis……187
5.4.3 Estimates of the effective reproductive number using genetic data..190
5.4.4 Inferring mixture of risk behavior……194
5.5 Discussion……198
5.6 Acknowledgements……...202
5.7 References……203

6 Conclusions…….....207
6.1 Characterizing the HIV epidemic in Eastern Europe……208
6.2 Using phylodynamics to assess prevention strategies for hard-to-reach populations……...212
6.3 Accurate choice of model parameters for phylodynamic inference…….213
6.4 Future directions……...216
6.4.1 Combining social network approach and phylodynamics…….....216
6.4.2 Further use of molecular epidemiology to improve HIV prevention and treatment outcomes in Ukraine……223
6.5 References……225
7 Appendices……229
Appendix 1……....229
Appendix 2…….....236
Appendix 3……....238
Appendix 4……....247
Appendix 5……....248

List of frequently used abbreviations

HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency Syndrome
PWID	People who inject drugs
MSM	Men who have sex with men
FSW	Female sex workers
MTC	Mother-to-child (transmission)
HET	Heterosexual (transmission)
HCV	Hepatitis C virus
HBV	Hepatitis B virus
STI	Sexually transmitted infection
RT	Reverse transcriptase
VL	Viral load
(HA)ART	(Highly-active) Anti-retroviral treatment
PI	Protease inhibitor
NRTI	Nucleoside-analog reverse transcriptase inhibitor
NNRTI	Non-nucleoside reverse transcriptase inhibitor
TasP	Treatment as Prevention
PrEP	Pre-exposure prophylaxis
CRF	Circulating recombinant form
(T)MRCA	(Time to) Most recent common ancestor
BSP	Bayesian Skyline Plot
BDM	Birth-death model
BDSKY	Birth-death Skyline
BEAST	Bayesian Evolutionary Analysis Sampling Trees
MCMC	Markov Chain Monte Carlo
DTA	Discrete trait analysis
CI	Confidence interval
HPD	Highest Posterior Density
ML	Maximum Likelihood
MLE	Marginal Likelihood Estimator
BF	Bayes Factor
HKY	Hasegawa-Kishino-Yano (model)
GTR	General time reversible (model)
PS	Path-sampling
SS	Stepping-stone sampling
SEP	Syringe exchange program
OST	Opioid substitution treatment
HR	Harm reduction
IDP	Internally displaced people
IBBS	Integrated Bio-Behavioral Survey
WHO	World Health Organization
UNAIDS	The Joint United Nations Programme on HIV and AIDS
MOH	Ministry of Health
NGO	Non-governmental organization
FSU	Former Soviet Union

1 Introduction

Parts of the introduction were published in:

(i) Tetyana I Vasylyeva, Samuel R Friedman, Dimitrios Paraskevis, Gkikas Magiorkinis. Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. Infect Genet Evol 2016,46:248-255.

and

(ii) Tetyana I Vasylyeva, Samuel R Friedman, Gkikas Magiorkinis. Prevention of early HIV transmissions might be more important in emerging or generalizing epidemics. Proc Natl Acad Sci U S A 2015,112:E1515.

Summary of authorship:

(i) TIV wrote the paper and created all figures. SRF and DP helped to conceptualize the paper, and edited the original draft. GM provided supervisory support.

(ii) TIV wrote the editorial comment. SRF and GM edited the original draft.

The human immunodeficiency virus (HIV) pandemic is one of the largest and most severe the world has ever recorded. The virus causes infection that, if untreated, leads to Acquired Immunodeficiency Syndrome (AIDS) and death. AIDS was first observed in the USA in 1981, when numbers of young men who have sex with men (MSM) were diagnosed with opportunistic infections or cancers unusual for this age group[1]. Two years later the viral agent of AIDS was identified by two independent research groups, led by French and American scientists[2,3]. Researchers were later able to show that HIV had been circulating in human population years before it was first detected. Specifically, the origin of HIV has been placed in Kinshasa, currently the capital of the Democratic Republic of Congo, in 1920s and its migration to human populations from chimpanzees is believed to have occurred a few decades earlier in southern Cameroon[4]. It was not until 1960s that HIV started its global voyage following, initially, social, cultural, and transportation network changes in Central Africa. Because of the unique nature of the virus and the pandemic that it ignited, since the early 1980s HIV research has grown into a large field that unites scientists who tackle the epidemiological, immunological, biological, sociological, and ethical issues associated with the disease. These combined efforts have resulted in a much clearer picture about all of the aspects of HIV prevention and treatment, but many questions remain unanswered.

1.1 HIV nomenclature and geographic distribution

There are two main types of HIV, namely HIV-1 and HIV-2[5] (Fig. 1.1). HIV-1 consists of four groups (M, N, O, and P), all of which except M have a very limited geographic distribution, appearing mostly in West-Central Africa or in emigrants from that region[6]. HIV-1 group M, responsible for the global HIV/AIDS pandemic,

has further classified into nine subtypes (A, B, C, D, F, G, H, J, and K) and at least 96 circulating recombinant forms, according to the Los Alamos HIV sequence database[7]. Some HIV-1 subtypes (e.g. A and F) are further classified into sub-subtypes[8]. The difference between different subtypes is established phylogenetically, based either on nucleotide sequences of different genomic regions of the same isolates or on full-length genomes[9]. For the virus subtypes within group M, genetic variability is 25% for the *env* gene and 15% for *gag*.

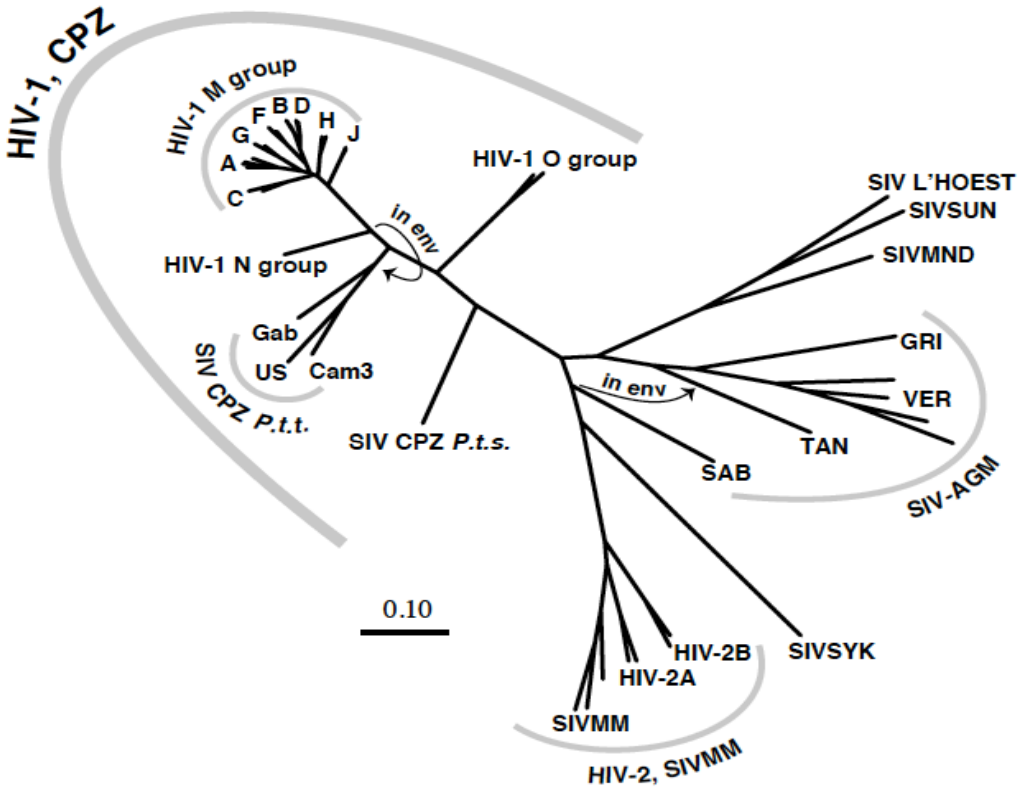


Figure 1.1. Phylogenetic relationships among HIV types, groups, subtypes, and simian immunodeficiency viruses (SIV) from different primate species: CPZ – Chimpanzee, P.t.s - Pan troglodytes shweinfuthii, P.t.t. - Pan troglodytes troglodytes, AGM – African Green Monkeys, MM - Sooty Mangaby, MND – Mandrill, Sun – Sun-tailed monkeys, SYK – Sykes mokeys. The tree was reconstructed using pol sequences.

The bar indicates the equivalent of the genetic distance in substitutions/site/year(from the Los Alamos HIV database[10]; the use is permitted under the LANL copyright).

Subtype C is the most prevalent HIV-1 subtype in the world (≈50% of all existing HIV cases[11]), and is the leading cause of HIV-1 infections in most African and South-American countries, as well as in India[9]. It is followed in relative prevalence by subtype A (≈12% of global HIV cases[11]), which is prevalent in some Central and Eastern African countries[9], as well as in the Former Soviet Union (FSU) countries of Eastern Europe and Central Asia[12], and by subtype B (≈10% of global HIV cases[11]), prevalent mostly in the "Western world" (the USA, Western Europe, Australia, etc)[13].

1.2 HIV biology

HIV belongs to the family of retroviruses, a group of viruses whose genomes are transmitted in an RNA form but are copied into DNA as part of their life-cycle[14]. For this purpose retrovirus genomes encode the enzyme reverse transcriptase (RT). HIV belongs to the genus lentivirus within the family Retroviridae and is transmitted in a single-stranded positive-sense RNA form[5]. The genome structure of HIV is shown in Figure 1.2. The HIV genome is about 9700 nucleotides long and at both 5' and 3' ends has long terminal repeats (LTRs), which are sequences that do not encode for viral proteins. The 5' LTR codes for the promoter for viral genes transcription. Moving from the 5' to the 3' LTR, the HIV genome comprises the *gag*, *pol*, and *env* genes[15]. The *gag* (*group-specific antigen*) gene encodes for proteins that form the inner capsid of the viral particle, the *pol* (*polymerase*) gene encodes for enzymes that are necessary for the replication of the virus (including RT), and the *env* (*envelope*) gene encodes the proteins of the outer envelope of the virus[16]. The accessory genes,

which include *nef*, *vif*, *vpu*, and *vpr*, were previously thought to be dispensable for infection, but are now known to have various functions responsible for virus evasion from adaptive and innate immunity, making them necessary to ensure viral persistence, replication, dissemination, and transmission[17].

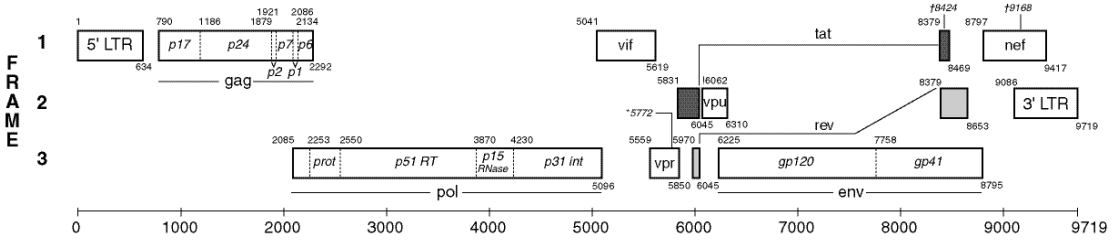


Figure 1.2. HIV-1 genome structure illustrated by HXB2, which is a standard reference HIV genome. Open reading frames are shown as rectangles. Different rows indicate different reading frames and include genes that are translated within that frame. The proteins of the regulator genes (tat and rev) are composed of two gene regions that belong to different reading frames (from the Los Alamos HIV database[10], the use is permitted under the LANL copyright).

Once transmitted, HIV enters the host cell through the binding of the HIV surface glycoprotein gp120, which is a part of the viral envelope, to the host's cellular CD4 receptors (Fig. 1.3)[16]. This opens an additional site for the binding of gp120 to the cell co-receptors, such as chemokine receptor 5 (CCR5) or chemokine receptor 4 (CXCR4 or fusin)[5], and enables the fusion of the viral envelope with the host cell membrane[16]. In the cytoplasm, the HIV genome is transformed from RNA into double-stranded DNA by means of the RT enzyme[18]. The HIV DNA is then transported via nucleopores to the host cell nucleus, where it gets integrated into the host chromosomal DNA through the action of an integrase enzyme. HIV integration creates 'proviruses' within the human genome and leads to the establishment of a

persistent infection[5]. Once integrated, HIV is transcribed and translated to produce viral proteins using host cell enzymes. These proteins are then transported back to the cytoplasm and cell surface, where they assemble and incorporate viral genomic RNA, thereby forming immature virion particles[16]. The maturation of these particles is undertaken by a viral protease enzyme, which cleaves the gag and gag-pol polyprotein into smaller proteins that form mature virions. HIV can infect all CD4+ cells, but it is the invasion and subsequent depletion of CD4+ T-helper lymphocyte cells, which are an important component of the human immune system, that makes HIV-infection so potent and leads, if untreated, to AIDS[19].

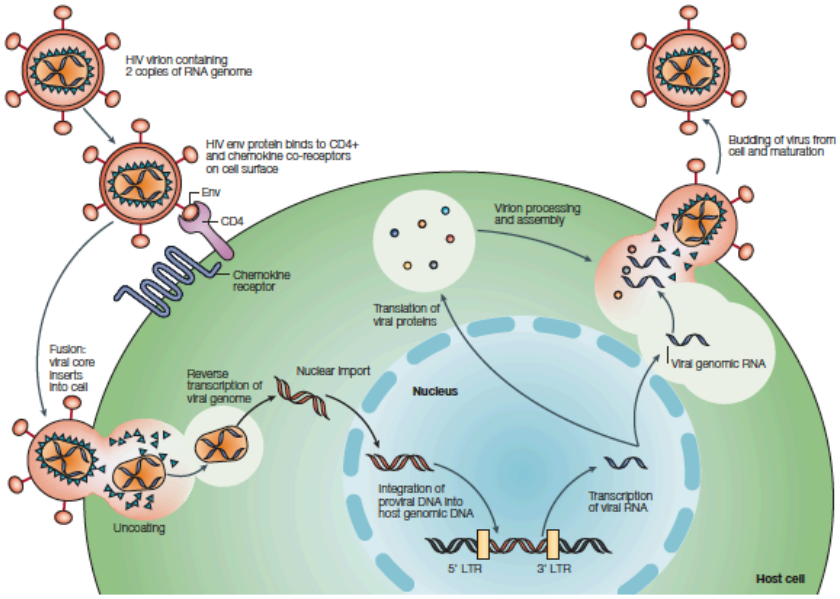


Figure 1.3. An illustration of the HIV life cycle (from Rambaut et al[19]. Licence to reproduce the material granted on December 18th 2018).

There are several reasons why HIV has been so successful in establishing a global pandemic and why it is still one of the major global threats to public health. Firstly, the HIV mutation rate, which is the rate at which mutations occur within the viral genome during genome replication, is extremely high, estimated *in vivo* to be on the

order of magnitude of 10^{-3} to 10^{-5} mutations per base per replication cycle, the highest estimate rate for any biological organism[20-22]. This high mutation rate combines with several other biological factors, such as strong immune selective pressure, high virus population size, and short virus generation time[23,24], to generate high rates of molecular evolutionary change (i.e. the rate at which new mutations are fixed in the virus population). Despite the fact that most mutations are likely to be selectively deleterious, the HIV evolutionary (or substitution) rate is estimated to be between 1.3×10^{-3} substitutions/site/year (s/s/y) (95% Highest Posterior Density (HPD) = 0.7–1.8 $\times10^{-3}$) and 3.5×10^{-3} s/s/y (95% HPD = 2.9–4.2 $\times10^{-3}$), depending on the virus subtype and genomic region[25]. These high rates of evolution make the development of a universal vaccine a moving target, which, despite recent advances in animal models and in human trials, has yet to be achieved[26]. Another important reason for the global dominance of HIV is that, despite the success of antiretroviral treatment (ART), drug treatment still has to be taken for life, hence HIV cannot be fully cured. In each infected person the virus establishes a reservoir of integrated proviruses in resting (or latent) $CD4^+$ T cells, which do not actively produce new copies of the virus, but which have the capacity to recreate infectious virions and re-establish productive infection if drug treatment is interrupted[27]. Lastly, after a brief period of non-specific symptoms during acute infection, HIV enters a long period of asymptomatic chronic infection stage (see section 1.3.2)[5]. Thus, the virus can remain undiagnosed for years, enabling multiple onward transmissions from undiagnosed and untreated individuals.

1.3 HIV epidemiology

The World Health Organisation (WHO) defines epidemiology as "the study of the distribution and determinants of health-related states or events" and how it can be

applied to control and prevent the spread of diseases (http://www.who.int/). Consequently, HIV epidemiology studies how, where and when HIV is transmitted, and how these transmissions can be prevented. HIV epidemiology research focuses on a combination of virus-specific factors, such as the virus life cycle and its interaction with the host immune system, and host-specific factors, such as human population structure and risk behaviours.

1.3.1 Transmission routes and transmission groups

HIV is typically transmitted through exchange of bodily fluids between an infected and uninfected individual and there are consequently three main routes of HIV transmission: sexual, parenteral, and perinatal. According to the route of transmission, epidemiologists define "transmission groups" or "risk groups" – groups of individuals with an elevated risk of HIV acquisition/transmission through one transmission route. Although the definition of transmission groups is sometimes uncertain and the list of groups is not exhaustive, it does allow epidemiologists to develop targeted HIV preventive strategies that aim to halt transmission in the most efficient way.

The majority of HIV-infected people in the world (>80%) acquired the virus through sexual transmission, through exchange of semen, vaginal secretions, and/or blood[28]. Unprotected vaginal intercourse is the most common route of infection[28], which is mostly attributed to sexual transmission of HIV in countries of Sub-Saharan Africa, where for many years people had less access to both information and preventive services. Additionally, due to some cultural norms and beliefs, women in this part of the world can experience difficulties in communicating or ensuring the use of preventive measures, such as condoms, with their partners[29]. MSM form another large sexual transmission group; this group accounts for the majority of new HIV infections in the US and Western Europe. While in most settings transmissions among MSM are

declining, some groups, such as young black MSM in the USA are still disproportionally affected by the virus[30].

The second most common route of HIV transmission is parenteral, i.e. through injection or through the contact with infected blood or blood products. Parenteral transmission is usually enabled through sharing of non-sterile instruments during illicit drug use or is of iatrogenic nature (blood transfusions, organ transplantations, clotting factors, etc). Iatrogenic transmissions were more frequent in the past[31], and even ignited nation-wide epidemics in some countries (e.g. in Romania[32]), and large epidemics in some groups (e.g. haemophiliacs)[33]. However, due to a massive improvement in the management of medical procedures, iatrogenic transmission is now limited to specific vulnerable populations (e.g. children) and to settings with poor medical services, as reviewed in Goldwater[34]. This improvement is mostly attributed to universal medical screening of blood and blood products for presence of HIV, Hepatitis C virus (HCV), Hepatitis B virus (HBV), and *Treponema pallidum* (syphilis), as recommended by WHO in 1991[35]. Since then this recommendation has been adopted globally, but only 58% of the WHO reporting countries perform the screening in a quality-assured manner[36]. For many years transmissions among people who inject drugs (PWID) have been a major driver of HIV epidemics in Central and Eastern European countries, and Central Asia[37]. More recently, transmissions in this group have risen in many Asian countries, the Pacific, and Middle East and the Northern Africa region[38]. Interestingly, in recent years, after years of MSM-driven epidemics, several "Western world" countries have been experiencing a series of HIV outbreaks in PWID. For example, in 2015 outbreaks in Indiana (USA) and Glasgow (UK) were in the spotlight after HIV emerged in historically low-prevalence communities[39,40]. The response to these epidemics was

very different to how emerging PWID-driven epidemics were addressed in the 1990s, since by 2015 a number of evidence-based approaches were available to halt HIV transmission in this group[41]. Specifically, these outbreaks were contained through timely implementation of so-called harm reduction programs. Harm reduction aims to provide means for safe drug injection, such as needle exchange programs or opioid substitution treatment programs (see section 1.3.4 for details). The timely implementation of harm reduction can halt the spread of HIV in PWID populations, as demonstrated by the abovementioned examples, but unfortunately in most settings harm reduction implementation has been significantly delayed, and/or implemented on a smaller scale[42]. This is due to the fact that enabling safe injections is still considered "controversial" and is not accepted in many places where more orthodox opinions, that drug use should not be tolerated, prevail.

Finally, mother-to-child transmissions (MTC) are enabled through exchange of blood and breast milk between an HIV-infected mother and her child during pregnancy, childbirth, and/or breast-feeding. MTC has been mostly eliminated in the developed world, but still poses a great challenge in resource-limited settings, such as Sub-Saharan Africa[43].

While vaginal intercourse is the most common transmission route, the risk of HIV transmission per coital act is low compared to other transmission routes[28]. As such, the probability of transmission is 1 in 200-2000 sexual acts for women and 1 in 700-3000 for men, although receptive anal intercourse brings higher risk of HIV transmission (1 in 20-300) for both genders[28]. The probability of parenteral HIV transmission is even higher at 1 in 2-150[28]. It is important to mention that the probability of transmission is significantly influenced by a number of factors, such as, but not limited to, viral load of a transmitter, the presence of other sexually

transmitted infections (STIs) or male circumcision[44]. Viral load (VL, copies of the virus per millilitre of blood) was found to be the main predictor of the risk of HIV transmission, particularly for a heterosexual unprotected contact[45]. STIs facilitate HIV transmission as they might increase HIV shedding in the genital tract for infected individuals and disrupt mucosal barriers to infection in susceptible recipient individuals[46]. Male circumcision reduces the risk of HIV infection in half for circumcised men compared to those non-circumcised; this protective effect is more apparent in high risk transmission groups where male circumcision cause a four-times reduction in the risk of HIV infection[47]. This effect is explained by the fact that the micro-traumatic lesions or mini-ulcerations of the foreskin can enable an easy HIV entry, and a foreskin has a higher concentration of the primary target cells for HIV transmission.

1.3.2 Stages of infection

One of the most important factors that determine the risk of HIV transmission during a contact between an infected (donor) and an uninfected (recipient) individual is the stage of HIV infection of the donor[48]. It has been shown that HIV infectiousness during the early acute stage of infection (primary infection), which equals approximately to the first three months after virus acquisition, is 26 times higher than during the asymptomatic chronic infection stage (clinical latency), due to the very high VL during acute infection, though estimates vary greatly[49]. The high VL during that phase is explained by rapid rates of HIV replication combined with a delayed immune response (Fig. 1.4)[50], which are associated with a steep reduction in the number of $CD4^+$ T cells. The first 1-2 months after infection are often referred to as the "window period" when an individual is HIV-negative by antibody tests. This is the period before detectable antibodies have been produced and the individual's status

is still "seronegative"; thus, the individual cannot be diagnosed using antibody tests, the virus can only be detected using antigen or HIV RNA tests (10 days after infection, since VL is undetectable in the first 10 days)[50]. Once detectable antibodies against the virus have developed (antibody seroconversion) the individual becomes "seropositive" and VL drops.

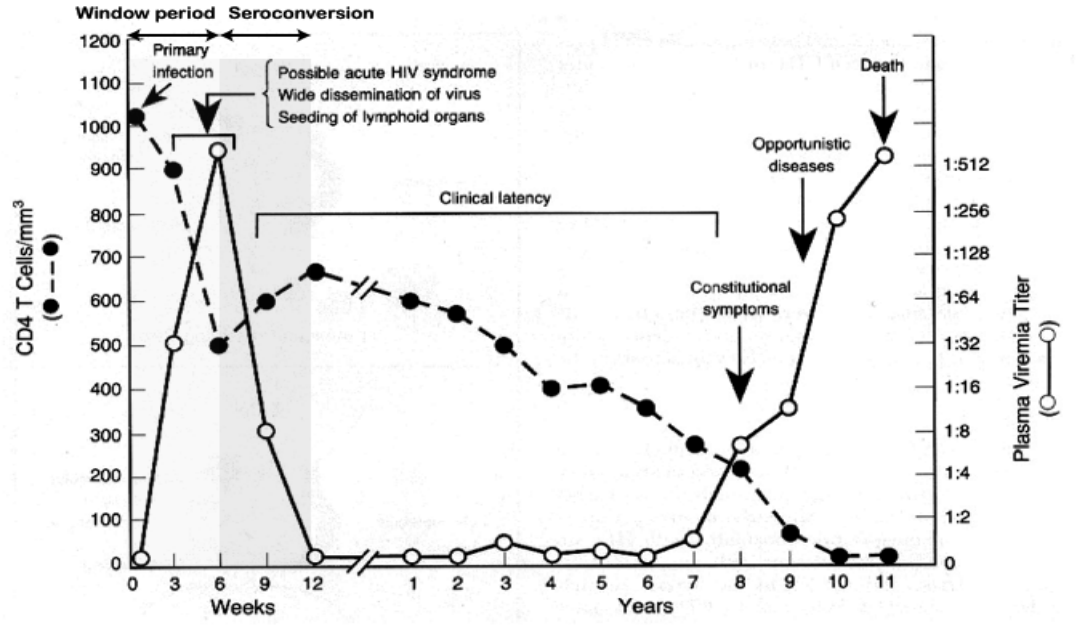


Figure 1.4. A typical course of HIV infection in an untreated individual. During the early period, after primary (acute) infection, there is widespread dissemination of virus and a sharp decrease in the number of CD4 T cells in peripheral blood, which leads to high VL. An immune response to HIV ensues, with a decrease in detectable VL followed by a prolonged period of clinical latency (the chronic infection stage) of highly variable length. The CD4 T-cell count continues to decrease during the following years, until it reaches a critical level below which there is a substantial risk of opportunistic diseases. At this stage (AIDS) VL peaks again. The "window period" is shown in light grey and the time around seroconversion is shown in darker grey. Reproduced with permission from Pantaleo et al[51], Copyright Massachusetts Medical Society.

HIV infectiousness is also 7 times greater during the AIDS stage (10-19 months before death) than during the asymptomatic chronic infection stage[49]. AIDS is associated with a critical decline in the number of CD4 T cells that allows widespread HIV dissemination and, consequently, a return to the very high VL levels previously seen in acute infection (Fig. 1.4). The period after the acute infection and before AIDS is referred to as the chronic infection stage, when the viral replication is controlled to some extent by immune responses. During this stage, the VL fluctuates around a semi-stable 'set-point' value. The value of this set-point determines time to progression to AIDS in untreated patients[52].

Consequently, there have been multiple reports suggesting that acute (recent) and late (AIDS) stages of HIV infection account for larger proportions of transmissions in a population compared to the asymptomatic stage (after the first 3-6 months and before AIDS)[53,54]. In particular, this phenomenon has been investigated through mathematical modelling[53], phylogenetic analysis[55], and by studying transmissions in serodiscordant couples (couples with different serostatuses, defined by a presence or absence of detectable antibodies)[48]. The estimated proportion of infections attributed to acutely infected individuals varies drastically, depending on the estimation methods used and the studied population. Acute infection has been estimated to account for 10-25% in MSM[56], while the estimates in heterosexuals (HET) range from 9% in the US[57] to 89% in Uganda[58]. In Chapter 1 I describe a method that I use to estimate this proportion to be 20-60% in PWID, depending on the stage of an epidemic.

Such a great variability in the estimates of the proportion of new infections attributed to acutely infected donors, combined with the significant efforts that would be required to target recently-infected individuals with preventive services, has led to an on-going debate about the contribution of recent infection to the spread of HIV and

whether recently infected individuals should be prioritised from a public health perspective, to which I had a chance to contribute[59]. A recent modelling study focused on the effect of early transmissions, relying on data from the South African generalized heterosexual epidemic, observed over a 30-year period[60]. The authors of the study showed that, over a long-term perspective, the effect of early transmission prevention is low[60]. Specifically, under various assumptions, failure to prevent transmissions from the recently infected group under various assumptions caused only a 20% reduction in the effectiveness of ART in preventing new infections after 30 years.

However, unlike the South African HIV epidemic, emerging or generalizing epidemics result from a breakthrough of the epidemic into a new pool of susceptibles, which in turn shrinks the epidemic generation time or serial interval (i.e., the average time between an individual becoming infected and transmitting to other individuals)[61]. In such outbreaks, it is more likely that transmission will be mainly led by recent infections. Similar effects are expected during the generalization of epidemics. For instance, the Latvian PWID-driven epidemic reinitiated exponential growth of the epidemic through HIV spillovers from high prevalence groups (PWID) to low prevalence groups through sexual transmissions[62]. Similarly, in the Greek HIV epidemic, the virus was mainly circulating among MSM with few PWID infections until 2009, when multiple unrelated HIV outbreaks emerged in the PWID group, resulting in a dramatic increase of recently infected individuals[63]. As mentioned above, not only the stage of epidemic, but also the main group driving transmission affects the proportion of new infections attributed to recently-infected individuals. For example, in PWID-driven epidemics, young or recent injectors who seroconvert may well be injecting unsafely with other new (and uninfected) injectors, which creates the

potential for contamination of shared injecting equipment with highly infectious blood among multiple susceptible injectors; the increase of transmissibility during early infection in this group seems to be enhanced by network and behavioural aspects on top of virus-host biology[64].

1.3.3 Epidemiological parameters of HIV spread

One of the most important aims of epidemiological studies is to describe and, when possible, to quantify HIV transmission. Knowing how fast and by what means the virus is transmitted at the population and risk group levels enables the planning of prevention and control measures. Two common epidemiological parameters used to quantify viral transmission are the *basic reproductive number* R_0 and the *generation time T*.

The basic reproductive number R_0 is "the average number of secondary infections produced when one infected individual is introduced into a host population where everyone is susceptible"; if $R_0 > 1$ the epidemic is growing and if $R_0 < 1$ the epidemic is declining[65]. When an epidemic is more advanced and it cannot be assumed that everyone is susceptible, then the *effective reproductive number* R_e can be estimated instead, which is usually lower than R_0. Without prevention R_e grows until the pool of susceptible individuals is temporarily exhausted. Estimating R_0 and R_e is crucial for the prognosis of an epidemic and for developing virus elimination strategies[66]. For example, epidemiologists use R_0 to calculate the coverage of vaccination necessary to achieve the eradication of a pathogen:

$$p_c = 1 - \frac{1}{R_0} \qquad (1.1)$$

where p_c is the critical vaccination proportion necessary to achieve the eradication of a pathogen[67].

Multiple approaches have been developed to estimate R_0, most of which can be broadly divided into two categories: those based on individual-level data and those based on population-level data. The most straightforward way to calculate R_0 is to track infected patients and their contacts from the onset of an outbreak and directly calculate the average number of secondary cases per an infected individual[68]. However, this approach is not feasible when studying infections with longer asymptomatic phases, like HIV, or studying epidemics in retrospect. More commonly used methods are based on population-level data that depend on mathematical models with known parameters and a set of assumptions[69]. Researchers typically construct mathematical models based on ordinary differential equations that embody certain assumptions about contact rate, duration of infectivity, infection probability per contact, etc, which are then used to estimate R_0 for various STIs, including HIV[68,70]. Population level models have been used to estimate R_0 for HIV since 1980s, but mostly for heterosexual populations and MSM. May & Anderson in 1987 estimated R_0 in the beginning of a heterosexual HIV epidemic to be ≈ 5[71]. Knolle in 1990 suggested that this value is much lower for low-risk heterosexuals and is below 1, ranging 0.1-0.9, depending on the number and age of sexual partners[72]. In 1993 Dietz estimated R_0 to be around 2.3 for HIV epidemics in heterosexuals, depending on the average number of sexual partners[73]. In 1994 R_0 was estimated for recently HIV-infected MSM to be in the range of 1-2[64].

Another approach is based on a likelihood-based procedure that can be used to estimate temporal changes in reproductive number from an observed epidemic curve[74]. More recently, genetic data have been also increasingly used to estimate R_0

and R_e for viral infectious diseases. Probabilistic models have been developed that jointly estimate effective population size (N_e) estimates, and birth and death rates directly from temporally calibrated viral phylogenies[75-77] (see section 1.4.2).

In infectious disease epidemiology, *generation time* (or *serial interval*), *T,* is the time between the infection of a primary infected individual and the average time to infection of one of his/her secondary cases[61]. Knowing the epidemic's generation time is important as this tells us when in the course of the epidemic interventions can be more effective to prevent transmissions. There are multiple methods to relate generation time to R_0, but the main two approaches differ only in the way they conceptualise generation time[78]. One method uses a distribution of generation times and is an extension of the Lotka-Euler equation[79]:

$$R_0 = e^{rT-(1/2)r^2\delta^2} \quad (1.2)$$

where r is the growth rate of an epidemic and δ is the standard deviation of the generation time distribution. If the distribution of generation time values is neglected and only the mean value (the "cohort" generation time) is considered, then there is an exponential relationship between the generation time and R_0[78]:

$$R_e = exp(rT) \quad (1.3)$$

1.3.4 HIV prevention

HIV prevention, meaning the prevention of transmission from an infected to a virus-free individual, is one of the most important components of successful HIV control. Prevention is often categorised into three major classes of action: behavioural, bio-

medical, and structural preventive strategies[80]. Combination of all three approaches is needed to ensure the most efficient, cost-effective, and sustainable HIV prevention[81].

Behavioural HIV prevention aims to change behaviours that can lead to HIV infection through education, empowerment, and skill-building[82]. Given the major transmission routes described above (see section 1.3.1), education about consistent condom use or reduction in the number of sexual partners are examples of behavioural prevention strategies that address sexual transmission of HIV[82]. Parenteral transmission is usually enabled through sharing of drug injecting instruments. Thus, behavioural preventive measures include education and enabling of the use of clean syringes, sponges, and other materials used for injection[81]. Harm reduction services represent a set of interventions that aim to eliminate transmission in PWID, not by preventing the use of drugs, but by making it safe for users and their peers[83]. Harm reduction services include (i) needle exchange programs, in which used syringes are replaced with clean ones provided by non-governmental organisations (NGOs) or at local pharmacies, (ii) opioid substitution treatment, where controlled quantities of drugs are provided by medical personnel at medical facilities, (iii) safe injection information booklets that give PWID information about safe injecting and sexual practices, including the prevention of disease transmission and overdoses, etc[42]. Harm reduction has proven to be effective in reducing transmission in PWID populations around the world and is now one of the most effective HIV preventive measures for this transmission group[41,42].

Biomedical prevention means the use of medication, medical procedures or materials to prevent HIV transmission. It includes the use of condoms, microbicides, male circumcision, treatment of co-factor STIs, and administration of anti-retroviral (ARV) drugs (ART). Azidothymidine (AZT) was the first ARV drug approved by Food and

Drug Administration (FDA) in the US in 1987 as ART monotherapy[84]. AZT belongs to the class of nucleoside reverse transcriptase inhibitors. Monotherapy wasn't particularly effective because the virus rapidly evolved anti-viral drug resistance mutations, and a rebound in viral load was inevitable[85]. In 1996 highly-active ART (HAART) was introduced, which dramatically changed the course of HIV treatment[86]. HAART includes ARV drugs from different classes, assigned based on the molecular mechanism and the part of the virus life cycle that they target: (i) protease inhibitors (PIs), which target the HIV protease, an enzyme responsible for polyprotein (gag and gag-pol) cleavage during virion maturation, (ii) nucleoside-analog reverse transcriptase inhibitors (NRTIs) and (iii) non-nucleoside reverse transcriptase inhibitors (NNRTIs), which target the reverse transcriptase, an enzyme responsible for converting the single-stranded HIV-RNA into double-stranded HIV-DNA, (iv) integrase inhibitors, which target the HIV integrase, an enzyme responsible for integration of viral genome into the human chromosome, (v) fusion inhibitors and (vi) coreceptor antagonists, which target HIV entry process by preventing binding and fusion with the cellular membrane[87]. Consistent ART use leads to undetectable (<50 copies/ml) levels of viral load, and, consequently, disables HIV transmission even when exchange of bodily fluids occurs between an HIV-infected and an uninfected person.

HAART can be also administered for preventive purposes, either to HIV-infected people (Treatment as prevention (TasP)), or to HIV-uninfected people (Pre-exposure prophylactics (PrEP) and Post-exposure prophylactics (PEP)). In recent years TasP has begun to play a major role in HIV prevention, as early treatment initiation has been shown to be both safe and beneficial for patients and also effective in preventing onward transmissions[88-90]. Some of the main concerns regarding TasP and PrEP

include the possibility of an increase of the level of transmitted drug-resistance[91] and the high costs of universal treatment and high coverage of PrEP, as well as associated ethical issues associated with resource allocation and access to TasP and PrEP in resource limited settings[92].

Structural preventive strategies include policy, advocacy, and environmental measures to impact communities. They include interventions designed to protect vulnerable populations at risk of HIV, based on their gender, socio-economical status, sexuality, and/or other factors that might be a basis for discrimination[80].

1.3.5 The HIV epidemic in Eastern Europe

Chapters 2, 3, and 4 of this thesis describe transmission dynamics, transmission group composition, and geographical spread of the most prevalent HIV-1 subtypes in Ukraine and Russia. I chose to study the HIV epidemic in Eastern Europe, and particularly the epidemic in the countries of the Former Soviet Union (FSU), because it presents an interesting scenario that is drastically different to that in other European countries. The first HIV cases in the Soviet Union were registered in the mid-1980s as sexual transmissions from foreigners[93]. During the following decade, sporadic HIV cases were diagnosed in MSM in Russia and Ukraine and were identified as HIV-1 subtype B, with some subtype C cases acquired through heterosexual contact, and a variety of subtypes (A, C, D, G) detected in PWID[94]. In the mid-1990s the first reports appeared of multiple subtype A infections in PWID populations in Russia and Ukraine[95,96]. Subsequent HIV-1 subtype A epidemics were registered in most post-Soviet and some neighbouring countries including Belarus, Moldova, Armenia, Kazakhstan, Kyrgyzstan, Turkmenistan, Georgia, Azerbaijan, Latvia and other Baltic countries, as well as in distant Russian provinces[97-105]. HIV-1 subtype A became the

most prevalent subtype in these countries; in many locations its relative prevalence surpasses 90%. In some of these locations an extreme founder effect was observed, suggesting that only one introduction into the local PWID population seeded further regional epidemics[96,106,107]. A few exceptions to this scenario include the Romanian HIV epidemic, which ignited with a nosocomial outbreak of HIV-1 subtype F[32,108] and the Estonian HIV-1 epidemic, created by CRF06_cpx transmission in PWID in the late 1990s[109]. The subtype A epidemic has been mostly contained within the FSU countries, though evidence of migration of A_{FSU} variant to Southern and Central Europe, probably through drug use communities and sex work, has been documented[110,111]. Rapid parenteral spread and large populations of susceptible individuals allowed the HIV epidemic in FSU countries to become the largest in the European region, accounting for two thirds of all HIV cases in Europe[112].

Bayesian phylogeographic analysis (see section 1.4.1) has been previously used to infer the origin of the FSU subtype A epidemic in Southern Ukraine[113]. Diez-Fuertes *et al.* reported that the virus likely originated in the Democratic Republic of Congo and arrived in Ukraine in the 1970s. The common ancestor of the A_{FSU} variant, which later caused the massive epidemic in FSU countries, emerged around 1984[113]. By 1993, this virus lineage had reached a highly-connected PWID population in the southern port city of Odessa, which is now recognized as the "birthplace" of the FSU epidemic[98]. Interestingly, subtypes A and B are believed to have entered the PWID population in southern Ukraine around the same time: A_{FSU} in Odessa and B_{FSU} in Mykolaiv (Russsian *Nikolaev*)[114]. Since there is no compelling evidence to believe that subtype B has a biological disadvantage compared to subtype A, it has been hypothesised that the striking difference in the spread of the two subtypes can be explained by cultural dissimilarities[114]. While Odessa has a large international port

and a large number of tourists, Mykolaiv was more isolated with some access restrictions during the Soviet period due to its close connection to the Soviet navy.

The co-circulation of the two subtypes in the Ukrainian PWID community led to the creation of the CRF03_AB circulating recombinant form (CRF), which further seeded an epidemic in PWID in Kaliningrad, Russia, in 1996, where the CRF remains the most prevalent HIV form today[107,115]. The dominance of subtype A is preserved in most of the locations discussed above, with the exception of Russia, where the relative proportions of subtype B, CRF63_02A1, and CRF02_AG are growing and are reaching 5-10% each[111].

In most settings in the FSU region, HIV-1 subtype A was associated with transmissions in people who inject drugs (PWID), while subtype B mostly accounted for infections in the MSM community[116]. The exception to this is the B_{FSU} strain that is distinct from other subtype B variants and was mostly found in PWID. In the early 2000s, B_{FSU} transmissions remained limited both in Russia and Ukraine[98,117]. The subtype B epidemic in MSM appeared to be sustained by multiple introductions from outside of the Eastern European region and was believed to be isolated from the PWID epidemic[116]. This belief was challenged recently when researchers reported that almost 20% of HIV infections in MSM in Russia were of subtype A or B_{FSU}, suggesting a shift from internationally- to locally-acquired HIV infections in MSM in the country[118].

From the early 2000s onwards, the epidemiological profile of HIV in Russia, Ukraine, and other countries in the region suggested a transition towards an increased proportion of infections attributed to heterosexual transmission. In Ukraine, for example, heterosexual transmission prevailed as the major transmission route since 2010[119], which is partly attributed to the success of harm reduction services in

PWID[120], resulting in a decrease in parenteral transmissions. Nevertheless, PWID and their sexual partners are thought to sustain the epidemic in heterosexuals, acting as so-called "bridging groups"[62]. In Russia this hypothesis is also supported by the finding that the majority of HET infections were identified as subtype A_{FSU}[117]. A recent HIV-1 subtype A study presented evidence that this mixing occurs not only between transmission groups, but also between countries, particularly in Central Asia and Russia[12].

1.4 Molecular epidemiology

Molecular epidemiology of infectious diseases aims to utilise molecular (nucleotide or amino acid) sequences to study the ecology and dynamics of pathogens[121]. Recent advances in nucleotide sequencing (i.e. high throughput sequencing technologies) allow faster and more affordable sequencing of pathogen genomes[122], resulting in the production of vast amounts of genetic data. This data-driven revolution has generated expectations with respect to more effective use of molecular sequences for scientific and public health purposes. Today's techniques have been developed to use pathogen genetic data to research the spread, distribution, treatment, and prevention of infectious disease epidemics[123,124].

In the context of viral epidemics, virus genetic sequences are often studied in combination with epidemiological information about the host. Most rapidly-evolving RNA viruses, including HIV, represent what are called measurably evolving populations, meaning that they accumulate detectable evolutionary change between sequences sampled at different points in time[125]. This allows the evolution of these viruses to be studied on the same time scale as ecological and epidemiological processes in their host populations[126].

When using the term "molecular epidemiology" scientists often refer to the field of evolutionary biology coined as phylodynamics. Phylodynamics studies how viral phylogenies are shaped by epidemiological, evolutionary, and immunological factors[127]. It has been extensively used to reconstruct the transmission dynamics of viruses, on a longer or more recent timescale, globally or within specific regions[13,128,129].

1.4.1 Phylogeography

Phylodynamic tools are based on mathematical models of population and evolutionary dynamics. To study the geographic structure of viral populations, phylogeographic models are commonly used. Typically, such models superimpose geographical information about the molecular sequences onto the phylogenetic tree in order to provide key information on the rates of virus movement across "traits" (e.g. regions, countries, villages, transmission groups). The most common "discrete trait" phylogeography models are the phylogenetic models that are similar to those used to model nucleotide substitution processes[130,131] (Fig. 1.5).

Phylogeographic methods can be used not only to describe how infectious diseases spread over particular territories, but also to hypothesise what factors (evolutionary and/or ecological, such as political and socio-economical) initiated these particular dissemination trends. Recently, a Bayesian model was developed that allows including covariates directly into the demographic inference framework implemented in the BEAST software (see section 1.4.4 for more details on BEAST)[132]. This model allows estimating the association between the effective population size (N_e) and covariates of interest, while accounting for the uncertainly in the evolutionary history. These methods have been applied at local, country, regional, and global levels to study multiple viruses, including HIV[4,133].

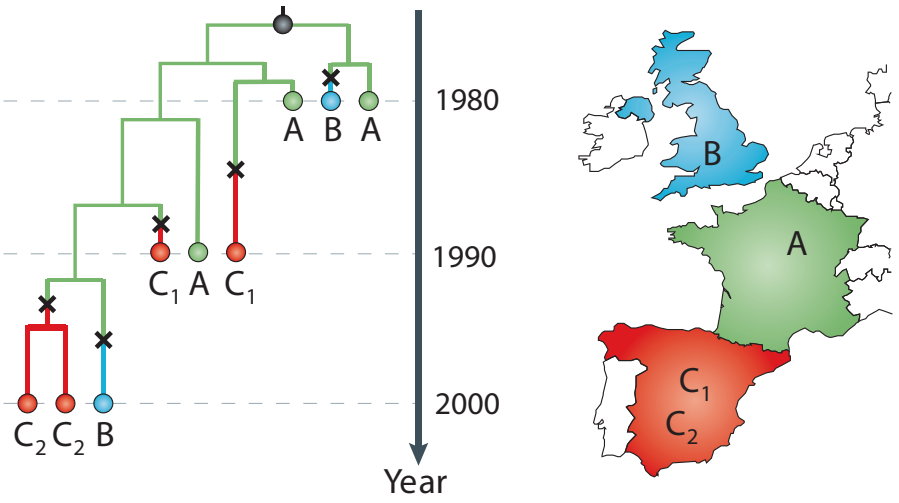


Figure 1.5. Phylogeographic analysis of sequences from France (green, A), the United Kingdom (blue, B) and two locations in Spain (red, C_1 and C_2). On the left, the colour of internal branches corresponds to the estimated location. Lineage migration events are marked on the phylogeny with crosses. The black dot indicates the most recent common ancestor of the sample (from Pybus & Rambaut[126]. Licence to reproduce the material granted on December 18th 2018).

1.4.2 Population dynamics models

Population dynamic models describe changes in the size of a given population, driven by biological and environmental processes. There are two large classes of population dynamic models used in phylodynamics, which differ in their underlying assumptions and applications: coalescent-based and birth-death models.

1.4.2.1 Coalescent-based models

Coalescent theory, which was first developed in 1982 by Kingman[134], helps to relate phylogenetic trees to the size of studied populations traced back through time, as long as some assumptions about the structure of these populations can be made (Fig. 1.6)[126]. Specifically, coalescent models require a small proportion of the studied

population to have been sampled. Populations are studied going back in time, starting from the most recently sampled individuals. In this case, the rate at which the common ancestor of any two lineages can be found (the coalescence rate) is inversely proportional to the size of the population[126,135]. Other factors can also influence the coalescence rate, such as population dynamics, selection, and recombination[136]. The last coalescent event represents the most recent common ancestor (MRCA) of the entire sample.

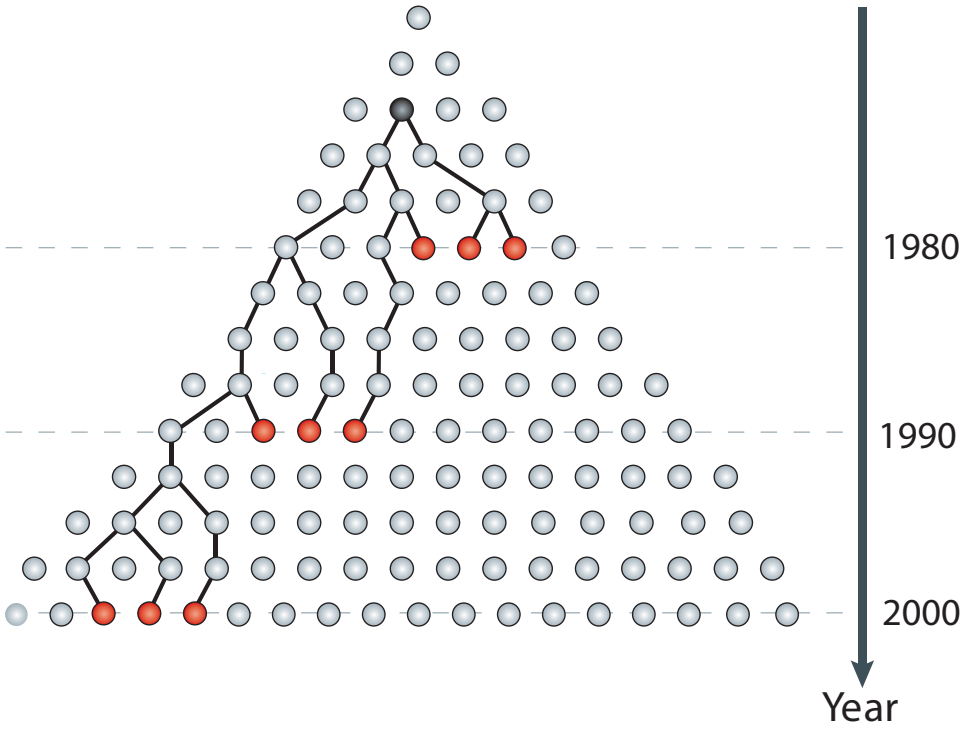


Figure 1.6. An illustration of the coalescent process for an epidemic growing through time; red circles represent sampled infections, grey circles represent unsampled infections, black circles represent the common ancestor of the sampled infections. Row widths represent sample size at the corresponding year of sample collection (from Pybus & Rambaut[126]. Licence to reproduce the material granted on December 18th 2018).

Coalescent-based approaches rely on specific models of population size and reproduction, which are often different from and are more simplistic than real-life scenarios, as they represent an ideal case based on a set of underlying assumptions. Specifically, the Wright-Fisher and Moran reproduction models are most frequently used. The Wright-Fisher population model represents reproduction under the following assumptions: (i) constant population size; (ii) non-overlapping generations, meaning that a generation dies out as soon as its offspring is produced; (iii) there is variation in reproductive success among individuals, but (iv) that variance is not inherited by offspring. The model assumes no selection or population subdivision. The main difference of the Moran model from the Wright-Fisher model is that it assumes overlapping generations.

A variety of coalescent-based approaches have been developed and applied to estimate past population dynamics. These methods assume an underlying "demographic model", the parameters of which can be inferred from sampled sequences or phylogenetic trees. Simple parametric demographic models include constant population size[137], exponential growth model[138], and logistic growth model[139]. Alternatively, more flexible, non-parametric models can be used, such as the Bayesian Skyline[140], Bayesian Skygrid[141], and Bayesian Skyride[142] models. These are piece-wise constant models of population size that are able to fit different demographic scenarios and estimate a posterior distribution of effective population size (N_e) through time. All these approaches have been successfully applied to study viral infectious diseases including HIV and have produced robust estimates that largely agree with available epidemiological information[4,141].

1.4.2.2 Birth-death models

Birth-death phylodynamic models are an alternative to coalescent models that assume that population dynamics follow a birth-death (BD) process[143]. This process can be described as going forward in time, starting from one infected person (lineage). Lineages split at "birth" (transmission) events and cease to exist at "death" (recovery) events. In simple BD models the birth and death rates are assumed to be constant among lineages and through time. The birth and death events occur independently and are assumed not to occur at the same time[144]. The "complete" phylogenetic tree generated by a BD process includes extinct lineages. However, only those lineages that are sampled can be included in the "reconstructed" tree that is built from sampled genetic sequences, i.e. the reconstructed tree excludes extinct lineages (Fig. 1.7). Even though the simple BD process in Fig. 1.7 assumes isochronous sampling (i.e. sampling at one point in time), extensions to the original phylodynamic BD model have been developed in which lineages are sampled through time (heterochronous sampling)[145].

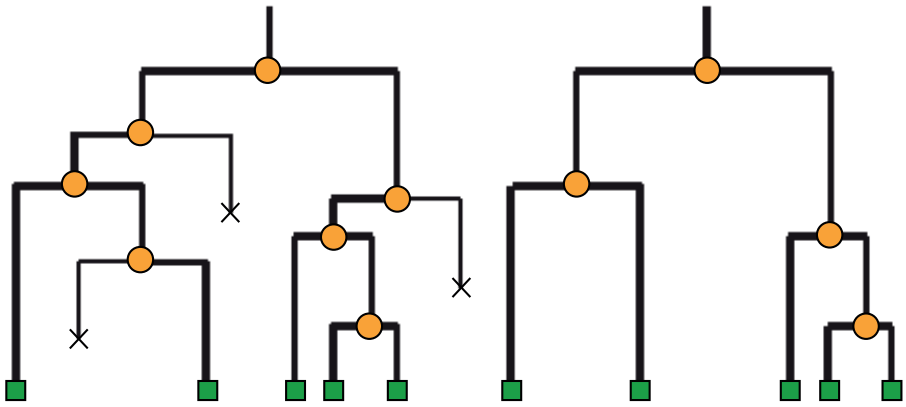


Figure 1.7. Illustration of a complete birth-death tree (left) and its corresponding reconstructed tree (right). Orange circles represent birth events, black crosses

represent death events, green squares represent sampling events; the width of the branches corresponds to their posterior probability (from Gernhard[144]. Licence to reproduce the material granted on December 18th 2018).

One of the important applications of BD models in molecular epidemiology is the estimation of the basic reproductive number R_0 directly from viral genetic sequences[75,76]. The simplest such approach involves a constant rate BD model, which assumes a constant transmission (birth) rate (λ), a constant becoming uninfectious through recovery or death (death) rate (μ), and a constant sampling (death) rate (ψ)[75]. R_e can then be derived as a ratio of birth and death rates:

$$R_0 = \frac{\lambda}{\mu+\psi} \qquad (1.4)$$

A more complicated model, BD Skyline plot (BDSKY), allows birth, death, and sampling rates to vary through time in a piece-wise constant manner[76]. This model allows researchers to estimate R_e through time for a given number of time intervals between the origin of an epidemic and the date of the most recent sample in the analysis.

1.4.3 Evolutionary models

Other important components of molecular epidemiology studies are models that describe how genetic sequences evolve through time. *Nucleotide substitution models* are used to estimate genetic distances (the number of substitutions per site) among sequences. These models predict the rate of nucleotide substitutions at a certain position and the distribution of substitutions across a sequence. Nucleotide

substitution models are used to account for changes that cannot be observed in a given sequence (i.e., reverse mutations or intermediate substitutions).

There are a variety of nucleotide substitution models that differ in their assumptions and the combinations of parameters they include. Nucleotide substitution models vary from simplest scenario, introduced by Jukes and Cantor (JC69)[146] in which base frequencies are equal and all types of substitutions are equally likely, to the general time reversible (GTR) model[147] with gamma distributed among-site rate variation, which is often used in modern phylodynamic analysis. The choice of an appropriate nucleotide substitution model is important for correct reconstructions of tree topologies[148] and accurate bootstrap calculations[149]. It has also been shown that codon-based models, which take into account different mutational patterns at different codon positions, often describe evolutionary change in nucleotide alignments better than standard models[150].

Furthermore, phylodynamic methods rely heavily on *molecular clock models*, which are based on the notion that the evolutionary rate of nucleotide sequences is relatively constant through time and can be described in mathematical terms by the amount of genetic divergence accumulating through time[151]. The *strict molecular clock model* assumes constant rate of evolution across all branches of a phylogeny and has a simple parameter "the evolutionary rate" measured in substitutions per site per year. In contrast, the *relaxed molecular clock* allows variation in the rate of evolution across branches of a phylogeny, regardless of the position of those branches within the tree[152]. The evolutionary rate that can be estimated by molecular clock models depends on many biological and epidemiological factors, including generation time, population size, and population dynamics[153].

1.4.4 Applications of molecular epidemiology to HIV

Molecular epidemiology was used to address a wide variety of questions, including to determine HIV diversity and structure in a given population, to estimate HIV transmission parameters, to infer changes in viral population size at a local or global scale, and to investigate transmission chains in detail, during outbreaks or for forensic investigations[4,40,76,132,154,155].

Recent examples of the use of genomic data include:

- Epidemiological surveillance. Molecular epidemiology has been used to monitor emerging HIV hotspots and to define growing clusters of HIV cases[156,157].
- Real-time outbreak investigations. In the past, the application of molecular epidemiology methods to infectious disease outbreaks was complicated by the task of sequencing viral genomes quickly as an outbreak develops. Consequently, molecular data have been mostly used in a retrospective manner to investigate and reconstruct outbreaks, such as a nosocomial HIV and HCV outbreak in Libya[158] or recent HIV outbreaks in PWID in Athens and Bucharest[159], in Indiana[39], and in Scotland[40]. However, molecular sequence data can be now produced and analysed faster and at a lower cost than before, so their use in real-life and real-time outbreak investigations might become more feasible for more diseases[160-162]. Further, portable sequencing technologies such as MinION have become available and allow pathogen sequencing in smaller clinics where diagnosis takes place[163]. Applications of these methods to HIV are being explored.

A popular software package used to undertake molecular epidemiology analyses is Bayesian Evolutionary Analysis Sampling Trees (BEAST)[164]. BEAST uses Markov Chain Monte Carlo (MCMC) sampling to analyse molecular sequences and infer parameters of evolutionary and phylodynamic models. It can be used to produce

molecular clock trees, based on strict or relaxed molecular clock models, and also to estimate population dynamics (see section 1.4.2). It combines nucleotide substitution, molecular clock, phylodynamic models to generate a posterior distribution of trees, where each tree is sampled proportionally to its posterior probability. A second version of BEAST called BEAST 2.0, implements a different range of evolutionary models[165]. I use both versions of BEAST in this thesis.

1.5 Thesis outline

In this thesis I explore multiple ways in which molecular epidemiology can be used to describe HIV population dynamics, to answer topical public health questions, and to inform HIV prevention programs, mostly for PWID in Eastern and Southern European region.

In Chapters 2, 3, and 4 I study a large HIV-1 epidemic in Eastern Europe, with a primary focus on Ukraine and infections in PWID. In Chapter 2 I describe how phylodynamic analysis and epidemiological compartmental modelling can be combined in order to describe HIV spread in PWID during the initial stage of an epidemic in the absence of prevention services. Specifically, I estimate epidemiological parameters (R_0 and epidemic generation time T) for HIV spread in PWID in Russia and Ukraine in the 1990s, and then explored the potential effect of different preventive strategies on epidemic development, given the estimated parameters. I also estimate the proportion of infections attributed to recently infected individuals during the initial stages of a PWID epidemic.

In Chapter 3 I analyse newly generated HIV-1 subtype A genetic sequences from Ukraine and study viral lineage movement among different regions of the country, particularly in the context of the recent political crisis and war in the eastern part of

the country. I apply a novel phylogeographic approach that allows estimating the change of the virus movement before and after the initiation of the war. I then describe host behavioural patterns in the regions that were the main recipients of the viral flow and I recommend prevention strategies to prevent the emergence of local outbreaks, emphasizing the importance of an effective harm reduction approach for PWID.

Whilst I study HIV-1 subtype A in Chapters 2 and 3, which is by far the most prevalent subtype in Eastern Europe, Chapter 4 focuses on the molecular epidemiology of subtype B in Ukraine. In this chapter I use the largest subtype B dataset from Ukraine, which also includes some newly generated sequence data, to describe the distribution of HIV subtype B by region and transmission group, and how this distribution has changed in the last 15 years. I also conduct a simulation analysis of the effect of sampling proportion and population size on the robustness of popular phylodynamic inference models, which estimate changes in the effective population size and the epidemic reproduction number from viral sequences.

Finally, in Chapter 5 I expand my research to study HIV-1 dynamics in Portugal. Portugal presents an interesting case from a molecular epidemiology point of view, because it has the most diverse HIV profile in Europe, and from a public health point of view, because Portugal has approached HIV epidemic in PWID differently compared to other European countries. I study subtypes B and G in Portugal, two of the most prevalent subtypes in the country, their transmission group composition and changes in transmission dynamics (R_e) over time. In this chapter I use a variety of birth-death models to study the effect of different preventive interventions on the growth dynamics of the HIV epidemic.

1.6 References

1. Gottlieb MS, Schanker MD, Fan PT, Saxon MD, Weisman JD, Centers for Disease Control (CDC). MMWR Morb Mortal Wkly Rep. 1981;30:250–2.
2. Barre-Sinoussi F, Chermann JC, Rey F, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 1983;220:868-71.
3. Gallo RC, Sarin PS, Gelmann EP, et al. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). Science 1983;220:865-7.
4. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56-61.
5. German Advisory Committee. Human Immunodeficiency Virus (HIV). Transfus Med Hemother 2016;43:203-22.
6. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. The challenge of HIV-1 subtype diversity. N Engl J Med 2008;358:1590-602.
7. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.
8. Robertson DL, Anderson JP, Bradac JA, et al. HIV-1 nomenclature proposal. Science 2000;288:55-6.
9. Buonaguro L, Tornesello ML, Buonaguro FM. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J Virol 2007;81:10209-19.
10. Los Alamos National Laboratory. The Human Retroviruses and AIDS 1999 Compendium. 1999.
11. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. Aids 2006;20:W13-W23.
12. Aibekova L, Foley B, Hortelano G, et al. Molecular epidemiology of HIV-1 subtype A in former Soviet Union countries. Plos One 2018;13:e0191891.
13. Magiorkinis G, Angelis K, Mamais I, et al. The global spread of HIV-1 subtype B epidemic. Infect Genet Evol 2016;46:169-79.
14. Cloyd MW. Human Retroviruses. In: Medical Microbiology. Galveston (TX)1996.
15. Watts JM, Dang KK, Gorelick RJ, et al. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 2009;460:711-6.
16. Wongstaal F. The Aids Virus What We Know and What We Can Do About It. Western J Med 1991;155:481-7.
17. Malim MH, Emerman M. HIV-1 accessory proteins-ensuring viral survival in a hostile environment. Cell Host Microbe 2008;3:388-98.
18. Gurtler L, Aepfelbacher M, Bauerfeind U, et al. Human Immunodeficiency Virus (HIV). Transfus Med Hemoth 2016;43:203-22.
19. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nature Reviews Genetics 2004;5:52-61.
20. Mansky LM, Temin HM. Lower in Vivo Mutation Rate of Human Immunodeficiency Virus Type-1 Than That Predicted from the Fidelity of Purified Reverse Transcriptase. Journal of Virology 1995;69:5087-94.
21. Cuevas JM, Geller R, Garijo R, Lopez-Aldeguer J, Sanjuan R. Extremely High Mutation Rate of HIV-1 In Vivo. PLoS Biol 2015;13:e1002251.
22. Nowak M. HIV mutation rate. Nature 1990;347:522.

23. Roberts JD, Bebenek K, Kunkel TA. The accuracy of reverse transcriptase from HIV-1. Science 1988;242:1171-3.
24. Michael NL. Host genetic influences on HIV-1 pathogenesis. Curr Opin Immunol 1999;11:466-74.
25. Patino-Galindo JA, Gonzalez-Candelas F. The substitution rate of HIV-1 subtypes: a genomic approach. Virus Evol 2017;3(2).
26. Barouch DH, Tomaka FL, Wegmann F, Stieh DJ, Alter G, et al. Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13-19). The Lancet 2018;392(10143).
27. Dahabieh MS, Battivelli E, Verdin E. Understanding HIV Latency: The Road to an HIV Cure. Annu Rev Med 2015;66:407-21.
28. Shaw GM, Hunter E. HIV transmission. Cold Spring Harb Perspect Med 2012;2(11).
29. Kharsany AB, Karim QA. HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. Open AIDS J 2016;10:34-48.
30. Center for Disease Control and Prevention (CDCP). HIV among Gay and Bisexual Men. CDCP Fact Sheet. 2017.
31. Gissequist DP. Estimating HIV-1 transmission efficiency through unsafe medical injections. Int J Std Aids 2002;13:152-9.
32. Patrascu IV, Dumitrescu O. The epidemic of human immunodeficiency virus infection in Romanian children. AIDS Res Hum Retroviruses 1993;9:99-104.
33. Lowe GD. Haemophilia, blood products and HIV infection. Scott Med J 1987;32:109-11.
34. Goldwater PN. Iatrogenic blood-borne viral infections in refugee children from war and transition zones. Emerg Infect Dis 2013;19.
35. World Health Organization. Consensus statement on screening blood donations for infectious agents through blood transfusion. Geneva: World Health Organization; 1991.
36. World Health Organization. Blood safety and availability: World Health Organization; 2017.
37. Jolley E, Rhodes T, Platt L, et al. HIV among people who inject drugs in Central and Eastern Europe and Central Asia: a systematic review with implications for policy. Bmj Open 2012;2.
38. UNAIDS. The GAP report 2014. People who inject drugs. 2014.
39. Peters PJ, Pontones P, Hoover KW, et al. HIV Infection Linked to Injection Use of Oxymorphone in Indiana, 2014-2015. New Engl J Med 2016;375:229-39.
40. Ragonnet-Cronin M, Jackson C, Bradley-Stewart A, et al. Recent and Rapid Transmission of HIV Among People Who Inject Drugs in Scotland Revealed Through Phylogenetic Analysis. Journal of Infectious Diseases 2018;217:1875-82.
41. MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and Hepatitis C in people who inject drugs: A review of reviews to assess evidence of effectiveness. Int J Drug Policy 2014;25:34-52.
42. Wodak A, McLeod L. The role of harm reduction in controlling HIV among injecting drug users. AIDS 2008;22 Suppl 2:S81-92.
43. Luzuriaga K, Mofenson LM. Challenges in the Elimination of Pediatric HIV-1 Infection. N Engl J Med 2016;374:761-70.
44. Powers KA, Poole C, Pettifor AE, Cohen MS. Rethinking the heterosexual infectivity of HIV-1: a systematic review and meta-analysis. Lancet Infect Dis 2008;8:553-63.

45. Quinn TC, Wawer MJ, Sewankambo N, et al. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. N Engl J Med 2000;342:921-9.
46. Fleming DT, Wasserheit JN. From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. Sex Transm Infect 1999;75:3-17.
47. Weiss HA, Quigley MA, Hayes RJ. Male circumcision and risk of HIV infection in sub-Saharan Africa: a systematic review and meta-analysis. AIDS 2000;14:2361-70.
48. Wawer MJ, Gray RH, Sewankambo NK, et al. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. J Infect Dis 2005;191:1403-9.
49. Hollingsworth TD, Anderson RM, Fraser C. HIV-1 transmission, by stage of infection. J Infect Dis 2008;198:687-93.
50. Cohen MS, Shaw GM, McMichael AJ, Haynes BF. Acute HIV-1 Infection. N Engl J Med 2011;364:1943-54.
51. Pantaleo G, Graziosi C, Fauci AS. The immunopathogenesis of human immunodeficiency virus infection. N Engl J Med 1993;328:327-35.
52. Mellors JW, Rinaldo CR, Jr., Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science 1996;272:1167-70.
53. Koopman JS, Jacquez JA, Welch GW, et al. The role of early HIV infection in the spread of HIV through populations. J Acq Immun Def Synd 1997;14:249-58.
54. Rapatski BL, Suppe F, Yorke JA. HIV epidemics driven by late disease stage transmission. J Acquir Immune Defic Syndr 2005;38:241-53.
55. Brenner BG, Roger M, Routy JP, et al. High rates of forward transmission events after acute/early HIV-1 infection. J Infect Dis 2007;195:951-9.
56. Xiridou M, Geskus R, de Wit J, Coutinho R, Kretzschmar M. Primary HIV infection as source of HIV transmission within steady and casual partnerships among homosexual men. Aids 2004;18:1311-20.
57. Pinkerton SD. How many sexually-acquired HIV infections in the USA are due to acute-phase HIV transmission? Aids 2007;21:1625-9.
58. Pinkerton SD. Probability of HIV transmission during acute infection in Rakai, Uganda. AIDS Behav 2008;12:677-84.
59. Vasylyeva TI, Friedman SR, Magiorkinis G. Prevention of early HIV transmissions might be more important in emerging or generalizing epidemics. Proc Natl Acad Sci U S A 2015;112:E1515.
60. Eaton JW, Hallett TB. Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence. Proc Natl Acad Sci U S A 2014;111:16202-7.
61. Svensson A. A note on generation times in epidemic models. Math Biosci 2007;208:300-11.
62. Graw F, Leitner T, Ribeiro RM. Agent-based and phylogenetic analyses reveal how HIV-1 moves between risk groups: Injecting drug users sustain the heterosexual epidemic in Latvia. Epidemics 2012;4:104-16.
63. Paraskevis D, Nikolopoulos G, Tsiara C, et al. HIV-1 outbreak among injecting drug users in Greece, 2011: a preliminary report. Eurosurveillance 2011;16:2-5.

64. Jacquez JA, Koopman JS, Simon CP, Longini IM, Jr. Role of the primary infection in epidemics of HIV infection in gay cohorts. J Acquir Immune Defic Syndr 1994;7:1169-84.
65. Anderson RM, May RM. Infectious diseases of humans : dynamics and control. Oxford: Oxford University Press; 1991.
66. Farrington CP. On vaccine efficacy and reproduction numbers. Math Biosci 2003;185:89-109.
67. Scherer A, McLean A. Mathematical models of vaccination. Br Med Bull 2002;62:187-99.
68. Breban R, Vardavas R, Blower S. Theory versus data: how to calculate R0? PLoS One 2007;2:e282.
69. Diekmann O, Heesterbeek JAP, Metz JAJ. On the Definition and the Computation of the Basic Reproduction Ratio R0 in Models for Infectious-Diseases in Heterogeneous Populations. J Math Biol 1990;28:365-82.
70. Dietz K. The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res 1993;2:23-41.
71. May RM, Anderson RM. Transmission dynamics of HIV infection. Nature 1987;326:137-42.
72. Knolle H. Age Preference in Sexual Choice and the Basic Reproduction Number of Hiv Aids. Biometrical J 1990;32:243-56.
73. Dietz K, Heesterbeek JA, Tudor DW. The basic reproduction ratio for sexually transmitted diseases. Part 2. Effects of variable HIV infectivity. Math Biosci 1993;117:35-47.
74. Wallinga J, Teunis P. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol 2004;160:509-16.
75. Stadler T, Kouyos R, von Wyl V, et al. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol 2012;29:347-57.
76. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 2013;110:228-33.
77. Magiorkinis G, Sypsa V, Magiorkinis E, et al. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. PLoS computational biology 2013;9:e1002876.
78. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. P Roy Soc B-Biol Sci 2007;274:599-604.
79. Dublin L. I. & Lotka AJ. On the true rate of natural increase, as exemplified by the population of the United States, 1920. J Am Stat Assoc 1925;150:305-39.
80. Merson MH, O'Malley J, Serwadda D, Apisuk C. The history and challenge of HIV prevention. The Lancet 2008;372:475-88.
81. Rotheram-Borus MJ, Swendeman D, Chovnick G. The past, present, and future of HIV prevention: integrating behavioral, biomedical, and structural intervention strategies for the next generation of HIV prevention. Annu Rev Clin Psychol 2009;5:143-67.
82. Bekker LG, Beyrer C, Quinn TC. Behavioral and biomedical combination strategies for HIV prevention. Cold Spring Harb Perspect Med 2012;2.
83. Desjarlais DC. Harm Reduction - a Framework for Incorporating Science into Drug Policy. Am J Public Health 1995;85:10-2.

84. Brook I. Approval of Zidovudine (Azt) for Acquired-Immunodeficiency-Syndrome - a Challenge to the Medical and Pharmaceutical Communities. Jama-J Am Med Assoc 1987;258(11):1517.
85. Vella S, Schwartlander B, Sow SP, Eholie SP, Murphy RL. The history of antiretroviral therapy and of its implementation in resource-limited areas of the world. AIDS 2012;26:1231-41.
86. Barbaro G, Scozzafava A, Mastrolorenzo A, Supuran CT. Highly active antiretroviral therapy: current state of the art, new agents and their pharmacological interactions useful for improving therapeutic outcome. Curr Pharm Des 2005;11:1805-43.
87. Arts EJ, Hazuda DJ. HIV-1 antiretroviral drug therapy. Cold Spring Harb Perspect Med 2012;2:a007161.
88. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. New Engl J Med 2011;365:493-505.
89. Danel C, Moh R, Gabillard D, et al. A Trial of Early Antiretrovirals and Isoniazid Preventive Therapy in Africa. New Engl J Med 2015;373:808-22.
90. Lundgren D, Babiker AG, Gordin F, et al. Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection. New Engl J Med 2015;373:795-807.
91. Hurt CB, Eron JJ, Cohen MS. Pre-Exposure Prophylaxis and Antiretroviral Resistance: HIV Prevention at a Cost? Clin Infect Dis 2011;53:1265-70.
92. Sugarman J. Bioethical Challenges With HIV Treatment as Prevention. Clin Infect Dis 2014;59:S32-S4.
93. Pokrovskiy V. HIV epidemic in Russia and neighbouring countries. J Int AIDS Soc 2014;17:19502.
94. Lukashov VV, Cornelissen MTE, Goudsmit J, et al. Simultaneous Introduction of Distinct Hiv-1 Subtypes into Different Risk Groups in Russia, Byelorussia and Lithuania. Aids 1995;9:435-9.
95. Novitsky VA, Montano MA, Essex M. Molecular epidemiology of an HIV-1 subtype A subcluster among injection drug users in the Southern Ukraine. AIDS Res Hum Retroviruses 1998;14:1079-85.
96. Bobkov A, CheingsongPopov R, Selimova L, et al. An HIV type 1 epidemic among injecting drug users in the former Soviet Union caused by a homogeneous subtype A strain. Aids Res Hum Retrov 1997;13:1195-201.
97. Kazennova E, Laga V, Lapovok I, et al. HIV-1 genetic variants in the Russian Far East. AIDS Res Hum Retroviruses 2014;30:742-52.
98. Saad MD, Shcherbinskaya AM, Nadai Y, et al. Molecular epidemiology of HIV Type 1 in Ukraine: birthplace of an epidemic. AIDS Res Hum Retroviruses 2006;22:709-14.
99. Lazouskaya NV, Eremin VF, Adema KW, Gasich EL, Baan E, Lukashov VV. The HIV type 1 epidemic in Belarus: predominance of Eastern European subtype A strains and circulation of subtype B viruses. AIDS Res Hum Retroviruses 2005;21:830-3.
100. Pandrea I, Descamps D, Collin G, et al. HIV type 1 genetic diversity and genotypic drug susceptibility in the Republic of Moldova. AIDS Res Hum Retroviruses 2001;17:1297-304.
101. Laga V, Vasilyev A, Lapovok I, et al. HIV Type 1 Subtype A1 Dominates in Armenia. Curr HIV Res 2015;13:219-25.
102. Bobkov A, Kazennova E, Selimova L, et al. A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: Identification of subtype A,

subtype B, and novel gagA/envB recombinants. Aids Res Hum Retrov 1998;14:669-76.
103. Balode D, Ferdats A, Dievberna I, et al. Rapid epidemic spread of HIV type 1 subtype A1 among intravenous drug users in Latvia and slower spread of subtype B among other risk groups. Aids Res Hum Retrov 2004;20:245-9.
104. Zarandia M, Tsertsvadze T, Carr JK, Nadai Y, Sanchez JL, Nelson K. HIV-1 genetic diversity and genotypic drug susceptibility in the Republic of Georgia. Aids Res Hum Retrov 2006;22:470-6.
105. Carr JK, Nadai Y, Eyzaguirre L, et al. Outbreak of a West African recombinant of HIV-1 in Tashkent, Uzbekistan. Jaids-J Acq Imm Def 2005;39:570-5.
106. Lukashov VV, Karamov EV, Eremin VF, Titov LP, Goudsmit J. Extreme founder effect in an HIV type 1 subtype A epidemic among drug users in Svetlogorsk, Belarus. AIDS Res Hum Retroviruses 1998;14:1299-303.
107. Liitsola K, Tashkinova I, Laukkanen T, et al. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. AIDS 1998;12:1907-19.
108. Mbisa JL, Hue S, Buckton AJ, et al. Phylodynamic and phylogeographic patterns of the HIV type 1 subtype F1 parenteral epidemic in Romania. AIDS Res Hum Retroviruses 2012;28:1161-6.
109. Adojaan M, Kivisild T, Mannik A, et al. Predominance of a rare type of HIV-1 in Estonia. J Acquir Immune Defic Syndr 2005;39:598-605.
110. Lai A, Bozzi G, Franzetti M, et al. HIV-1 A1 Subtype Epidemic in Italy Originated from Africa and Eastern Europe and Shows a High Frequency of Transmission Chains Involving Intravenous Drug Users. Plos One 2016;11.
111. Beloukas A, Psarris A, Giannelou P, Kostaki E, Hatzakis A, Paraskevis D. Molecular epidemiology of HIV-1 infection in Europe: An overview. Infect Genet Evol 2016;46:180-9.
112. European Center for Disease Prevention and Control/WHO Regional Office for Europe. HIV/AIDS surveillance in Europe 2010. Stockholm: European Centre for Disease Prevention and Control;2011.
113. Diez-Fuertes F, Cabello M, Thomson MM. Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. Infect Genet Evol 2015;33:197-205.
114. Nabatov AA, Kravchenko ON, Lyulchuk MG, Shcherbinskaya AM, Lukashov VV. Simultaneous introduction of HIV type 1 subtype A and B viruses into injecting drug users in southern Ukraine at the beginning of the epidemic in the former Soviet Union. AIDS Res Hum Retroviruses 2002;18:891-5.
115. Liitsola K, Holm K, Bobkov A, et al. An AB recombinant and its parental HIV type 1 strains in the area of the former Soviet Union: low requirements for sequence identity in recombination. UNAIDS Virus Isolation Network. AIDS Res Hum Retroviruses 2000;16:1047-53.
116. Dukhovlinova E, Masharsky A, Toussova O, et al. Two Independent HIV Epidemics in Saint Petersburg, Russia Revealed by Molecular Epidemiology. AIDS Res Hum Retroviruses 2015;31:608-14.
117. Bobkov AF, Kazennova EV, Selimova LM, et al. Temporal trends in the HIV-1 epidemic in Russia: predominance of subtype A. J Med Virol 2004;74:191-6.
118. Kazennova E, Laga V, Gromov K, et al. Genetic Variants of HIV Type 1 in Men Who Have Sex with Men in Russia. Aids Res Hum Retrov 2017;33:1061-4.
119. Ministry of Health of Ukraine. HIV-infection in Ukraine. Informational bulletin N45. Kyiv: Ukrainian Center for Control for Socially Dangerous Diseases of

Ministry of Health of Ukraine; Institute of Epidemiology and Infectious Diseases named after L.V. Gromashevskiy; 2016.
120. Vitek CR, Cakalo JI, Kruglov YV, et al. Slowing of the HIV Epidemic in Ukraine: Evidence from Case Reporting and Key Population Surveys, 2005-2012. Plos One 2014;9(9).
121. Foxman B, Riley L. Molecular epidemiology: focus on infection. American journal of epidemiology 2001;153:1135-41.
122. Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. The Journal of investigative dermatology 2013;133:e11.
123. Kuhnert D, Wu CH, Drummond AJ. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. Infection, genetics and evolution 2011; 11:1825-41.
124. Hartfield M, Murall CL, Alizon S. Clinical applications of pathogen phylogenies. Trends in molecular medicine 2014;20:394-404.
125. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends Ecol Evol 2003;18:481-8.
126. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 2009;10:540-50.
127. Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol 2013;9:e1002947.
128. Gray RR, Tatem AJ, Lamers S, et al. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. AIDS 2009;23:F9-F17.
129. Pybus OG, Barnes E, Taggart R, et al. Genetic history of hepatitis C virus in East Asia. J Virol 2009;83:1071-82.
130. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol 2009;5:e1000520.
131. Salemi M, Lamers SL, Yu S, de Oliveira T, Fitch WM, McGrath MS. Phylodynamic analysis of human immunodeficiency virus type 1 in distinct brain compartments provides a model for the neuropathogenesis of AIDS. J Virol 2005;79:11343-52.
132. Gill MS, Lemey P, Bennett SN, Biek R, Suchard MA. Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates. Syst Biol 2016;65:1041-56.
133. Paraskevis D, Pybus O, Magiorkinis G, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology 2009;6:49.
134. Kingman JFC. The coalescent. Stochastic Processes and their Applications 1982;13:235-48.
135. Wakeley J. Coalescent Theory: An Introduction Systematic Biology 2009;58:162–5.
136. Fu YX, Li WH. Coalescing into the 21st century: An overview and prospects of coalescent theory. Theor Popul Biol 1999;56:1-10.
137. Kingman J. The coalescent. Stochastic Processes and Their Applications 1982;13:235-48.
138. Griffiths RC, Tavare S. Sampling Theory for Neutral Alleles in a Varying Environment. Philos T Roy Soc B 1994;344:403-10.
139. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 2000;155(3): 1429-1437.

140. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 2005;22:1185-92.
141. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol 2013;30:713-24.
142. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol 2008;25:1459-71.
143. Stadler T. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J Theor Biol 2009;261:58-66.
144. Gernhard T. The conditioned reconstructed process. Journal of Theoretical Biology 2008;253:769-78.
145. Stadler T. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 2010;267:396-404.
146. Jukes TH, Cantor CR. Evolution of protein molecules. In: HN M, ed. Mammalian Protein Metabolism. New York: Academic Press; 1969:21-132.
147. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol 1984;20:86-93.
148. Hoff M, Orf S, Riehm B, Darriba D, Stamatakis A. Does the choice of nucleotide substitution models matter topologically? Bmc Bioinformatics 2016;17.
149. Buckley TR, Cunningham CW. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol Biol Evol 2002;19:394-405.
150. Shapiro B, Rambaut A, Drummond AJ. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 2006;23:7-9.
151. Pybus OG. Model selection and the molecular clock. PLoS Biol 2006; 4:e151.
152. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology 2006;4:e88.
153. Holmes EC, Dudas G, Rambaut A, Andersen KG. The evolution of Ebola virus: Insights from the 2013-2016 epidemic. Nature 2016;538:193-200.
154. Bezemer D, Cori A, Ratmann O, et al. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLoS Med 2015;12:e1001898.
155. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. Proc Natl Acad Sci U S A 2016;113:2690-5.
156. Poon AF, Gustafson R, Daly P, et al. Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study. Lancet HIV 2016;3:e231-8.
157. Wertheim JO, Murrell B, Mehta SR, et al. Growth of HIV-1 Molecular Transmission Clusters in New York City. J Infect Dis 2018.
158. de Oliveira T, Pybus OG, Rambaut A, et al. Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. Nature 2006;444:836-7.
159. Paraskevis D, Paraschiv S, Sypsa V, et al. Enhanced HIV-1 surveillance using molecular epidemiology to study and monitor HIV-1 outbreaks among intravenous drug users (IDUs) in Athens and Bucharest. Infect Genet Evol 2015;35:109-21.
160. Dudas G, Rambaut A. Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. PLoS Curr 2014;6.

161. Faria NR, Sabino EC, Nunes MR, Alcantara LC, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med 2016;8:97.
162. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a Tool for Large Scale Molecular Epidemiology of HIV-1 and Other Rapidly Evolving Pathogens. Mol Biol Evol 2018;35:1812-9.
163. Laver T, Harrison J, O'Neill PA, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 2015;3:1-8.
164. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012;29:1969-73.
165. Bouckaert R, Heled J, Kuhnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10:e1003537.

2 Reducing HIV infection in people who inject drugs is impossible without targeting recently-infected subjects

This chapter was published in:

Tetyana I Vasylyeva, Samuel R Friedman, Jose Lourenco, Sunetra Gupta, Angelos Hatzakis, Oliver G Pybus, Aris Katzourakis, Pavlo Smyrnov, Timokratis Karamitros, Dimitrios Paraskevis, Gkikas Magiorkinis, Reducing HIV infection in people who inject drugs is impossible without targeting recently-infected subjects. AIDS 2016,30:2885-2890.

Summary of authorship:

TIV wrote this chapter, conducted phylogenetic analyses, constructed the transmission model, analysed the demographic data of the samples, and created all figures in this chapter. JL and SG helped with the compartmental model development and mathematical derivations. TK and DP composed the reference dataset. GM provided supervisory support. All others helped with the interpretations of the results.

2.1 Abstract

Although our understanding of viral transmission among people who inject drugs (PWID) has improved, we still know little about when and how many times each drug injector transmits HIV throughout the duration of infection. Here, we describe HIV dynamics in PWID to evaluate which preventive strategies can be efficient. Due to the notably scarce interventions, HIV-1 spread explosively in Russia and Ukraine in 1990s. By studying this epidemic between 1995 and 2005, we characterised naturally occurring transmission dynamics of HIV among PWID. We combined publicly available HIV *pol* and *env* sequences with prevalence estimates from Russia and Ukraine under an evolutionary epidemiology framework to characterise HIV transmissibility between PWID. We then constructed compartmental models to simulate HIV spread among PWID. In the absence of interventions each injector transmits on average to 10 others. Half of the transmissions take place within one month after primary infection, suggesting that the epidemic will expand even after blocking all the post-1st month transmissions. Primary prevention can realistically target the first month of infection and we show that it is very efficient to control the spread of HIV in PWID. Treating acutely infected on top of primary prevention is notably effective. Since a large proportion of transmissions among PWID happen within 1 month after infection, reducing and delaying transmissions through scale-up of harm reduction programs should always form the backbone of HIV control strategies in PWID. Growing PWID populations in the developing world, where primary prevention is scarce, constitutes a public health time bomb.

2.2 Introduction

The HIV pandemic has historically been dominated by sexual transmission. Transmission among Men who have Sex with Men (MSM) contributed to the virus' early spread in Western Europe and the Americas, while heterosexual transmission initially dominated in most of the developing world (e.g. Sub-Saharan Africa, South-Eastern Asia). Transmission among People Who Inject Drugs (PWID) has, to date, mostly posed a secondary burden worldwide, with the striking exceptions of Eastern European and Central Asian countries[1].

Much work has been done to study how HIV spreads in PWID populations and, as a consequence, we now have interventions that proved to be effective to slow down HIV epidemics[2]. Still, our knowledge of HIV transmission dynamics within PWID communities is limited in some aspects. For instance, we still do not know how fast the virus spreads and how much time after infection do we have to implement the evidence-based interventions before the virus is transmitted. Research to date has not measured forward transmission (i.e. number of transmissions that can be attributed to an infected PWID), and our knowledge is limited to the force of infection (i.e. the probability of a PWID getting infected over time)[3].

With almost one million HIV-infected individuals and a PWID population of around two million people[1], Russia and Ukraine provide a unique opportunity for studying HIV transmission dynamics in this transmission group. The epidemic in these countries has been accompanied by very limited harm reduction services[4] (represented mostly by small syringe exchange programs (SEPs) in the late 1990s and early 2000s), making it a large-scale "natural experiment" of HIV spread among PWID. By studying the conditions that allowed the fastest growth of an HIV epidemic among PWID, we can attempt to understand epidemics emerging in countries of the Middle

East and Northern Africa[5,6], South-Eastern Asia[7], Europe[8,9], and the USA[10]. Critically, while harm reduction approaches among PWID since the 1980s have had significant effects in the developed world[11], it remains unclear whether these preventive strategies should be prioritised over newer approaches such as treatment as prevention (TasP)[12] in resource-limited settings.

Since genome sequences of isolated viral strains contain information about how viruses spread in a community[13], evolutionary epidemiology has been increasingly used alongside other epidemiological methods. Viral genetic data have previously been used to estimate the basic reproductive number (R_0) for Hepatitis C (HCV)[14], and for some populations affected by HIV[15,16]. Phylodynamics and mathematical modelling have been previously combined in an integrated evolutionary epidemiology approach to estimate both R_0 and generation time (T), to show that onward HCV transmission from PWID in Greece was more likely during the early years of infection[17]. Since R_0 and the generation time of HIV have not previously been estimated together for PWID populations, we used the same approach to describe the transmission dynamics of the HIV epidemic among PWID in Ukraine and Russia that started in the mid-1990s. We then used the estimated R_0 to infer transmission rate coefficients and to inform simple compartmental models in order to assess what factors affect whether HIV epidemics become established in PWID.

2.3 Methods

2.3.1 Molecular evolutionary analysis

2.3.1.1 Nucleotide sequences and initial phylogenetic analysis

We compiled three HIV-1 sequence alignment datasets with sequences available from the Los Alamos HIV sequence database[18]. Firstly, a *reference dataset A* of 2199 (all

available) sequences of minimal length of 300 nucleotides (nt) from the *pol* region of HIV genome accessed in 2011 from HIV sequence database (sequence names are in Appendix 1). Sequences came from 36 countries (Afghanistan, Albania, Azerbaijan, Burkina Faso, Burundi, Benin, Democratic Republic of Congo, Chad, Congo, Cameroon, Cuba, Cyprus, Czech Republic, Ethiopia, France, Gabon, Georgia, Ghana, Equatorial Guinea, Kenya, Kazakhstan, Latvia, Mali, Nigeria, Russia, Rwanda, Sudan, Slovenia, Senegal, Togo, Tanzania, Ukraine, Uganda, Uzbekistan) sampled in 1985-2010. We aligned the sequences using MEGA 6.0 software[19] and then manually edited. We constructed phylogenetic trees with Mega 6.0 with simple evolutionary models to reduce computational complexity (i.e. assuming uniform rates among sites, using neighbour-joining approach with Tamura-Nei model[20], and same parameters were kept for other phylogenetic trees constructed during these analyses). Interestingly, the strains that were isolated from countries of the former Eastern block clustered together with high bootstrap support (>0.9). Given that previously published data suggested that the epidemic in Russia and Ukraine started in Odessa, Ukraine, and then expanded through the whole region[21,22], the following analysis was conducted for sequences from Russia and Ukraine together as coming from the same HIV epidemic. Consequently, all sequences from Russia and Ukraine from the *pol* region were sub-sampled from dataset A and sequences available from later years were added. This formed *dataset B* with 418 *pol* sequences from Russia and Ukraine sampled in 1997-2013.

Further, 92 sequences from Russia and Ukraine from the *env* region sampled in 1993-2011 were downloaded from the same database and formed the *dataset C*. Alignments were performed in Mega 6.0 and manually edited. Only one sequence from the same patient was included: if there was evidence that two sequences came from the same

person (based on the sequence names or the absence of genetic distance between the two sequences), one of them was deleted from the dataset. Accession numbers of all sequences are in Appendix 1.

Both of the phylogenetic trees (constructed with *env* and *pol* alignments) suggested that the subtype A epidemic in Russia and Ukraine was introduced once for both of these countries and thus allowed us to run all further analyses considering them to be part of the same epidemic.

2.3.1.2. Phylodynamic analysis

We ran population dynamics analyses using the Bayesian Evolutionary Analysis Sampling Trees (BEAST) 1.8.1 software[23]. Initially, we analysed dataset C (*env*) creating a subgroup of sequences that had both *env* and *pol* genes. We estimated the time to most recent common ancestor (TMRCA) for this group and used it as a prior on TMRCA to inform the analysis of dataset B (*pol*). We used the GTR nucleotide substitution model and assumed sites heterogeneity with Gamma distribution with 4 invariant sites. We applied the lognormal uncorrelated relaxed clock model to account for the evolutionary rate variation between branches[24]. We used a Bayesian Skyline plot model with 10 intervals as a coalescent tree prior, which allowed us to estimate the product of an effective population size (*Ne*, corresponds to the number of individuals in a population who produced offspring) and a generation time (*T*, the expected time from acquiring infection to onward transmission).

We ran a Markov chain Monte Carlo (MCMC) for the datasets B and C separately: 100,000,000 generations for the dataset C with *env* sequences (burn in 20 x 10^6 generations), and two separate 100,000,000 generations runs for the dataset C with *pol* sequences (burn in 20 x 10^6 and 40 x 10^6 generations). Trees were sampled every 1000 generations. We used the LogCombiner program from the BEAST 1.8.1

package to combine the two runs from *pol* region, and the program Tracer from the same package to check for convergence and determine whether appropriate mixing of the MCMC sampler had been achieved in the posterior target distribution (effective sample size>100).

We obtained annual estimates of HIV prevalent cases for all the available years (1990-2011) from the UNAIDS website[25].

2.3.2 Estimates of epidemiological parameters

We estimated the number of secondary infections per primary infection in a completely susceptible population assuming that the population is large enough to follow a deterministic Susceptible-Infected-Removed model[26]:

$$R_0 = \frac{1}{\varpi}\left[\frac{1}{t} ln \frac{N_e(t)}{N_e(0)}\right] + 1 \qquad (2.1)$$

where $N_e(t)$ is the effective population size at time t ($N_e \times T$ was estimated from the skyline plot reconstructed from *env* sequences), $N_e(0)$ is the effective population size at the baseline of the exponential growth phase, and ϖ is the average removal rate.

For our model we estimated the average removal rate as:

$$\varpi = \frac{jq}{j+q} \qquad (2.2)$$

where j is the rate of progression from the recent to long-term phase of infection (1/0.5 a year, 2.0), q is the overall mortality rate of HIV infected individuals (in the absence of prevention assumed to average 1/10 years, 0.1). Thus, the number of secondary infections per primary infection in our model is described as:

$$R_0 = \frac{j+q}{jq}\left[\frac{1}{t} ln \frac{N_e(t)}{N_e(0)}\right] + 1 \quad (2.3)$$

Further, the ratio $N(t)/N_e(t)$ represents the variation in offspring numbers among individuals[27,28]:

$$\mathrm{var}(Z) = \frac{N(t)}{N_e(t)} \quad (2.4)$$

where var(Z) is the variation in the number of secondary infections per individual (R_0), and N(t) is the estimate of HIV prevalence at time t.

The phylodynamic transmission parameter (*PTP*) has been previously introduced[17] as a value that reflects the transmissions within a population by integrating generation time (T) and the variance in the number of secondary infections (*var(Z)*):

$$PTP = \frac{\mathrm{var}(Z)}{T} \quad (2.5)$$

thus, using equations (2.4) and (2.5) we get:

$$\mathrm{var}(Z) = \left(\frac{N(t)}{N_e(t)\times T}\right) \times T = PTP \times T \quad (2.6)$$

As it has been previously shown in [17], in a system where a proportion u transmits while the rest ($1 - u$) is not transmitting, the distribution of the number of secondary infections Z in the whole population (transmitters and non-transmitters; PWID and general population in our case) can be assumed to be approximated by a zero-inflated Poisson distribution such as the mean number of secondary infections per transmitter

($R_{0,a}$) (per an injector) is related to the overall mean (expected) number of secondary infections in the population (R_0) by:

$$R_0 = uR_{0,a} \tag{2.7}$$

while the variance of secondary infections would be provided by:

$$\mathrm{var}(Z) = uR_{0,a}[1 + (1-u)R_{0,a}] \tag{2.8}$$

Thus, from equations (2.6) and (2.8) we estimate the generation time T (the expected time from initial HIV infection (index case) to transmission to other individuals (secondary infections)) to be

$$T = uR_{0,a}\frac{1+(1-u)R_{0,a}}{\frac{N(t)}{N_e(t)\times T}} \tag{2.9}$$

We used the exponential phase of the epidemic growth for calculations of generation time as it is expected to be nearly constant over this period.

All parameters and their explanations are in Table 2.1.

Table 2.1. A summary of parameters used for the epidemiological transmission dynamics estimates and for the following compartmental modelling. "Source" column indicates how the values were obtained. "Value" column reflects the values for the outbreak scenario with recently-infected individuals in treatment at the start of the simulation (see text for the values for other scenarios); if N/A is stated – values change through time or are non-numerical; rates correspond to individuals/year. BSP – Bayesian Skyline Plot.

Para-meter	Explanation	Source	Value
c	Duration of recent phase	Bellan *et al*[31]	0.5 year
h	Proportion of recently infected individuals on treatment	Assumed	0.25
I_c	Proportion of long-term infected individuals	Assumed	0.0
I_r	Proportion of recently infected individuals	Assumed	0.0001
I_t	Proportion of individuals on treatment		
j	Rate of progression from the recent to the long-term phase of infection	Bellan *et al*[31]	2.0
k	Proportion of long-term infected individuals on treatment	Assumed	0.5
m	Rate with which recently infected individuals are engaged into treatment	Assumed	4.0
n	Rate with which long-term infected individuals are engaged into treatment	Assumed	0.25
N(t)	Estimate of HIV prevalence at time *t*	UNAIDS[25]	N/A
$N_e(0)$	Effective population size at the baseline of the exponential growth phase	BSP	N/A
$N_e(t)$	Effective population size at time t	BSP	N/A
p	Duration of infectivity	Longini *et al*[36]	10 years
PTP	Phylodynamic transmission parameter	Magiorkinis *et al*[17]	N/A
q	Overall mortality rate of HIV infected individuals	Longini *et al*[36]	0.1
R_0	Number of secondary infections	Estimated	7.0
$R_{0,a}$	Mean number of secondary infections per transmitter (per an injector)	Estimated	10.11
S	Proportion of individuals susceptible to HIV in a PWID population	Assumed	0.9999
T	Generation time	Estimated	30 days
u	Proportion of transmitters in HIV-infected population	Estimated	0.7
β_1	Transmission rate coefficient that determines virus transmissibility from individuals recently infected with HIV	Bellan *et al*[31]	$5.3x\beta_2$
β_2	Transmission rate coefficient that determines virus transmissibility from long-term infected individuals	Bellan *et al*[31]	N/A
μ	Rate at which HIV individuals on treatment die	Assumed	0.014
var (Z)	Variation in R_0	Derived	N/A

2.3.3 Compartmental Modelling

We used compartmental Susceptible-Recently Infected-Chronically Infected (Susceptibles – individuals that are not infected, but are at risk of infection; Recently Infected – individuals that acquired infection within the last 6 months; Chronically Infected – individuals that acquired infection longer than 6 months ago) model to describe dynamics of HIV epidemic among PWID in Russia and Ukraine. We used R-project software to estimate epidemiologic parameters given different rates of moving from one compartment to another. Deterministic compartmental modelling has been shown to be effective for epidemiological parameters estimation for epidemics with $R_0>1$ and large susceptible populations[29], which is the case for our settings. In our model we acknowledge different infectivity during a "recent" (within the first six months after infection) and a "long-term" (after the first six months after HIV acquisition) infectivity periods. First, we constructed a compartmental model of a natural epidemic without treatment option available (Fig. 2.1).

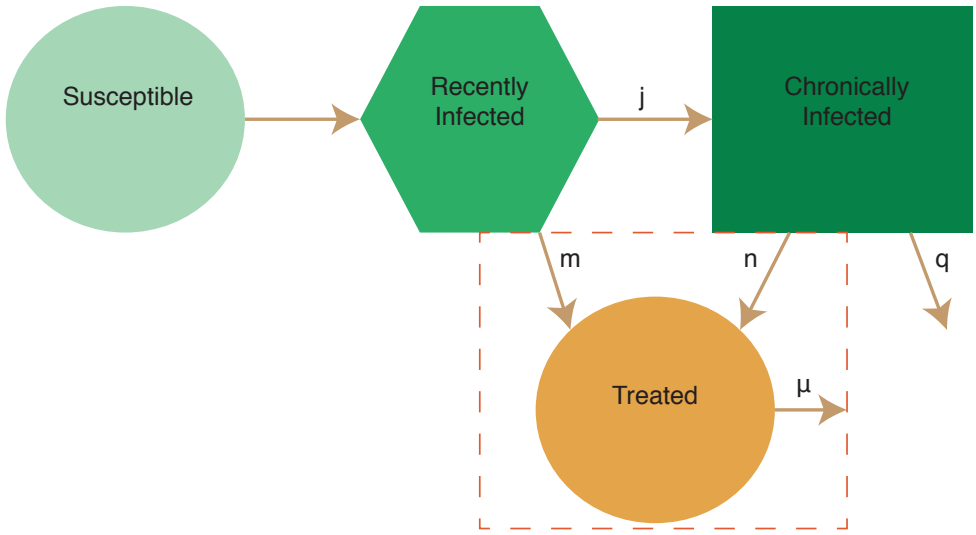


Figure 2.1. Compartmental model that describes natural HIV progression with and without a treatment compartment (the treatment compartment is highlighted with a

red dotted line in the figure). Each box represents compartments within a hypothetical PWID population. The symbols stand for the rates with which individuals move from one group (compartment) to another: j – rate of progression from recently infected to long-term phase of infection; q, μ – removal rates that indicate mortality of HIV infected individuals (or quitting drug use) with untreated and treated HIV infection, respectively; m, n – rates of involvement in treatment of recently and long-term infected individuals respectively.

This model predicts the number of infected and susceptible individuals over the natural course of the epidemic, for 20 years after HIV introduction into a population of PWID and is described by the following differential equations:

$$\frac{dS}{dt} = -I_r\beta_1 S - I_c\beta_2 S + qI_c \quad (2.10)$$

$$\frac{dI_r}{dt} = I_r\beta_1 S + I_c\beta_2 S - jI_r \quad (2.11)$$

$$\frac{dI_c}{dt} = jI_r - qI_c \quad (2.12)$$

where S is the proportion of individuals susceptible to HIV in the PWID population, I_r is the proportion of recently infected individuals (<6 months since infection), I_c is the proportion of long-term infected individuals(>6 months since infection). β_1 is a transmission rate coefficient that determines virus transmissibility from individuals recently infected with HIV, whilst β_2 determines virus transmissibility from individuals who have been infected for at least 6 months; j is a rate of progression from the recent to long-term phase of infection (if c is the duration of recent phase and c=0.5 years, then j=1/c=1/0.5=2); q is a removal rate that indicates both the overall

mortality of HIV infected PWID with untreated infection and the rate they quit injecting drugs and are not engaging in risky injecting behaviours anymore (if p is the duration of infectivity and p=10 years, then q=1/p=1/10=0.1).

We have calculated the R(t) which is the number of secondary infections per infected individual from the natural epidemic model using next-generation matrix[30]. The eigenvalue λ of the matrix FV^{-1} represents the R_0 for the model. In this approach the matrix F represents the rate of appearance of new infections in a compartment, which for our model can be described by the matrix:

$$F = \begin{pmatrix} \beta_1 S & 0 \\ \beta_2 S & 0 \end{pmatrix} \quad (2.13)$$

Similarly, the matrix V represents the rate of transfer of individuals into a compartment by other means (i.e, recently-infected individuals transfer to chronically-infected compartment after 6 months) and is described as:

$$V = \begin{pmatrix} -j & j \\ 0 & -q \end{pmatrix} \quad (2.14)$$

Knowing the F and V allowed us to estimate the reproduction number in our population as:

$$\lambda FV^{-1} = R(t) = \frac{q\beta_1 + j\beta_2}{jq} \mathrm{S}(t) \quad (2.15)$$

where S(t) is the change in the compartment of susceptible individuals over time. We have used the value of $R_{o,a}$ estimated from the genetic data (10.11 for 10 years of

infectivity (p, q=1/p=0.1) and 6 months duration of the recent phase (c, j=1/c=2)) to estimate the transmission rate coefficients (β_s). We have also used the data from another study that showed that transmission hazard during the recent period is 5.3 (95% CI 0.79-57) times higher compared to the long-term infection period (β_1/β_2)[31]. This estimate comes from a study of an epidemic driven by sexual transmissions among heterosexual individuals, but was used in our model as there is no similar estimate for PWID transmissions. To estimate the uncertainty of our model we have estimated the credibility intervals for the model compartments using the 95% credibility intervals from the skyline plot estimated from the *env* sequences. We have also conducted sensitivity analysis on the difference in β_s.

To get the estimates of different compartments of the population throughout time we used rates known from the literature. The removal rate q is calculated by the infectivity period p (assuming the natural history of HIV infection in absence of treatment the time to AIDS/death p ~10 years, thus 10 years from the moment of infection to removal, q=1/10) and the transition rate j is calculated by the duration of recent infection phase c ~ 0.5 years, j=1/0.5=2)

We have also assumed that the removal rate (infected individual dying from any cause or quitting using drugs) is equal to the rate at which individuals become susceptible (start drug use in this case); consequently, we assumed that population size is constant throughout time. As the input parameters for the first model, we assumed that the proportion of recently infected individuals in the population at initial time point was 0.0001, consequently, the proportion of susceptibles was 0.9999.

We further advanced this model to include the compartment of treated individuals (highlighted with a red dotted line in Fig. 2.1). This model predicts the impact of treatment provided to individuals at different stages of disease (this model was

applied to a new epidemic/outbreak and also to an established epidemic) and is described as:

$$\frac{dS}{dt} = -I_r\beta_1 S - I_c\beta_2 S + q\, I_c + \mu I_t \quad (2.16)$$

$$\frac{dI_r}{dt} = I_r\beta_1 S + I_c\beta_2 S - jI_c - m\, h\, I_r \quad (2.17)$$

$$\frac{dI_c}{dt} = jI_r - qI_c - nkI_c \quad (2.18)$$

$$\frac{dI_t}{dt} = mhI_r + nkI_c - \mu I_t \quad (2.19)$$

where S is the proportion of individuals susceptible to HIV in the population, I_r is the proportion of recently HIV infected individuals, I_c is the proportion of long-term infected individuals, I_t is the proportion of individuals on treatment. β_1 is a transmission rate coefficient that determines virus transmissibility from individuals recently infected with HIV, whilst β_2 determines virus transmissibility from individuals who have been infected for more than 6 months; j is the rate of progression from recent to long-term phase of infection; m is the rate with which recently infected individuals are engaged into treatment; n is the rate with which long-term infected individuals are engaged into treatment; h is the proportion of recently infected individuals on treatment; k is the proportion of long-term infected individuals on treatment; μ is the rate at which HIV individuals on treatment die.

For the model with treatment compartment, μ is calculated assuming that the life expectancy of an HIV-infected person with suppressed viral load would be similar to the general population in Russia and Ukraine in 1990s (70 years, μ =1/70). To simplify, we also assume that those individuals that enter a treatment program, stay in treatment, have undetectable viral load and do not infect susceptible individuals. In

reality, individuals might drop the antiretroviral treatment program or the efficiency of the antiretroviral treatment might be suboptimal. Here we do not explore the effect of treatment drop-outs and virological failures in interventions; thus, the treatment coverage that we are modelling is the ideal scenario. Current HAART regimens achieve virological response in 80-90% of the individuals in treatment in developing countries[32-34], thus we would normally expect higher than the modelled coverage to be required in real-life interventions (i.e, 56-63% instead of 50% in the current model) to observe the dynamics suggested by the model.

We modelled several scenarios for a new epidemic/outbreak and for a well-established epidemic. First, we have compared scenarios when none of the recently infected individuals are on treatment (h=0) and 50% of long-term infected individuals receive treatment within the first 4 years after they got infected (n=0.25, k =0.5) with the scenario when 25% of recently infected individuals and 50% of long-term infected individuals receive treatment (within the first 3 months and the first 4 years, respectively; h=0.25, m=4.0, n=0.25, k=0.5). As the initial parameters, for the settings with a new epidemic/outbreak, we assumed that the proportion of recently infected individuals in the population was 0.0001 and the proportion of susceptible was 0.9999. For the scenario of advanced epidemic we assumed that the PWID population consists of 0.01% of recently infected, 10% of individuals on treatment, 20% of long-term infected individuals not on treatment, and the rest 69.99% are susceptible. Then we also introduce a scenario, when the $R(t)$ is reduced by 60% through the introduction of preventive interventions, thus, reducing the β_s by 60%.

For all the scenarios we have conducted sensitivity analysis on the proportion of individuals on treatment.

2.4 Results

2.4.1 Phylogenetic analysis

The phylogenetic tree, which included globally representative subtype A reference strains, shows that the strains from the epidemic in Russia, Ukraine and other countries of the former Eastern bloc formed a distinct clade within the global HIV-1 epidemic (Fig. 2.2). This supports previous reports that the Russian and Ukrainian epidemics were isolated from the rest of the world and spread as a joint epidemic[21,22]. We thus integrated molecular and epidemiological data from these areas and analysed them as a single epidemic.

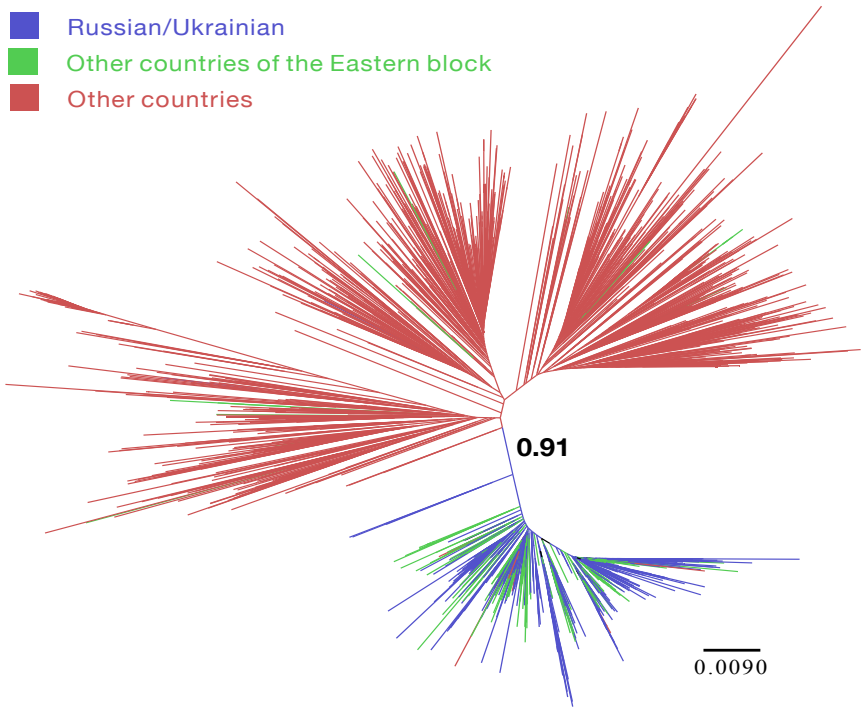


Figure 2.2. The phylogenetic tree constructed using the reference dataset A of pol sequences. The colours represent the country of sampling. The value at the node represents the bootstrap support for the clade comprised of sequences from Russia, Ukraine, and other countries of the former Eastern bloc. The bar indicates the equivalent of the genetic distance in substitutions/site/year.

According to both the skyline plot and the estimated HIV prevalent cases curve, the exponential phase of the epidemic began around 1995 and ended around 2005 (Fig. 2.3). The skyline plot estimated from the pol sequences showed a dramatic drop and rebound in $N_e \times T$ around 2002-03 (Fig. 2.3).

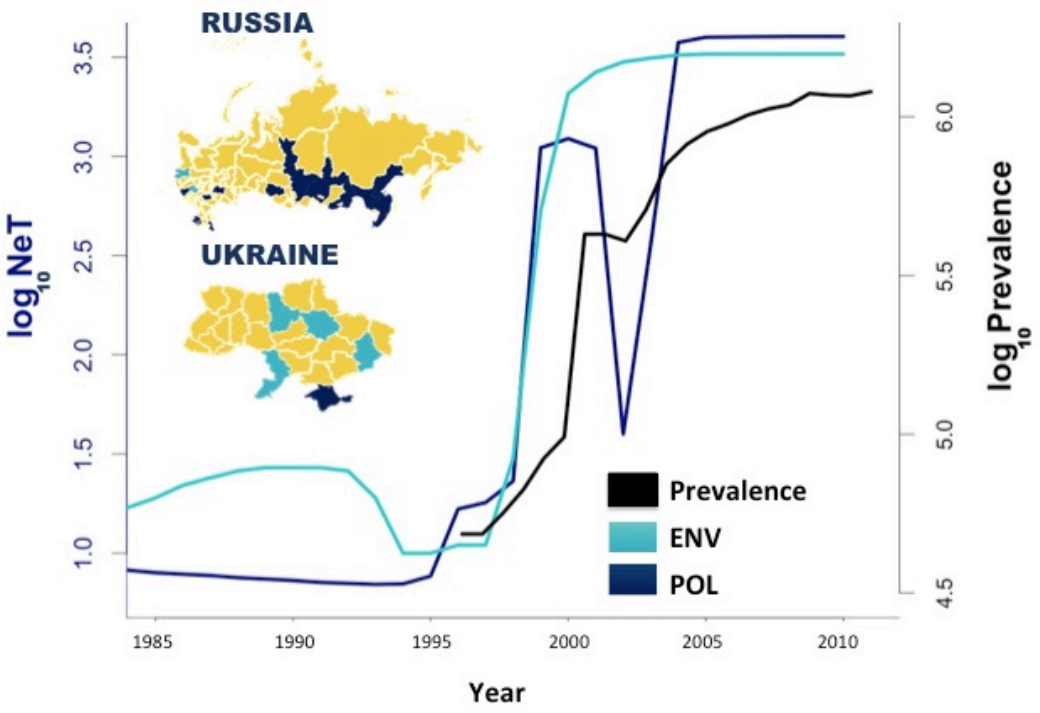


Figure 2.3. Bayesian Skyline plot estimated from pol (dark blue line) and env (cyan line). Maps show representation of sequences by geographical regions within the countries. The UNAIDS estimates of HIV prevalent cases are shown in black line.

The drop was significant given the 95% highest posterior density (HPD) intervals and appeared in Skyline plots reconstructed from all three random sub-samples of *pol* sequences (data not shown); it was not, however, apparent in the skyline plot estimated from *env* sequences. We concluded that the discordance among *env* and *pol* skylines could be attributed to their geographic distribution, with sequences from *pol*

being more representative of the Russian and Ukrainian PWID population than sequences from *env* (19 vs 7 geographic regions, respectively, see map on Fig. 2.3). We estimated the *PTP* to be 421.5 (95% CI 248.6-594.4) through a linear regression of the estimated number of prevalent cases for Ukraine and Russia over the median $N_e \times T$. We then used *PTP* together with the changes in the effective population size to calculate the R_0 and generation time *T* as described above and previously reported[17].

2.4.2 Phylodynamic and epidemiological parameter estimates

We calculated $R_{0,a}$ using different values of the duration of infectivity *p* (5-15 years), where $1/p=q$ is the overall mortality rate of untreated HIV infected PWID (Fig. 2.4).

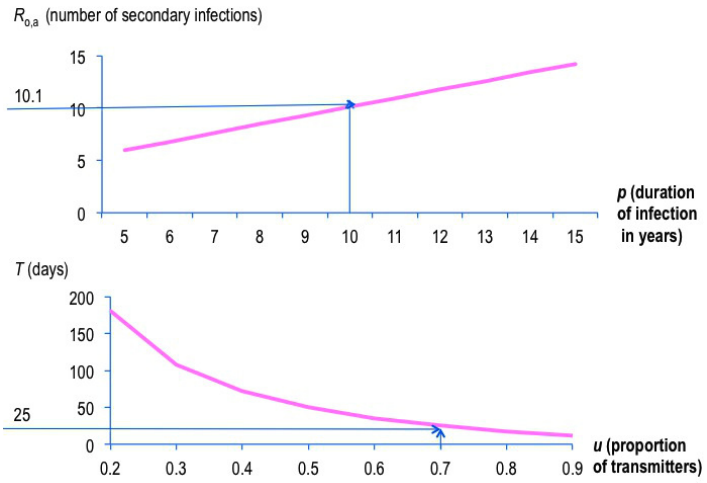


Figure 2.4.Estimates of the epidemiologic parameters. Top: $R_{0,a}$ values given different duration of infectivity. Bottom: Estimates of the range of generation time values given different proportion of transmitters, assuming the duration of infection to be 10 years and $R_{0,a}$ to be 10.1.

$R_{0,a}$ estimates changed from 6 for p=5 to 14 for p=15. Since HAART was introduced in 1996, treatment coverage of PWID in Russia and Ukraine remained low (only 2% of HIV infected PWID received treatment in early 2000s[35]). Thus, a plausible duration

of infectivity for an untreated HIV-infected PWID at that time is unlikely to be more than 8-10 years[36], suggesting that the epidemic in PWID had an R_0 in the range 8-10.

We assumed that the majority of sexual (including both MSM and heterosexual) transmissions among non-PWID gave rise to a negligibly low number of secondary infections compared to PWID. Heterosexual transmissions were mostly through the sex trade, which is highly correlated with drug use in Russia and Ukraine[37]. It is difficult to quantify MSM transmission since MSM contacts were highly stigmatized in Russia and Ukraine and were underreported in the surveillance system[38]: in 1990s infections attributed to both MSM and heterosexual contacts were coded and reported together. We considered $T<10$ days as biologically implausible (viral load is undetectable)[39].

One hypothesis for the explosive growth of the HIV epidemic in Russia and Ukraine is that a particular practice—using blood as a buffer/purifier in the process of domestic drug production—was popular among some PWID[37]. Contamination of large quantities of drugs could facilitate the spread of infection and would allow transmission to multiple recipients generated by a small proportion of PWID. This would translate, within our model framework, into a small proportion of transmitters (those "supplying" blood) driving an epidemic characterised by a very long generation time (around a year, Fig. 2.4). Translating these dynamics in real-life data means that single individuals should be providing blood for 1 year to purify drugs that were used by 70 PWID. Such a scenario is clinically unlikely due to the large amount of blood that a single individual would need to provide within a long period of time.

Thus we suggest that the most plausible scenario for this HIV epidemic is one characterised by a very short generation time (< 1 month), a large proportion of transmitters (70%) and a high number of secondary infections (7 new infections per

infected individual). This is compatible with the hypothesis that unsafe injections practiced by a large proportion of PWID (40-90% of Russian/Ukrainian PWID shared instruments or front/backloaded syringes in 1995-2005[40-42]) may have forced the rapid epidemic spread, which also agrees with previously reported findings[43]. The fact that a very short generation time was accompanied by a very high value of R_0 is also biologically plausible, since the first month after the infection is the period of the highest infectivity.

2.4.3 Modeling the establishment of HIV epidemics in PWID

In the natural epidemic model without treatment or prevention, HIV prevalence among PWID was predicted to reach 86% (credibility interval (CI) 81-88%) after 20 years (Fig. 2.5). The proportion of transmissions attributed to recently infected individuals rapidly decreased over the first years of the epidemic and stabilized at around 20% after 7 years (Fig. 2.6).

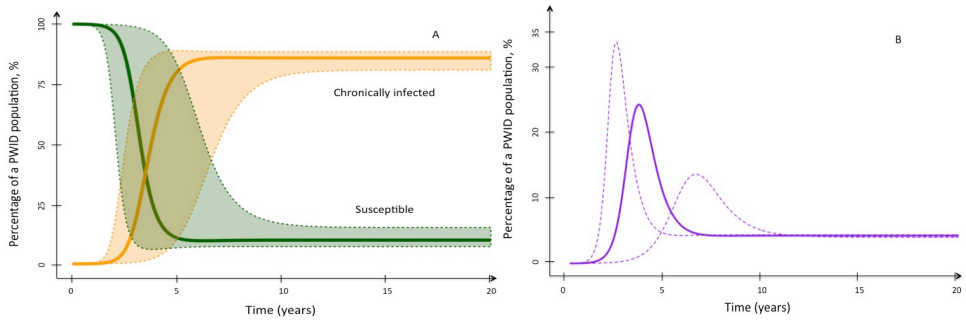


Figure 2.5. The results of the natural epidemic model: A) The dynamics in the compartments of susceptible (green) and chronically infected (yellow); B) The dynamics in the compartment of recently infected (purple). Solid lines indicate mean estimates, dashed lines represent credibility intervals. For the initial parameters, we assumed that the proportion of recently infected individuals in the population was 0.0001 and the proportion of susceptibles was 0.9999. The credibility intervals estimated from the Bayesian skyline plot constructed from the env sequences.

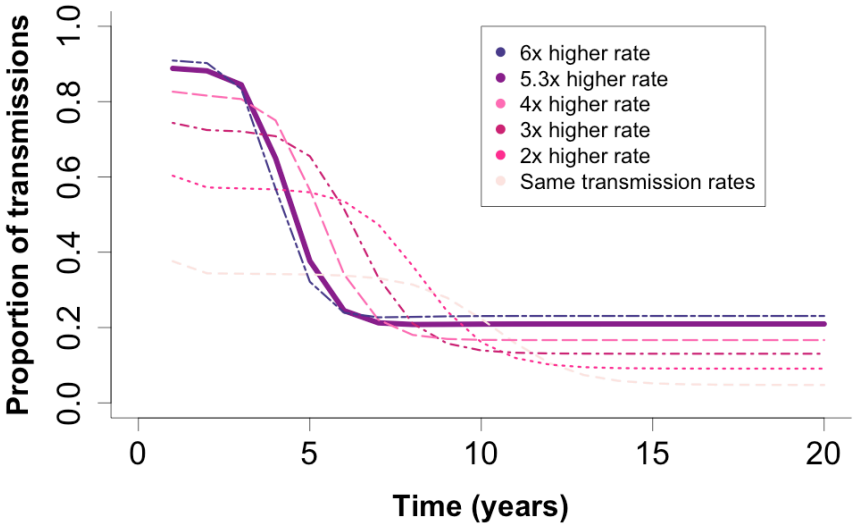


Figure 2.6. The proportion of new HIV infections in the population attributed to recently infected individuals in the natural epidemic model. We assumed different transmission rates: from the same transmission rate during the recent and the long-term phase of the infection to the 6 times higher transmission rate during the recent phase. In our model we assumed that the transmission rates is 5.3 times higher during recent infection.

In absence of prevention, but with treatment, the model showed that HIV prevalence would still reach 83% (CI 73-87%) after 20 years if 50% of long-term infected individuals receive treatment on average 4 years after getting infected (Fig. 2.7, A). If treatment was provided to a small proportion (25%) of recently infected individuals, after 20 years we did not observe a dramatic reduction in prevalence (72%, CI 62-80%) (Fig. 2.7, B).

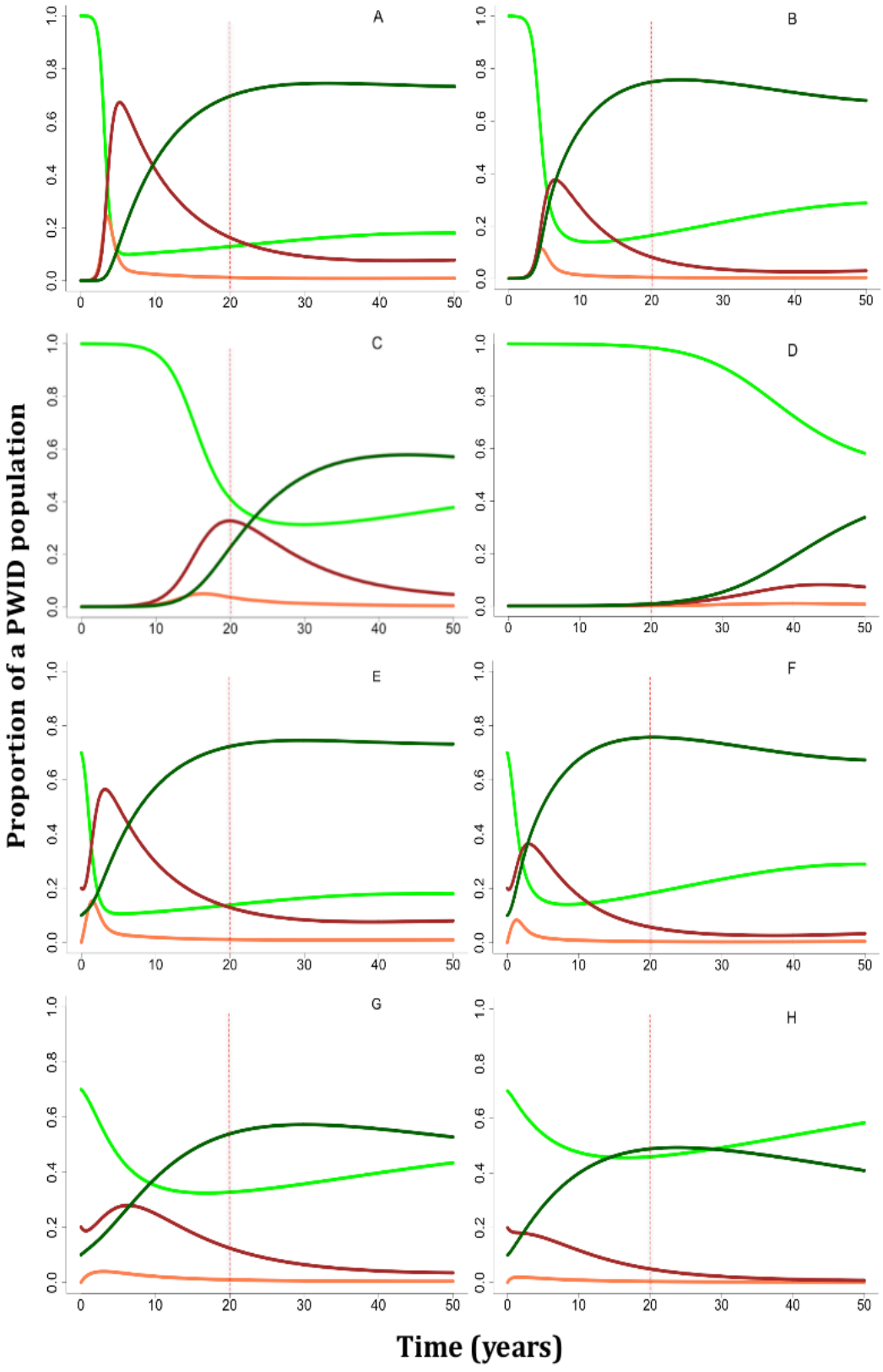

A
B
C
D
E
F
G
H
Proportion of a PWID population
Time (years)

Figure 2.7. Results of the compartmental model representing HIV spread under different treatment scenarios. Lines represent changes in different compartments of the model: orange – recently infected individuals, dark red – long-term infected individuals, light green – susceptible individuals, dark green – individuals in treatment. The vertical dashed red line indicates 20 years into an epidemic. Panels A-D represent HIV spread during an outbreak. Panels A and B represent scenarios when 0 and 25% of recently infected individuals, respectively, and 50% of long-term infected individuals (for both A and B) are provided with treatment while no other preventive services are suggested. Panels C and D represent scenarios when, additionally to providing treatment, other harm reduction interventions are introduced, so that $R_{0,a}$ is reduced to 4. Panels E-H represent HIV spread in a well-established epidemic. Panels E and F represent scenarios when 0 and 25% of recently infected individuals, respectively, and 50% of long-term infected individuals (for both E and F) are provided with treatment while no other preventive services are suggested. Panels G and H represent scenarios when, additionally to providing treatment, other harm reduction interventions are introduced, so that $R_{0,a}$ is reduced to 4.

We next investigated the scenario in which R_0 was reduced from the very beginning of an outbreak by scaling up harm reduction interventions and changing risky behaviors. By reducing $R_{0,a}$ by 60% (studies show that harm reduction programs can help to reduce risky practices by 60%[44,45] and incidence in PWID by 80%[46]) to $R_{0,a}$ = 4, we were able to prevent the epidemic from establishing, provided that treatment was offered to at least 25% of recently-infected PWID (Fig. 2.7, D). There was no intervention scenario to prevent the epidemic from being established when treatment

was not provided to recently infected individuals (Fig. 2.7, C). For the scenarios with the well-established epidemic (30% baseline HIV prevalence) (Fig. 2.7, panels E-H) the best-case scenario was also to treat at least a small proportion of recently-infected individuals while scaling up prevention.

2.4.4 Sensitivity analyses

We conducted sensitivity analyses to test the robustness of our estimates against assumptions on infectivity and the difference in transmission rate coefficients (β_s in Fig. 2.1). The infectivity period p influenced the estimates significantly changing $R_{0,a}$ from around 6 secondary infections per infected individual assuming p=5 years to around 14 secondary infections per infected individual for p=15 years (Fig. 2.4). We chose p = 10 years as the most plausible scenario of natural history of progression towards AIDS for settings with lack of prevention and limited treatment options (as Russia and Ukraine in late 1990s)[35]. Additionally, p<5 years resulted in a biologically implausible generation time (10 days or less) given that the proportion of transmitters was more than 70%.

For the compartmental modelling, we also tested for sensitivity against assumptions on the infectivity period and it was not sensitive for $p \geq 4$. Furthermore, the sensitivity analysis on the changes of the transmission rate coefficients throughout the infectivity period showed that even if the transmission rate coefficient during the recent phase of infection is only slightly higher or equal to the rate coefficient throughout the remainder of infectivity period, the proportion of new infections attributable to recently infected individuals stabilized at around 5% after 15 years (Fig. 2.6). As expected, the proportion of transmitters changed the estimates of generation time: the bigger the proportion of transmitters – the lower the generation time (Fig. 2.4).

Additionally, we have tested a renewal of the PWID population in the compartmental model (Fig. 2.8). More specifically we have modelled a scenario where individuals in antiretroviral treatment quit drug injection with a rate of 1 every 20 person-years. We then add new injectors to the compartment of "Susceptibles" at the same rate.

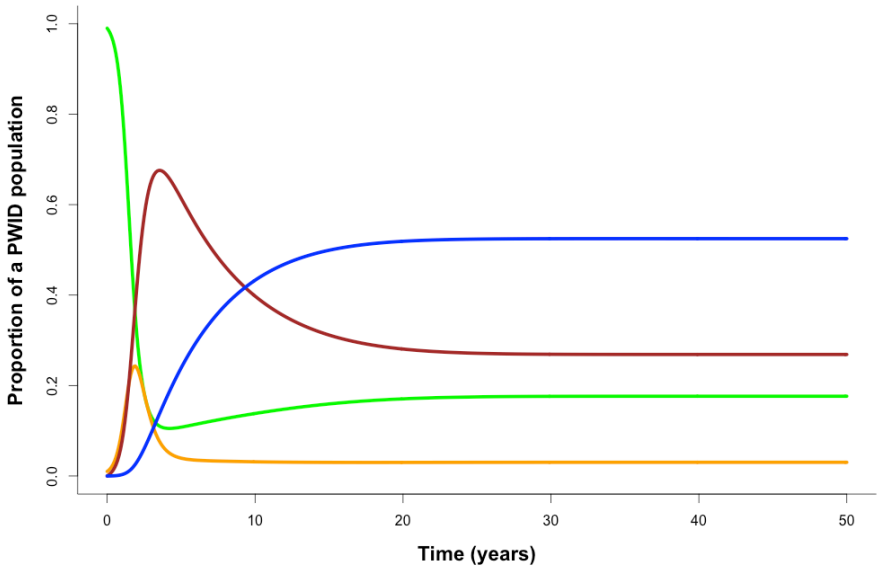


Figure 2.8. Results of the compartmental model representing HIV spread under a treatment scenario when the PWID population is renewed (quit under antiretroviral and recruit in the Susceptibles). Lines represent changes in different compartments of the model: orange – recently infected individuals, dark red – long-term infected individuals, light green – susceptible individuals, blue – individuals in treatment. In this scenario, 0% of recently infected individuals and 50% of long-term infected individuals (similar scenario to Fig. 2.7, A) are provided with treatment while no other preventive programs are introduced. Additionally, the susceptible PWID population is renewed as a result of quitting injection whilst on antiretroviral treatment: a) PWID quit injecting when introduced and sustained in antiretroviral treatment with a rate of 1 PWID quitting every 20 person-years of treatment, b) susceptible injectors are recruited with the same rate.

In this scenario, 0% of recently infected individuals and 50% of long-term infected individuals are provided with treatment while no other preventive programs are introduced, similarly to scenario presented in the Figure 2.7, A. In this case, the dynamics and prevalence stay the same, but the proportion of people in compartments of "Infected" and "Treated" changes. Thus, after 20 years 52% of PWID are expected to be in treatment; 28% of them chronically and 2% recently-infected, not receiving ART; summing up the overall prevalence in the population at 82%, compared to 83% (CI 73-87%) in our original scenario with no drug injection quitting rate.

2.4.5 Limitations: Assumptions on network effects

One of the basic assumptions of compartmental modelling is that it assumes panmictic populations – populations with random mixing within compartments, such that individuals have no preferences when choosing partners and establishing connections that can lead to transmissions. However, it is well documented that populations of PWID have network structure and a social network approach might provide a more realistic description of the epidemic dynamics of spread among PWID[47]. In our modelling we assume constant rate at which recent and long-term infected individuals come in contact with susceptible individuals over time. The long-term infected individuals after some time have lower infectiousness, which under a panmictic model reduces the transmission rate as the probability of contact between a susceptible and a highly infectious recently infected is reduced, creating a suboptimal saturation/protection of the epidemic spread. Under some network settings this might however not be true. Some network structures might allow only contact between susceptibles and long-term infected individuals and, thus, prevent susceptible individuals from contact with the highly infectious acutely infected creating a "firewall" effect[48-50].

We believe that this effect is less likely to bias estimates that come from a very large nation-wide epidemic that involves more than 1,000,000 individuals, as higher and/or lower transmission rates would be averaged over a large sample. However, further research is needed that will aim at accounting for the network structure of a population when using this method to describe HIV transmission dynamics in PWID.

2.5 Discussion

We analysed the uncontrolled and extensive HIV epidemic in Russia and Ukraine that began in the mid-1990s with an evolutionary epidemiology approach. We estimated that at the time each injector transmitted to at least 5 individuals within the first month after getting infected. We thus showed that in the absence of prevention HIV transmissions in PWID happen rapidly after primary infection.

Can we realistically target the first month of the infection? With primary prevention, such as SEPs and safe practices trainings, we intervene from the very first day of infections. Since the infectivity drops after the first month as a result of lower viral load and established immune responses, if we were to increase generation time, we would reduce the probability of infection per contact; we would then have more time to intervene with treatment.

Indeed we show that TasP has a place in controlling an outbreak only when we introduce large-scale primary prevention and appeared to be more effective when we targeted recently-infected individuals. Other studies show the importance of recently infected individuals in sexually driven HIV epidemics[51-53], but no consensus was reached on whether aiming HIV prevention efforts specifically at the recently infected group would be game-changing[54,55]. While such evidence for PWID is largely

missing, we show that prevention of early transmissions provides major gains when we reduce transmissibility by reducing risky injecting practices.

Our findings may provide insights on the dynamics of recently emerging HIV outbreaks such as those in Athens[8], Romania[9], or Indiana[56]. Crucially, HIV epidemics in PWID have been recently emerging in resource-limited countries in the Middle East and North Africa (particularly Afghanistan, Egypt, and Morocco)[5], as well as in some Southeast Asian countries (like Indonesia and Philippines)[7,57]. PWID populations in these locations were historically small, but their growth created the potential for new HIV outbreaks. Our study suggests that if primary prevention is not introduced in countries with increasing prevalence of drug use, the epidemics in this group will expand rapidly when started and remain uncontrolled until sufficient primary prevention is put in place. Given that intravenous drug use is increasing faster than preventive responses in the developing world, it is a matter of time that these public health time bombs will explode again.

Changes in drug-taking behaviour might also influence transmission dynamics and introduce new pools of susceptible individuals within PWID populations. In our results, the drop of N_eT around 2002 might be partly attributed to the shift in the Russian-Ukrainian drug scene towards stimulant drugs (associated with risky injecting behaviours[58]). Thus, settings with changing drug scenes, including changes in drug of choice or demographic changes, should also be addressed with basic prevention scale-up.

Our modelling suggested that increasing SEP coverage and reducing risky injecting practices in addition to treatment provision to half of the infected PWID will yield an additional 15% decrease in HIV prevalence after 20 years. Another modelling study that was informed by data from Ukraine showed that a 40% reduction in HIV

prevalence within 10 years period should be expected if at least 60% of the unmet needs of harm reduction services (including SEPs and opioid substitution treatment) and treatment were covered[59], but this study did not model differences in transmissibility among recently and chronically infected. Our modelling suggested that such a steep decrease in HIV prevalence can be obtained only if we additionally provide at least 25% of recently infected individuals with treatment.

Apart from epidemiological reasons, early treatment initiation provides clear benefits for the patients themselves[60,61]. Thus, a combination of the expansion of existing harm-reduction interventions and the development of new approaches to identify and treat recently infected people should be applied to settings with PWID-driven HIV epidemics[62]. Unfortunately, reaching those recently infected is extremely challenging, especially in settings where preventive responses have been limited due to the socio-political context in which the outbreak arises. In the 1990s social stigmatization of PWID in Russia and Ukraine was reflected in governmental and health care sectors, which limited the implementation of harm reduction services[4,63]. Other developing countries experiencing HIV outbreaks among PWID face similar problems; for example, despite decriminalization of drug use, police harassment of PWID in Indonesia undermines the success of SEPs[64]. Political support to reduce stigmatization of PWID could significantly increase access to health services and mitigate HIV spread in this vulnerable group.

2.6 References

1. UNAIDS. UNAIDS report 2014: The Gap Report. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS; 2014.
2. MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and Hepatitis C in people who inject drugs: a review of reviews to assess evidence of effectiveness. Int J Drug Policy 2014;25:34-52.
3. Thomas M, Sahu D, Raj Y, Pandey A. A Probability Model for Estimating the Force of Transmission of HIV Infection and Its Application. American Journal of Mathematics and Statistics 2014;4:171-7.
4. The global state of Harm Reduction. Towards an integrated response. London: Harm Reduction International; 2012.
5. Mumtaz GR, Weiss HA, Thomas SL, et al. HIV among people who inject drugs in the Middle East and North Africa: systematic review and data synthesis. PLoS medicine 2014;11:e1001663.
6. Ghasemian A, Shokuhi SK, KazemiVardanjani AR. Human Immunodeficiency Virus (HIV) Distribution in Middle East Region. Avicenna J Clin Microb Infec 2016:3(2).
7. Afriandi I, Aditama TY, Mustikawati D, Oktavia M, Alisjahbana B, Riono P. HIV and injecting drug use in Indonesia: epidemiology and national response. Acta medica Indonesiana 2009;41 Suppl 1:75-8.
8. Paraskevis D, Nikolopoulos G, Tsiara C, et al. HIV-1 outbreak among injecting drug users in Greece, 2011: a preliminary report. Euro surveillance 2011;16(36).
9. Niculescu I, Paraschiv S, Paraskevis D, et al. Recent HIV-1 Outbreak Among Intravenous Drug Users in Romania: Evidence for Cocirculation of CRF14_BG and Subtype F1 Strains. AIDS research and human retroviruses 2014.
10. Conrad C, Bradley HM, Broz D, et al. Community Outbreak of HIV Infection Linked to Injection Drug Use of Oxymorphone - Indiana, 2015. Mmwr-Morbid Mortal W 2015;64:443-4.
11. Wodak A, McLeod L. The role of harm reduction in controlling HIV among injecting drug users. Aids 2008;22 Suppl 2:S81-92.
12. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 Infection with Early Antiretroviral Therapy. New Engl J Med 2011;365:493-505.
13. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome biology 2014;15:538.
14. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. Science 2001;292:2323-5.
15. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. Genetics 2009;183:1421-30.
16. Nishiura H. Correcting the actual reproduction number: a simple method to estimate R(0) from early epidemic growth data. Int J Environ Res Public Health 2010;7:291-302.
17. Magiorkinis G, Sypsa V, Magiorkinis E, et al. Integrating phylodynamics and epidemiology to estimate transmission diversity in viral epidemics. PLoS computational biology 2013;9:e1002876.
18. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.

19. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular biology and evolution 2013;30:2725-9.
20. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution 1993;10:512-26.
21. Saad MD, Shcherbinskaya AM, Nadai Y, et al. Molecular epidemiology of HIV Type 1 in Ukraine: birthplace of an epidemic. AIDS Res Hum Retroviruses 2006;22:709-14.
22. Nabatov AA, Kravchenko ON, Lyulchuk MG, Shcherbinskaya AM, Lukashov VV. Simultaneous introduction of HIV type 1 subtype A and B viruses into injecting drug users in southern Ukraine at the beginning of the epidemic in the former Soviet Union. AIDS Res Hum Retroviruses 2002;18:891-5.
23. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology 2007;7:214.
24. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology 2006;4:e88.
25. UNAIDS. Indicator registry. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS; 2015.
26. Keeling Mj RP. Modeling Infectious Diseases in Humans and Animals. Princeton, New Jersey: Princeton University Press; 2008.
27. Tavare S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. Genetics 1997;145:505-18.
28. Kingman JFC. On the genealogy of large populations. J Appl Probability 1982;19:27-43.
29. Popinga A, Vaughan T, Stadler T, Drummond AJ. Inferring epidemiological dynamics with Bayesian coalescent inference: the merits of deterministic and stochastic models. Genetics 2015;199:595-607.
30. Diekmann O, Heesterbeek JA, Roberts MG. The construction of next-generation matrices for compartmental epidemic models. J R Soc Interface 2010;7:873-85.
31. Bellan SE, Dushoff J, Galvani AP, Meyers LA. Reassessment of HIV-1 Acute Phase Infectivity: Accounting for Heterogeneity and Study Design with Simulated Cohorts. PLoS Medicine 2015;12:e1001801.
32. Mosha F, Ledwaba J, Ndugulile F, et al. Clinical and virological response to antiretroviral drugs among HIV patients on first-line treatment in Dar-es-Salaam, Tanzania. J Infect Dev Ctries 2014;8:845-52.
33. Ugbena R, Aberle-Grasse J, Diallo K, et al. Virological response and HIV drug resistance 12 months after antiretroviral therapy initiation at 2 clinics in Nigeria. Clin Infect Dis 2012;54 Suppl 4:S375-80.
34. Abdissa A, Yilma D, Fonager J, et al. Drug resistance in HIV patients with virological failure or slow virological response to antiretroviral therapy in Ethiopia. BMC Infect Dis 2014;14:181.
35. Mathers BM, Degenhardt L, Ali H, et al. HIV prevention, treatment, and care services for people who inject drugs: a systematic review of global, regional, and national coverage. Lancet 2010;375:1014-28.
36. Longini IM, Jr., Clark WS, Byers RH, et al. Statistical analysis of the stages of HIV infection using a Markov model. Statistics in medicine 1989;8:831-43.
37. Artyuh OR, Balakireva OM, Bochkova LV, et al. Analytical report. Monitoring of behavior of injecting drug users as a component of second generation

epidemiological surveillance. Kyiv: International HIV/AIDS Alliance in Ukraine; 2005.
38. Beyrer C, Wirtz AL, Walker D, Johns B, Sifakis F, Baral SD. The Global HIV Epidemics among Men Who Have Sex with Men Introduction. Dir Dev 2011:1-4.
39. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. The immune response during acute HIV-1 infection: clues for vaccine development. Nature reviews Immunology 2010;10:11-23.
40. Booth RE, Lehman WE, Dvoryak S, Brewster JT, Sinitsyna L. Interventions with injection drug users in Ukraine. Addiction 2009;104:1864-73.
41. Grund J-PC. A Candle Lit from Both Sides: The Epidemic of HIV Infection in Central and Eastern Europe. Westport, Ct. USA: Greenwood Press; 2001.
42. Rhodes T, Lowndes C, Judd A, et al. Explosive spread and high prevalence of HIV infection among injecting drug users in Togliatti City, Russia. Aids 2002;16:F25-31.
43. Magiorkinis G. PD, Pybus OG., Karamitros T., Vasylyeva T., Bobkova M., Hatzakis M. HIV-1 epidemic in Russia: an evolutionary epidemiology analysis. Lancet. Online2014.
44. Huo D, Ouellet LJ. Needle exchange and injection-related risk behaviors in Chicago: a longitudinal study. JAIDS 2007;45:108-14.
45. Bluthenthal RN, Kral AH, Gee L, Erringer EA, Edlin BR. The effect of syringe exchange use on high-risk injection drug users: a cohort study. Aids 2000;14:605-11.
46. Des Jarlais DC, Marmor M, Friedmann P, et al. HIV incidence among injection drug users in New York City, 1992-1997: Evidence for a declining epidemic. Amer J Public Health 2000;90:352-9.
47. Friedman SR, Neaigus A, Jose B, et al. Sociometric risk networks and risk for HIV infection. Amer J Public Health 1997;87:1289-96.
48. Friedman SR, Kottiri BJ, Neaigus A, Curtis R, Vermund SH, Des Jarlais DC. Network-related mechanisms may help explain long-term HIV-1 seroprevalence levels that remain high but do not approach population-group saturation. Amer J Epidem 2000;152:913-22.
49. Khan B, Dombrowski K, Saad M, McLean K, Friedman S. Network Firewall Dynamics and the Subsaturation Stabilization of HIV. Discrete Dyn Nat Soc 2013;2013:720818.
50. Dombrowski K, Curtis R, Friedman S, Khan B. Topological and Historical Considerations for Infectious Disease Transmission among Injecting Drug Users in Bushwick, Brooklyn (USA). World J AIDS 2013;3:1-9.
51. Brenner BG, Roger M, Routy JP, et al. High rates of forward transmission events after acute/early HIV-1 infection. J Inf Dis 2007;195:951-9.
52. Powers K, Ghani A, Miller W, et al. The contribution of HIV-discordant couples to HIV transmission in Lilongwe, Malawi. JIAS 2012;15:79-.
53. Pinkerton SD. How many sexually-acquired HIV infections in the USA are due to acute-phase HIV transmission? Aids 2007;21:1625-9.
54. Eaton JW, Hallett TB. Why the proportion of transmission during early-stage HIV infection does not predict the long-term impact of treatment on HIV incidence. Proc Natl Acad Sci USA 2014;111:16202-7.
55. Vasylyeva TI, Friedman SR, Magiorkinis G. Prevention of early HIV transmissions might be more important in emerging or generalizing epidemics. Proceedings of the National Academy of Sciences of the United States of America 2015;112:E1515.

56. Rich JD, Adashi EY. Ideological Anachronism Involving Needle and Syringe Exchange Programs: Lessons From the Indiana HIV Outbreak. Jama 2015;314:23-4.
57. Farr AC, Wilson DP. An HIV epidemic is ready to emerge in the Philippines. Journal of the International AIDS Society 2010;13:16.
58. Booth RE, Lehman WE, Kwiatkowski CF, Brewster JT, Sinitsyna L, Dvoryak S. Stimulant injectors in Ukraine: the next wave of the epidemic? AIDS and behavior 2008;12:652-61.
59. Strathdee SA, Hallett TB, Bobrova N, et al. HIV and risk environment for injecting drug users: the past, present, and future. Lancet 2010;376:268-84.
60. Temprano Anrs Study Group, Danel C, Moh R, et al. A Trial of Early Antiretrovirals and Isoniazid Preventive Therapy in Africa. N Engl J Med 2015;373:808-22.
61. Group ISS, Lundgren JD, Babiker AG, et al. Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection. N Engl J Med 2015;373:795-807.
62. Vasylyeva TI, Friedman SR, Smyrnov P, Bondarenko K. A new approach to prevent HIV transmission: Project Protect intervention for recently infected individuals. AIDS Care 2015;27:223-8.
63. Csete J, Human Rights Watch Organization. Lessons not learned : human rights abuses and HIV/AIDS in the Russian Federation. New York: Human Rights Watch; 2004.
64. Mboi N. Indonesia. A consummate insider pushes ideas from outside Indonesia. Science 2014;345:162-3.

3 Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine

---

This chapter was published in:

Tetyana I Vasylyeva, Mariya Liulchuk, Samuel R Friedman, Iana Sazonova, Nuno R. Faria, Aris Katzourakis, Nataliia Babii, Alla Scherbinska, Julien Thézé, Oliver G Pybus, Pavlo Smyrnov, Jean L Mbisa, Dimitrios Paraskevis, Angelos Hatzakis, Gkikas Magiorkinis. Molecular epidemiology reveals the role of war in the spread of HIV in Ukraine. PNAS 2018 115(5):1051-1056.

Summary of authorship:

TIV wrote this chapter, conducted phylogenetic and statistical analyses. IS, PS, and SRF collected and prepared behavioural data. JT gave advice on phylogenetic analysis. ML, NB, and AS were involved in the sequencing of the genetic data. NRF, OGP, and GM provided supervisory support. All others helped with the interpretations of the results.

3.1 Abstract

Ukraine has one of the largest HIV epidemics in Europe, historically driven by people who inject drugs (PWID). The epidemic showed signs of stabilisation in 2012 but the recent war in eastern Ukraine might be reigniting virus spread. We investigated the movement of HIV-infected people within Ukraine before and during the conflict. We analysed HIV-1 subtype-A *pol* nucleotide sequences sampled during 2012-2015 from 427 patients of 24 regional AIDS Centers and used phylogeographic analysis to reconstruct virus movement among different locations in Ukraine. We then tested for correlations between reported PWID behaviours and reconstructed patterns of virus spread. Our analyses suggest that Donetsk and Lugansk, two separatist-controlled cities in eastern Ukraine, were the main exporters of the virus to the rest of the country. Additional analyses showed that viral dissemination within the country changed after 2013. Spearman correlation analysis showed that incoming virus flow was correlated with the number of HIV-infected internally displaced people. Additionally, there was a correlation between more intensive virus movement and locations with a higher proportion of PWID practicing risky sexual behaviours. Our findings suggest that effective prevention responses should involve internally displaced people and people who frequently travel to war-affected regions. Scale-up of harm reduction services for PWID will be an important factor in preventing new local HIV outbreaks in Ukraine.

3.2 Significance statement

Current conflict in the heavily HIV-affected regions of eastern Ukraine, which caused the relocation of 1.7 million people, may have increased the exportation of the virus from war-affected regions to other parts of the country. We show that the large-scale war-initiated movement of people (including those HIV-infected) was associated with patterns of HIV flow reconstructed from gene sequence data. Our results further suggest that virus dissemination was directed to the locations with the highest prevalence of risky sexual practices in PWID. Recently the Ukrainian HIV epidemic has been transitioning from being PWID-concentrated to a sexually driven one. Our study helps to understand the factors that might facilitate this shift.

3.3 Introduction

The HIV epidemic in Ukraine is one of the largest in Europe: in 2015 the country had an estimated 220,000 people living with HIV and the highest rate of new infections in Europe (~30 cases per 100,000 people per year)[1]. After independence in 1991, Ukraine struggled with political, economic, and social crises[2,3], which led to the expansion of drug trafficking and drug use[4]. The number of people who inject drugs (PWID) grew quickly, starting from southern and eastern Ukraine (Odessa, Mykolaiv, Dnipro, and Donetsk regions)[5]. Until 1993 the country had <40 new HIV cases per year[6], mostly attributed to unprotected commercial sex[7]; the first HIV transmissions attributed to intravenous drug use were registered in the south of Ukraine in 1993. In 1994 the number of new cases rocketed to 12,000 and by 1997 HIV in PWID was reported in all administrative regions of Ukraine[8]. This was the beginning of one of the largest HIV PWID epidemics recorded, characterised by multiple transmissions within the first month after infection and driven by risky injecting practices (see Chapter 2 and [9,10]).

In 2002 Ukraine introduced a harm reduction approach[11] that has been shown to reduce risky injecting practices and HIV transmission among PWID[12]. This included syringe exchange programs, condom distribution, and, since 2005, opioid substitution treatment programs (OST)[13]. Consequently, the fraction of newly diagnosed cases attributed to illicit drug use declined; since 2010 more newly diagnosed infections have been attributed to sexual than parenteral transmission[14]. In 2012 the number of newly registered HIV cases dropped for the first time in the history of the epidemic in Ukraine - an indicator that the epidemiological situation had stabilised[14].

However, this epidemiological equilibrium was disrupted by military conflict that began in Ukraine following a political insurgency and the Maidan movement in late

2013[15]. In early 2014 the Crimean Peninsula was annexed, and the ongoing military conflict started in eastern Ukraine, in the Donetsk and Lugansk regions. During the conflict, health care provision is continuously interrupted[16] and there have been documented shortages of antiretroviral treatment (ART), closures of OST sites, and a scale-down of harm reduction services in general[15,17-19]. HIV incidence has grown 15-54% in some areas affected by war[20]. Notably, the conflict caused massive internal human migration: since 2014 over 1.7 million people have been internally displaced in Ukraine[21], out of a total population of 45 million[22]. HIV prevalence in general population stands at >0.5% suggesting that >8,000 of these displaced individuals might be HIV-infected. However, only 1,153 HIV-infected internally displaced people (IDP) have registered with AIDS centres at their new places of residence[23]. Earlier, a WHO report highlighted that IDP are not properly covered by HIV surveillance in the area[24]. Large-scale human migration is likely to have had an impact on epidemic dynamics within and among the regions of Ukraine. Within cities and well-established transmission networks, chronically infected people protect individuals susceptible to infection from highly-infectious recently infected cases, creating the so-called "firewall" effect[25,26]. Changes to established networks could disrupt this effect and trigger further viral transmission or even provoke local outbreaks. Furthermore, as infected people move to new places, they might disseminate new viral strains. In addition, new groups of susceptible individuals may be generated, e.g. disadvantaged IDP, affected by war, with potentially poor access to care. In general, we might expect an increased export of HIV from eastern Ukraine (historically the most heavily HIV-affected region) to other locations.

There have been some reports of the response of Ukrainian non-governmental organisations to the conflict[19]. However, given the disruptive nature of the conflict

and only sporadic evidence of its consequences, there is need for additional research about the current epidemiological situation. Here we analyse 427 HIV-1 subtype A *pol* sequences collected from almost all administrative regions of Ukraine in order to investigate how viral infections have been redistributed among geographical locations in recent years. We then combine these findings with epidemiological and behavioural data to explore the current epidemiological situation in those regions where increased viral gene flow was detected.

3.4 Methods

3.4.1 Data

3.4.1.1 Sequences sampling and phylogenetic analysis

We used 448 HIV-1 genetic sequences sampled in 2012-2015 from patients of 24 Ukrainian Regional Centers to Prevent and Fight AIDS (AIDS Centers) (52% males) and available from the Ukrainian HIV drug resistance database. 427, 19, and 2 of these sequences were identified as subtypes A, B, and circulating recombinant form 03_AB, respectively. All of the available high-quality (<5% of ambiguous nucleotides) subtype A sequences were used in the subsequent phylogenetic and phylogeographic analyses.

We aligned the sequences using the software MEGA 7.0[27] and then manually edited the alignment deleting 43 codon positions associated with drug resistance (RT: 41, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 115, 116, 151, 179, 181, 184, 188, 190, 210, 215, 219, 225, 230; PR: 23, 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 76, 82, 83, 84, 85, 88, 90)[28]. The final alignment consisted of 866 codons that correspond to the 2253 - 3216 positions of the HXB2 HIV reference sequence (*pol* gene). We used the

REGAv3[29] and COMET[30] subtyping tools to identify subtypes and recombinant sequences.

To explore the position of the 427 subtype A sequences from the Ukrainian HIV database (accession numbers KY370106-KY370532) within the global subtype A epidemic, we merged our sequences with a globally representative subtype A reference dataset, which consisted of 2,199 sequences, using Clustal W[31]. The dataset consisted of all available sequences of minimal length of 300 nucleotides (nt) from the *pol* region of HIV genome accessed in 2011 from the HIV sequence database (sequence names are in the Appendix 1 (dataset A)). Sequences came from 36 countries (Afghanistan, Albania, Azerbaijan, Burkina Faso, Burundi, Benin, Democratic Republic of Congo, Chad, Congo, Cameroon, Cuba, Cyprus, Czech Republic, Ethiopia, France, Gabon, Georgia, Ghana, Equatorial Guinea, Kenya, Kazakhstan, Latvia, Mali, Nigeria, Russia, Rwanda, Sudan, Slovenia, Senegal, Togo, Tanzania, Ukraine, Uganda, Uzbekistan) sampled in 1985-2010. We then reconstructed an ML phylogenetic tree with 2,626 sequences (427 Ukrainian and 2,199 reference) using RAxML[32] to check the position of our sequences within the global subtype-A diversity. We further selected 36 subtype A HIV-1 reference sequences sampled in 1987-2013 publicly available from the Los Alamos database[33] to improve the molecular clock calibration. Sequences were selected to ensure the best geographical and temporal representation. These reference sequences come from the following countries: Afghanistan, Armenia, Belarus, Benin, Burundi, Cameroon, Cuba, Cyprus, Ethiopia, Georgia, Kazakhstan, Kenya, Kyrgyzstan, Mali, Russian Federation, Rwanda, Senegal, South Africa, Sudan, Tanzania, Uganda, Zambia. We aligned these reference sequences (N=36) to the Ukrainian dataset (N=427) in a combined dataset using Clustal W. We further constructed an ML phylogenetic tree in

RAxML with these 463 sequences. We have tested the temporal signal of the combined dataset with the TempEst program[34]. This combined dataset was further used to produce the posterior tree distribution for the phylogeographic analysis. The ggtree package in R was used to visualize the phylogenetic trees[35].

3.4.1.2 Human migration and behaviours

Data on risky injecting behaviours came from the Integrated Bio-Behavioural Surveillance (IBBS) of PWID (N=9,002) implemented in Ukraine in 2013. IBBS has a cross-sectional survey design and was conducted via respondent-driven sampling. In IBBS-2013 participants were asked about their injecting and sexual behaviours and tested for HIV and Hepatitis C virus (HCV) antibodies using rapid tests CITO TEST HIV 1/2/07. We conducted the secondary analysis of data collected in 25 regional capital cities of Ukraine (only data corresponding to the 24 administrative units for which genetic sequences were available were used). Only PWID in regional capitals were surveyed, since Ukraine has a highly urbanised HIV epidemic: 77% of all HIV cases in 2013 were registered among urban population[36]. The cities, selected as the study sites for the IBBS, represent all geographical areas of Ukraine. These areas represent varying HIV-prevalence rates among PWID and varying PWID population sizes[37,38]. Sample size was calculated for each study site by combining group size estimation and HIV-prevalance level from the previous IBBS study conducted in 2011[39]. The inclusion criteria for study participants included: 1) injection drug use in the last 30 days before the study 2) 14 years old or older 3) permanent resident of the city where IBBS was conducted.

We accessed per region statistics from publicly available sources as follows:

1) the number of internally displaced persons (IDP) was available from the website of the CEDOS analytical center (http://www.cedos.org.ua/). This website

summarises data delivered by the UN High Commissioner for Refugees and the Ukrainian Ministry of Social Policy[40];

2) HIV prevalence estimates (cases per 100,000 population) in the general population were available from the website of the Ministry of Health of Ukraine[41];

3) the number of HIV-positive IDP registered with an AIDS center at the place of their current residence was available from the National AIDS Center website (http://ucdc.gov.ua/);

4) the population size of every region as of 2016 was available from the website of the State Statistics Service of Ukraine[42].

3.4.2 Phylogeographic analysis

3.4.2.1 Dating phylogenetics

We conducted molecular clock and discrete trait analyses in BEASTv1.8.4[43,44]. We used the general time-reversible (GTR) nucleotide substitution model with a gamma distributed rate variation among sites. We used the lognormal uncorrelated relaxed clock model to account for rate heterogeneity among lineages[45]. We used a Bayesian Skyline plot model with 10 intervals to reconstruct population dynamics. Analyses were run for 200,000,000 generations (burn-in of 30,000,000) until obtaining appropriate mixing and convergence of the Markov Chain Monte Carlo sampler in the posterior target distribution (effective sample size>150). We used the LogCombiner[43] to subsample the posterior tree distribution, and after removing 15% of burn-in, the resulting empirical tree distribution (N=1,700 trees) was used in subsequent phylogeographic analysis after the reference sequences were pruned using PAUP[46]. The maximum clade credibility tree for the Ukrainian dataset was reconstructed in TreeAnnotator.

3.4.2.2 Discrete trait analysis

For the discrete trait phylogeographic analysis we grouped the 427 Ukrainian sequences from 24 different administrative regions into 7 geographic locations as shown in Table 3.1. We grouped the data into 7 locations to minimize over-parameterisation of the model and to ensure that every geographic location had at least 10 sequences. We also made sure that this grouping was appropriate in cultural and social context of Ukraine.

Table 3.1. Distribution of sequences per regions and locations, Ukrainian dataset

Sequences IDs	Sampling region	Number of sequences	Allocated geographical group (location)	Number of sequences (full dataset)	Number of sequences (sensitivity analysis)
1-14	Zhytomyr	14	Center	82	57
15-32	Vinnitsya	18	Center		
33-36	Sumy	4	Center		
37-44	Poltava	8	Center		
45-54	Hmelnitskiy	10	Center		
55-62	Chernigiv	8	Center		
63-79	Cherkassy	17	Center		
80-82	Harkiv	3	Center		
83-92	Crimea	10	Crimea	10	10
93-108	Lugansk	16	East	97	57
109-189	Donetsk	81	East		
190-209	Kyiv Region	20	Kyiv	77	57
210-266	Kyiv City	57	Kyiv		
267-323	Odessa	57	Odessa	57	57
324-339	Mykolaiv	16	South	82	57
340-346	Zaporizzhya	7	South		
347-351	Kirovograd	5	South		
352-361	Herson	10	South		
362-405	Dnipro	44	South		
406-407	Zakarpattya	2	West	22	22
408	Volyn	1	West		
409-413	Rivne	5	West		
414-419	Lviv	6	West		
420-427	Ivano-Frankivsk	8	West		
	Total	427	Total	427	317

We ran discrete phylogeographic analysis with the robust counting method as implemented in BEAST and using the empirical tree distribution as described before[47-50]. This approach infers the expected number of viral migrations for every pathway on a branch-by-branch basis. The location exchange process throughout the phylogeny was estimated using asymmetric (non-reversible) discrete-time Markov chains, which quantified the number of exporting and importing virus lineage movements for each location. We estimated the viral gene flow as the number of migration events, which is the number of viral lineage movements from one location to another.

In the first step of the analysis we included the full Ukrainian dataset, N=427. In an attempt to account for the bias that might arise from unequal allocation of sequences per location, we ran additional analyses with reduced number of sequences per location. Specifically, we conducted 10 replicate analyses where we reduced the number of sequences in each location to the third smallest number in the dataset (N=57 for Center, East, Kyiv, Odessa, and South, N=10 for Crimea and N=22 for West). For each one of these analyses we randomly selected different subsets of the sequences for each location, with the same N=314 for each subset. Phylogeographic analyses using the empirical tree distributions were then run for 5,000,000 generations (with a 10% burn-in).

3.4.2.3 Phylogeographic transition

To check for changes in the viral gene flow that might be associated with the military conflict in the East or any Maidan-related effects, we applied models with time-dependent lineage transition rates, i.e. "epoch analysis" as implemented in BEAST[51]. Specifically, we compared the 'single epoch' model (in which pairwise rates of lineage movement remained constant between 2012-2015) with the 'two epoch'

model (in which pairwise rates of lineage movement could be different before and after the start of the conflict in 2013). This approach allows testing for a change in viral gene flow after a specified time period. For the purposes of this analysis, we specified 2 epochs for the "two-epoch" model, before and after December 31st, 2013. We were not able to specify the official date of the beginning of war (April 6th, 2014) as the transition time, because the dates in our analysis were not resolved at a monthly scale, only sampling year data were available. We used path-sampling (PS)[52] and stepping-stone sampling (SS)[53] marginal likelihood estimator (MLE) techniques to select which model fits the data best, "single epoch" or "two epoch". We followed the standard terminology in Bayes Factor (BF) interpretation: the strength of evidence for a particular model is substantial when BF > 3, strong if BF > 10, very strong if BF > 30 and decisive if BF > 100[54].

We first compared the two models as implemented in our primary discrete trait analysis: grouping the data into 7 discrete locations. Both PS and SS MLE were higher for "single epoch" model, but with a very low statistical support (BF=3) (Table 3.2).

Table 3.2. Marginal Likelihood estimated using path-sampling and stepping-stone methods. Bayes Factor provides statistical support for one of the two models considered (1 epoch vs 2 epochs).

Sample	Path-sampling		Bayes Factor	Stepping-stone		Bayes Factor
	1 epoch	2 epochs		1 epoch	2 epochs	
Full (7 locations)	-866.65	-870.47	3.83	-865.11	-868.82	3.71

Since we tested the time-heterogeneity over a very short period of time (4 years, 2 years in each epoch), we decided to reduce the number of locations in the analysis to

avoid the over-parameterisation of the analysis. For this purpose, we created 10 subsets of data where the sequences were grouped into 2 locations: "East" and "Other". For each of the subsets, we included all the sequences from East and the same number of sequences subsampled from all of the other locations.

3.4.3 Spearman correlation analysis

We performed Spearman correlation test to check if the numbers of viral lineage migration exportation and importation events were correlated with some of the behavioural and epidemiological characteristics of locations. We ran two separate analyses of correlation to individually explore factors that are associated with the number of exportation and importation events. All the p-values for the correlation analyses were adjusted for multiple comparisons using false discovery rate method[55]. For every location, we looked at the following:

1) The sum of exportation viral lineage migration events, from a location to all the other locations. For example, the sum of exportation viral lineage migration events for Kyiv equals all the migration events from Kyiv to any other location.

2) The sum of importation viral lineage migration events, from a location to all the other locations. For example, the sum of importation viral lineage migration events for Kyiv equals the sum of all the migration events to Kyiv from any other location.

We checked if the viral gene flow was correlated with the behavioural and epidemiological characteristics of the locations constructed as following:

1. For continuous variables, we looked at mean and median values per location. For example, mean number of commercial partners per an injector.

2. For categorical variables, we looked at the % of respondents who answered "Yes" or "Always" to a question about a specific behaviour. For example,

the proportion of PWID who answered "Always" to the question "How often did you use someone else's used syringe?"

All injecting practices of PWID were reported for the last 30 days before the survey and all sexual practices of PWID were reported for the last 90 days. For each of the locations that were used in the phylogeographical analysis, we calculated values of a set of behavioural and epidemiological variables. For every variable we present summary values of this variable for each of the administrative regions in the respective location. The list of explanatory variables include:

1. Injecting practices that might increase the HIV transmission risk ("Risky injecting practices"):
    - Median frequency of injections per week in regions that are part of a location (i.e. for East, median duration of drug use for PWID in Donetsk and Lugansk)
    - Median number of PWID that PWID used drugs with at the same place
    - Proportion of PWID who sometimes gave used syringes to other injectors
2. Injecting practices that might protect against HIV transmission ("Safe injecting practices"):
    - Proportion of PWID who used sterile syringe at the last injection.
3. Sexual practices that might increase the HIV transmission risk ("Risky sexual practices"):
    - Mean number of casual sexual partners per injector
    - Mean number of commercial partners per injector
    - Mean number of all sexual partners (overall number of sexual partners)
    - A proportion of PWID who participated in group sex.
4. Sexual practices that might protect against HIV ("Safe sexual practices"):

- Proportion of PWID who always used condoms with casual partners
- Proportion of PWID who always used condoms with commercial partners
- Proportion of PWID who used condoms at the last sexual intercourse

5. Regional level epidemiological characteristics:
    - HIV prevalence in PWID as measured in the IBBS-2013
    - Number of internally displaced people (IDP) relocated to and registered in a location
    - Number of HIV-positive IDP relocated to and registered in a location
    - Cumulative number of HIV-diagnosed people in a location (HIV prevalence in general population) as of 2013
    - Population size (number of residents) in a location (combined number of residents of all the regions in this location) as of 2013.

Apart from the PWID behavioural variables, we checked the correlation between the viral gene flow and 6 additional epidemiological characteristics of the locations:

- HIV prevalence in PWID as measured in IBBS-2013;
- HCV prevalence in PWID as measured in IBBS-2013;
- Number of IDP relocated to and registered in a location;
- Number of HIV-positive IDP relocated to and registered in a location;
- Overall cumulative number of HIV-diagnosed people (HIV prevalence per location);
- Population size (number of residents) per location.

3.5 Results

3.5.1 HIV-1 dissemination among locations in Ukraine

We analysed 427 HIV-1 subtype A *pol* gene sequences available from the Ukrainian HIV drug resistance database collected between 2012 and 2015 from 22 administrative regions of Ukraine, from the capital of Kyiv, and from the Autonomous Republic of Crimea, representing 24 of the 27 Ukrainian administrative units. For the phylogeographic analysis we grouped the sequences into 7 geographic locations: Center, Crimea, East (including Donetsk and Lugansk – two regions in eastern Ukraine involved in military conflict), Kyiv, Odessa, South, and West. Hereafter we refer to these groups as "locations" and to the Ukrainian administrative units as "regions".

An initial maximum likelihood (ML) phylogenetic tree included 2199 reference and 427 Ukrainian subtype A sequences (Fig. 3.1). The tree revealed that the Ukrainian samples formed a distinct monophyletic clade within sequences from countries of the former Soviet Union (FSU), which in turn formed a clade within HIV-1 subtype A.

Further, we selected 36 HIV subtype A sequences sampled in 1986-2014 and added them to the Ukrainian sequences to improve molecular clock calibration (N=463; the combined dataset). The ML phylogenetic tree reconstructed from the combined dataset is demonstrated in Figure 3.2.

The combined dataset showed sufficient temporal signal (root-to-tip correlation coefficient = 0.5), supporting the use of molecular clock models. Phylodynamic analysis showed that the most recent common ancestor of the sequences in the Ukrainian dataset was 2001 (95% Highest Posterior Density (HPD): 1999 – 2002), for the FSU countries was 1998 (95% HPD: 1997 – 2000), and for the combined dataset was 1977 (95% HPD: 1972 - 1981). The resulting empirical posterior distribution of

dated phylogenies was used in subsequent phylogeographic analyses, after the removal of the non-Ukrainian reference sequences (N=427; the Ukrainian dataset). Given the *k*=7 sampling locations, we investigated virus dissemination among the *k(k*-1)=42 possible pathways of virus spread using discrete phylogeographic models[44,56]. More specifically, we investigated the main exporters of viral lineages, i.e. the locations from which the majority of migration events originated.

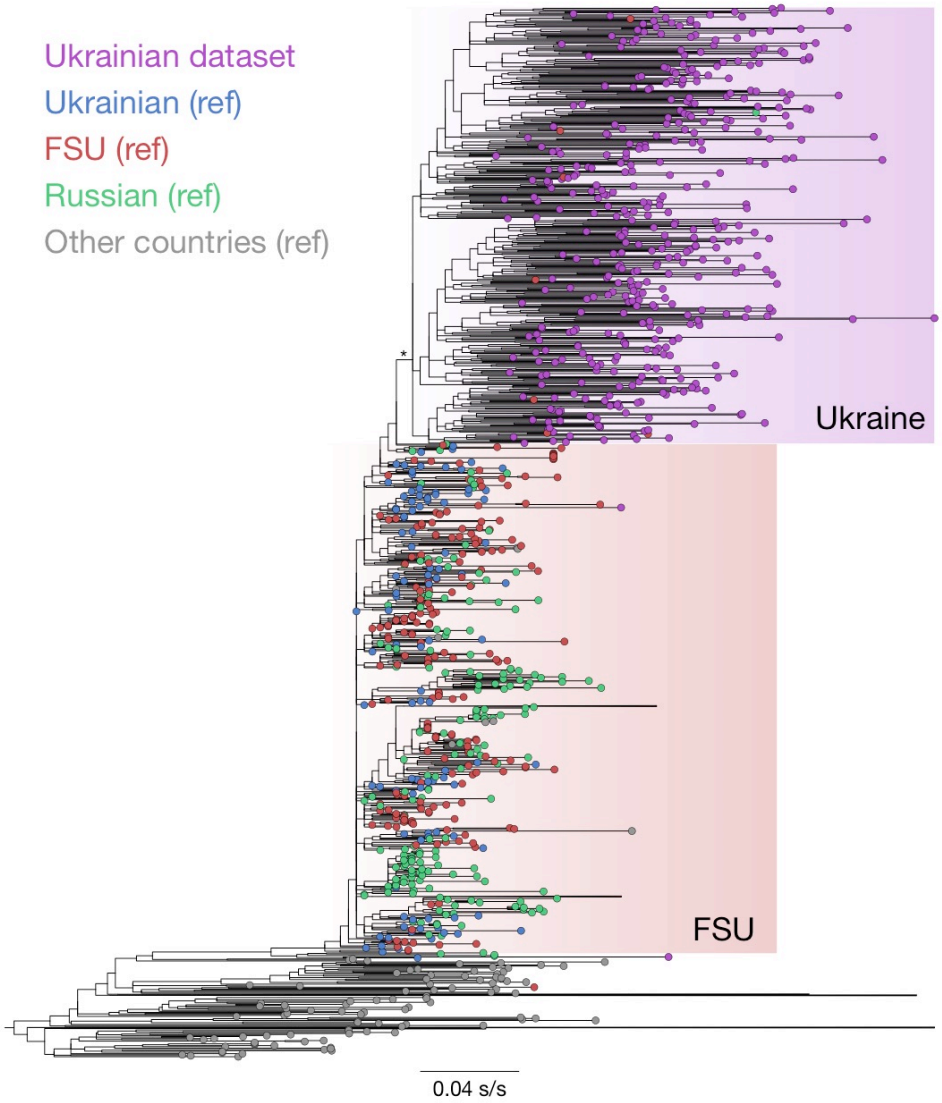


Figure 3.1. Maximum Likelihood (ML) phylogenetic reconstruction of global HIV-1 subtype A genetic diversity based on genetic sequences (N=2,626) from the Ukrainian

dataset (N=427) and publicly available sequence (N=2,199, reference dataset). The bar represents the genetic distance. The asterisk marks the Ukrainian dataset further used for phylogeographical analysis.

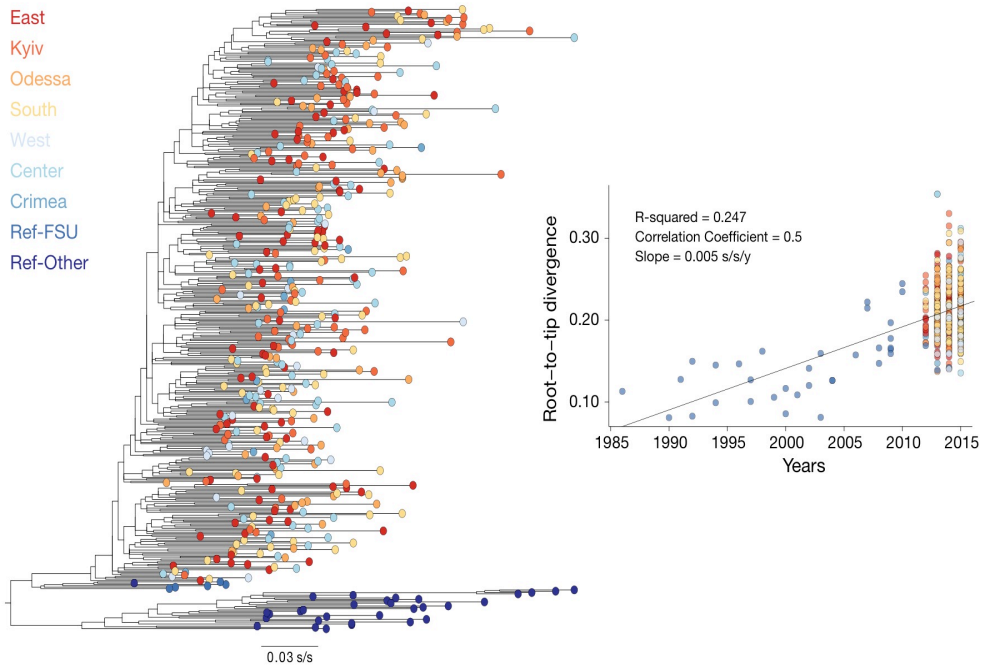


Figure 3.2. ML phylogenetic tree reconstructed in RaXML from the combined dataset (N=463). The combined dataset consisted of the Ukrainian dataset (N=427) and the reference sequences (N=36); reference sequences were included to improve the temporal signal. Results of the Tempest analysis that indicate the strength of the signal in the combined dataset are presented on the right.

In our phylogeographic analysis, the East (Donetsk and Lugansk regions) was the main exporter, accounting on average for 92% (95% HPD: 87%-96%) of the migration events in the Ukrainian dataset (Table 3.3, Appendix 2, and Fig. 3.3). Particularly, viral lineage movements were most frequently observed from the East to the Center, South, Kyiv, and Odessa locations, accounting for 26% (95% HPD: 24%-

28%), 24% (95% HPD: 21%-26%), 22% (95% HPD: 20%-24%), and 17% (95% HPD: 16%-18%) of migration events, respectively. Very few migration events were observed as originating from the Crimea and West locations (<1% each), but West received a significant proportion of translocated viral lineages (7% of viral lineage movement events, 95% HPD: 6% - 8%).

Table 3.3. Lineage migration events in the Ukrainian HIV-1 dataset, 2012-2015.

HPD = Highest Posterior Density.

From	Exportation events		Importation events	
	N [95% HPD]	%	N [95% HPD]	%
Center	4.83 [0, 11]	1.59	78.76 [74, 84]	25.96
Crimea	0.52 [0, 2]	0.17	10.39 [10, 12]	3.43
East	279.73 [264, 292]	92.27	1.83 [0, 5]	0.6
Kyiv	7.00 [1, 13]	2.31	67.65 [61, 74]	22.31
Odessa	4.13 [0, 9]	1.36	50.88 [46, 55]	16.78
South	5.36 [1, 12]	1.76	72.24 [65, 78]	23.83
West	1.74 [0, 5]	0.57	21.60 [19, 24]	7.12
All, N	303.18 [289, 315]			

Given that locations were represented by different numbers of sequences, we explored estimates of among-location viral lineage movement under condition of equal representation of locations in the analysis. Specifically, we randomly down-sampled sequences in each location to match the number of sequences in the third smallest location (Odessa, N=57). This resulted in five locations with the same number of sequences (Center, East, Kyiv, Odessa, and South) and two locations with a smaller number of sequences (Crimea and West, N=10 and N=19, consequently), which were unlikely to be significant contributors to virus gene flow, based on the preceding analyses. We did not reduce the number of sequences to the smallest sample size (N=10) in order to ensure acceptable statistical power.

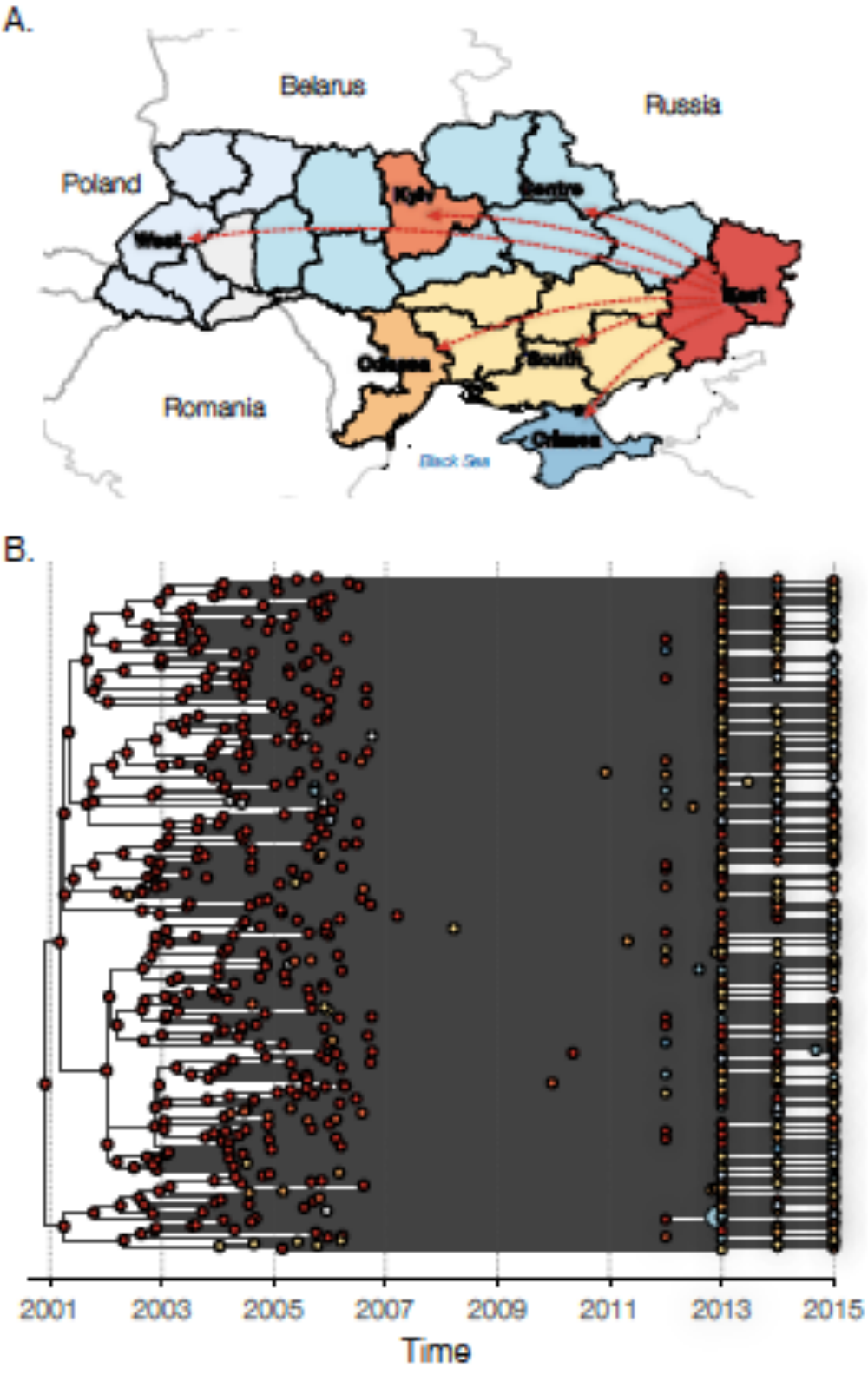

A.
Belarus
Russia
Poland
Romania
Black Sea
B.
2001
2003
2005
2007
2009
2011
2013
2015
Time

Figure 3.3. (A) A map of Ukraine. Colours indicate the grouping of the sequences in locations used in phylogeographic analysis. The arrows indicate the directionality of virus gene flow movement from the East.

(B) Results of the discrete trait phylogeographic analysis based on the Ukrainian dataset (N=427). Colours indicate the ancestral state locations of the internal nodes reconstructed by robust count phylogeographic method and the sampling locations of the tips.

We conducted 10 replicate analyses with different random sub-samples of the data set. We found that the East remained as the main viral lineage export location in 70% of the analyses (7/10 sub-samples). In those replicate analyses, where the East was the main exporter of the viral lineages, it accounted on average for 91% of migration events. In the other 3 sub-samples, Center, Kyiv, and South were the main exporters each, respectively. For the number of migration events estimated for each subsampled dataset see Appendix 3.

3.5.2 Temporal change in virus lineage movement after the beginning of the conflict

We aimed to investigate whether viral gene flow among regions has changed after the initiation of war, a phenomenon we call hereafter “a phylogeographic transition”. We estimated rates of viral lineage movement using models with time-dependent lineage transition rates. Specifically, we compared the ‘single epoch’ model (in which lineage migration rates remained constant between 2012-2015) with the ‘two epoch’ model (in which lineage migration rates were allowed to be different before and after the start of the conflict in 2013) (Table 3.4).

Table 3.4. Results of the Marginal Likelihood estimations.

Sample	Path-sampling		Bayes Factor	Stepping-stone		Bayes Factor
Epoch model	1 epoch	2 epochs		1 epoch	2 epochs	
Sub-sample 1	-175.04	-144.03	**31.01**	-174.66	-143.91	**30.75**
Sub-sample 2	-179.38	-146.56	**32.82**	-179.00	-146.46	**32.44**
Sub-sample 3	-176.18	-144.20	**31.98**	-175.83	-144.10	**31.73**
Sub-sample 4	-175.20	-143.60	**31.60**	-174.84	-143.51	**31.34**
Sub-sample 5	-178.77	-145.99	**32.78**	-178.44	-145.91	**32.43**
Sub-sample 6	-177.69	-145.52	**32.17**	-177.27	-145.37	**31.89**
Sub-sample 7	-176.36	-144.58	**31.78**	-176.02	-144.49	**31.53**
Sub-sample 8	-175.81	-143.94	**31.87**	-175.41	-143.84	**31.57**
Sub-sample 9	-178.96	-146.19	**32.77**	-178.58	-146.10	**32.48**
Sub-sample 10	-176.73	-144.24	**32.49**	-176.41	-144.15	**32.26**

The results of the epoch analysis showed that the 'two epoch' model fitted the data better than the 'single epoch' model, as demonstrated by a Bayes factor (BF) support of >30 in favour of the more complex model. Standard terminology defines the strength of support for a particular model as "very strong" for BF>30[54]. This result was consistent across all 10 replicate data sets generated by random subsampling.

3.5.3 Correlation between observed virus flow and PWID behaviours

We further investigated whether among-location viral lineage movement was correlated with important epidemiological characteristics of the HIV-1 epidemic in Ukraine. Here, we calculated Spearman correlation coefficients between exporting and importing viral gene flow and all behavioural covariates of interest. In particular, we were interested in the total number of lineage migration events that originated in a given location (exportation events) and the total number of lineage migration events received by a given location (importation events). Behavioural data and regional epidemiological characteristics were aggregated to the same geographic units (locations) as used in the phylogeographic analysis.

The behavioural variables came from (i) the national survey of PWID (Integrated Bio-Behavioural Surveillance, IBBS), (ii) the regional HIV surveillance systems, and (iii) data on the war-associated movement of IDP. We chose the IBBS data that were collected closer to the phylogeographic transition (i.e. IBBS-2013) as an indicator of behaviours at the time when viral movement is hypothesised to have changed.

We find that injecting behaviours of PWID were not associated with the observed viral migration flow. Risky sexual behaviours of PWID were, on the contrary, associated with the numbers of importation events. Specifically, the proportion of PWID who always use condoms with casual and commercial partners was associated with a lower number of importation events (Spearman correlation p-values =0.03 and =0.05, respectively, after adjustment for multiple comparisons using false discovery rate method[55]) (Table 3.5).

The number of importation events per location was also negatively correlated with the proportion of PWID who reported participating in group sex (p-value =0.04). Importantly, we found that the number of exportation events to a location was positively correlated with the HIV prevalence in general population (p-value=0.05) and the number of IDP relocated to that location (p-value <0.01). At the same time the number of HIV-infected IDP registered at that location was positively correlated with a higher number of importation events (p-value=0.03).

Table 3.5 (next page). Correlation coefficient (R) between behavioural variables and the number of migration events. Variables in bold indicate p-value ≤0.05.

Variables	Number of exp. migration events		Number of imp. migration events	
	R	p-value	R	p-value
Risky injecting practices				
Median frequency of injections per week	0.04	0.94	-0.35	0.44
Median number of PWID that PWID used drugs with at the same place	-0.45	0.31	0.06	0.90
Proportion of PWID who sometimes gave used syringes to other injectors	-0.43	0.34	0.18	0.70
Safe injecting practices				
Proportion of PWID who used sterile syringe at the last injection	-0.21	0.64	0.68	0.09
Risky sexual practices				
Mean number of casual sexual partners per injector	0.38	0.40	-0.23	0.61
Mean number of commercial partners per injector	0.29	0.53	0.61	0.15
Mean number of all sexual partners (overall number of sexual partners)	0.36	0.43	0.18	0.70
Proportion of PWID who reported participating in group sex	0.21	0.64	**-0.82**	**0.04**
Safe sexual practices				
Proportion of PWID who always used condoms with casual partners	0.04	0.94	**-0.86**	**0.03**
Proportion of PWID who always used condoms with commercial partners	0.07	0.88	**-0.75**	**0.05**
Proportion of PWID who used condoms at the last sexual intercourse	0.36	0.43	0.03	0.99
Regional level epidemiological characteristics				
HIV prevalence in PWID, %	0.07	0.88	0.04	0.94
Number of IDP relocated to and registered in a location	**0.96**	**<0.01**	0.14	0.76
Number of HIV-positive IDP relocated to and registered in a location	0.43	0.34	**0.86**	**0.03**
HIV prevalence in general population	0.75	**0.05**	0.18	0.70
Population size in a location	0.32	0.48	0.71	0.07

3.5.4 Time to failure and drug resistance prevalence

We performed a basic drug resistance analysis on our samples but did not observe a change in the frequency of drug resistance mutations since the beginning of the war, apart from a small decrease in L90 mutation (Table 3.6 and Fig. 3.4). Interestingly, some of the mutations have very high prevalence, >50%, both before and after war (i.e. M184 and G190), suggesting that they are polymorphic in the sampled population (patients who failed treatment).

We observed an increase (albeit non-significant) in the proportion of sequences with a K65R mutation since the initiation of war, which also coincides with an increased percentage of tenofovir desoproxil fumarate (TDF)-based regimens prescribed in the general HIV population (Table 3.7). K65R is a signature mutation for TDF resistance associated with early virological failure[57]. However, if we compare the prevalence of this mutation between the beginning (2012) and the end (2015) of our observation period, then the increase is statistically significant (Chi-square test p-value=0.03).

Based on the recent scale-up of TDF in the Ukrainian HIV population, we would expect the time to treatment failure in our sample to be approximately 12-24 months, comparable to the time observed in the early days of highly active ART[58]. The increase in patients with an absence of primary resistance mutations within the sampling period was insignificant (Table 3.6), suggesting that cases with severe adherence issues were uniformly distributed before and after war.

Table 3.6. Distribution of drug-resistance associated mutations

Mutation	2012-2013, N	2012-2013, % of patients that have this mutation	2014-2015, N	2014-2015, % of patients that have this mutation	Chi-square test of significance, p-value
Non-nucleoside RT inhibitors					
L100	1	1	5	2	0.43
K101	68	42	113	43	0.84
K103	26	16	29	11	0.14
V106	0	0	3	1	0.20
E138	21	13	39	15	0.57
Y181	42	26	91	34	0.08
Y188	2	1	0	0	0.10
G190	92	57	151	57	1.00
M230	2	1	3	1	1.00
Nucleoside RT inhibitors					
M184	109	67	165	62	0.30
K65	37	23	80	30	0.12
K70	28	17	53	20	0.44
L74	14	9	35	13	0.21
Y115	16	10	30	11	0.75
M41	11	7	13	5	0.39
D67	34	21	42	16	0.19
L210	6	4	5	2	0.22
T215	24	15	28	11	0.23
K219	22	14	39	15	0.78
Protease inhibitors					
I47	0	0	2	1	0.20
I54	2	1	1	0	0.10
L90	**3**	**2**	**0**	**0**	**0.02**
M46	5	3	7	3	1.00
V82	4	2	8	3	0.53
Patients with no drug-associated mutations	28	17	52	20	0.224
N of patients	162		265		

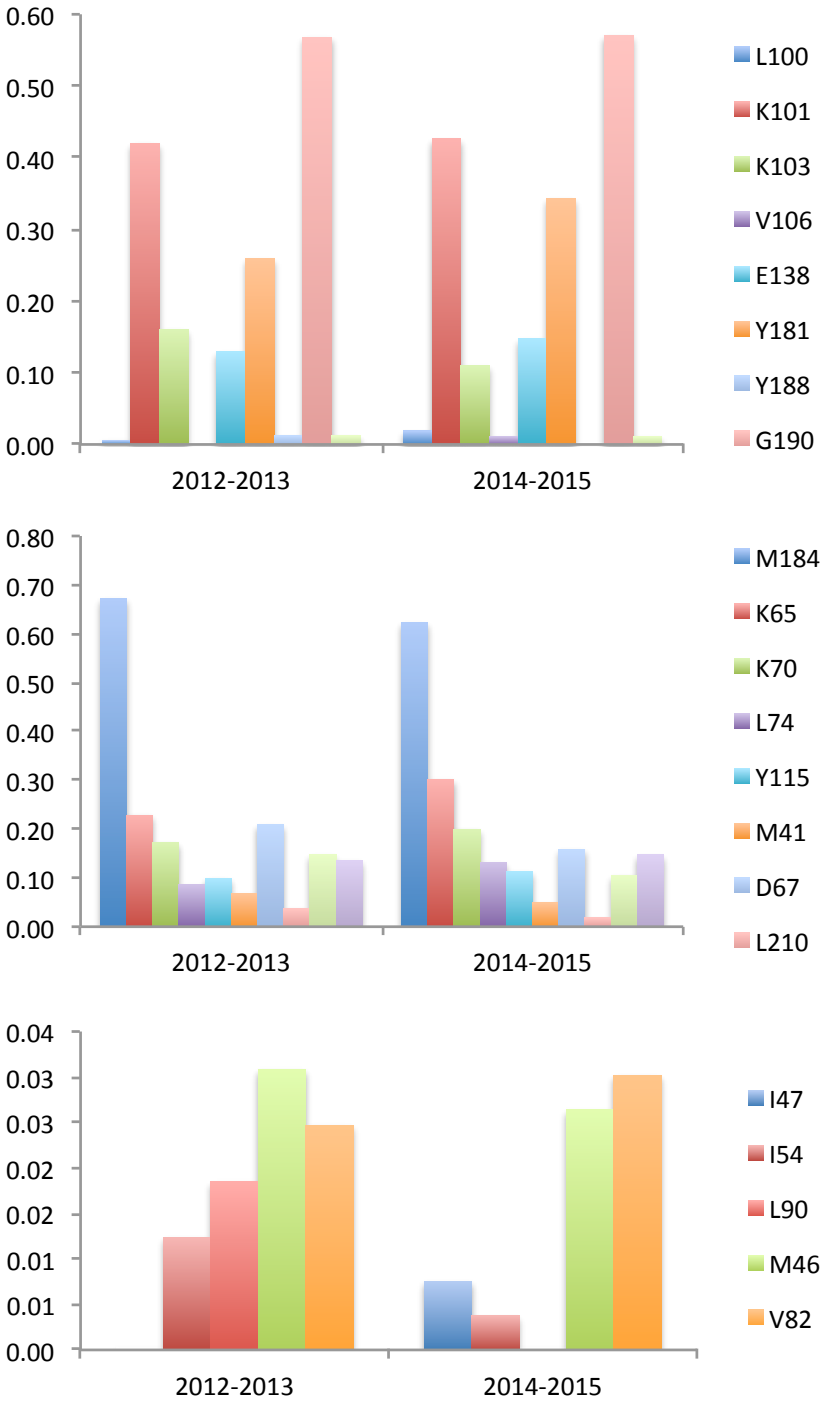


Figure 3.4. Proportion of patients with drug-resistance associated mutations in patients before and after the initiation of war. NNRTI – non-nucleoside reverse transcriptase inhibitors; NRTI - nucleoside reverse transcriptase inhibitors; PS – protease inhibitors.

Table 3.7. % of patients prescribed with tenofovir based regimen. TDF - Tenofovir Disoproxil Fumarate; FTC - Emtricitabin; EFV - Efavirenz; NVP - Nevirapine; Lpv - Lopinavir/Ritonavir; DRV - Darunavir; RTV - Ritonavir

Treatment regimen	Year				
	2012, %	2013, %	2014, %	2015, %	2016, %
TDF/FTC+EFV	5,1	9,0	11,0	15,4	20,4
TDF/FTC+NVP	0,3	0,6	0,6	0,8	0,7
TDF/FTC+Lpv/rtv	3,0	5,0	6,5	7,8	7,9
TDF/FTC+DRV+RTV	0,02	0%	0,02	0,02	0,03

3.5.5 Limitations

The sequences in our analysis do not come from a random sample of patients and only cover a tiny proportion of HIV-infected people in Ukraine. As of 2017, 79,000 HIV-infected individuals receive ART in Ukraine[59]; 7,000 (9%) of them experience treatment failure and require drug resistance test. Due to a limited availability of the tests (≈500 per year), physicians decide who gets to be tested, and this decision is based on the current state and medical history of a patient: patients with multiple changes of drug schemes get a priority. Every region has a quota of tests per year, based on the HIV prevalence and treatment coverage in the region. Unfortunately, even though the number of tests is insufficient to cover the needs of all patients, some physicians also choose to not use their quota. Thus, some test systems remain unused and the actual coverage is <500 tests per year.

We had to merge our data into 7 locations that represent the country. While there might be a further variability in behaviours within a location, we were not able to conduct a more detailed statistical analysis. It was necessary to group sequences into 7 locations because of the small number of sequences from some administrative regions. We then had to request the aggregated behavioural data that correspond to the

locations from the phylogeographical analysis. This possibly resulted in lessened statistical power to detect significant correlations between the virus movement and behaviours and limited our ability to control for potential bias.

3.6 Discussion

Ukraine has a high-prevalence HIV-epidemic that has been driven historically by PWID[1]. The conflict that the country has experienced since 2014 has resulted in large-scale internal migration and may be re-shaping the Ukrainian HIV epidemic. Here we explored the HIV dissemination in Ukraine by analysing geographical patterns of viral genetic flow in recent years. We then investigated if locations with higher rates of viral lineage import or export also have higher prevalence of risky behaviours.

Our results suggest that viral lineages have moved from the regions occupied by the war to the rest of Ukraine: many strains sampled in Ukraine during 2012-2015 appear to have originated, directly or indirectly, from Donetsk and Lugansk (i.e. the ancestral sequence location of the Ukrainian dataset was found to be in the East location). The East was the main exporter of viral lineages, and this remained the case when the number of sequences per location was made more balanced. The most intensive virus lineage movement was observed from the eastern to the central and southern regions, and to Odessa and Kyiv. The East location is heavily affected by HIV: 24% of all nationally registered HIV cases, and the largest number of HIV positive people who use injection drugs (45,000), live, or have lived in the Donetsk and Lugansk regions. Donetsk is the largest city in the area and has some of the highest HIV prevalences in Ukraine among PWID (33.5% in 2015), sex workers (17% in 2015), and in general population (0.9% of pregnant women were HIV positive in Donetsk in 2015[20,24,60]).

Our data supports a transition in the pattern of viral lineage movement around the end of 2013, the time of the beginning of several major events in Ukraine: the Maidan movement in Kyiv, the annexation of Crimea, and an initiation of the war in the east. Whilst the movement of protesters to Maidan, and of the deployment of volunteers and soldiers to the Ukrainian army (followed by the return of both to their respective places of residence) probably had an effect, civilian movement away from the affected regions was probably the most impactful[21]. This movement involved a large number of people who relocated from the area controlled by separatist forces to locations elsewhere, which corresponds to the observed virus movement patterns. Furthermore, observed viral gene flow from the occupied eastern regions is unlikely to be a result of recent transmissions from IDP or movements of recently-infected people. We believe the observed pattern instead reflects the redistribution of pre-existing infections within the country. The gene sequences in the Ukrainian HIV database come from patients who experienced ART failure and whose blood samples were sent for HIV drug resistance testing. In Ukraine before 2017, ART followed an older treatment guideline (i.e. patients received ART when CD4 count <350), hence it's likely that the patients included in our study were infected some years before their virus was sequenced. Overall, patients in Ukraine are diagnosed with HIV late in infection – 41% of the newly diagnosed patients are in the 3rd or 4th stage of HIV[61]. Thus, patients in our analysis are unlikely to reflect recent transmission events and the viral lineage movement from Donetsk and Lugansk is most likely an indication of previously infected patients that have moved. This scenario suggests that many people (e.g. residents) who were initially infected in the eastern regions have moved to the central and southern regions due to the war; resulting in a redistribution of HIV viral lineages within Ukraine.

The number of virus lineage exportations correlate well with the overall number of IDP registered in a region. Thus, we see that more people moved to the locations from which the virus is originating. This observation is explained by the fact that a large proportion of IDP moved away from the front line that divided government and separatist-controlled regions, but stayed within Donetsk and Lugansk regions, just in different counties. Surprisingly, this observation was not repeated for the HIV-infected IDP population: higher numbers of HIV-infected IDP were registered in the locations that imported more viral lineages. Thus, even though more IDP moved to the main exporting locations (East), more HIV-infected people moved to the main importing locations in our analysis (Center, Kyiv, South, Odessa). It is crucial to understand the nature of this difference: we either see more HIV-infected IDP move fully outside of the eastern regions compared to non-infected people; or we are simply not linking the HIV-infected IDP who stayed within the Donetsk and Lugansk regions to care.

The stress of the displacement might result in treatment failures for some HIV-infected patients and cause overrepresentation of IDP in our sample. Patients who had to relocate because of the conflict may be more likely to reduce treatment adherence or drop out of treatment for some time. With the non-nucleoside RT inhibitors that are currently in use in Ukraine for most first line regimens, mutations can occur within weeks of non-adherence. Indeed, the recent introduction of TDF followed by the recent increase of K65R, suggests that the time to failure in our sample is recent. On the other hand, in Ukraine 9-10% of ART patients experience virological failure which, given 79,000 patients in treatment, would lead to ~7,000 resistance tests a year. However only ~500 tests are available annually. Therefore, every region receives a quota of tests, in proportion to the number of infected people in the region.

Physicians then decide who is tested, based on the current state and medical history of a patient. Patients with multiple changes of drug schemes are prioritised. Thus a patient failing ART due to apparent adherence issues is unlikely to be prioritised for resistance testing over patients with failures under treatment. Indeed the percentage of sequences with no resistance mutations at the time of failure did not change significantly over time (17% and 20% before and after the beginning of war, respectively), suggesting that severe adherence issues are not overrepresented in our data.

Regrettably, the viral genetic data from patients of the AIDS centers in Ukraine do not have associated transmission group information. Thus, we cannot be confident as to what risk population our sample mostly represents. We know that HIV-infected PWID have poor access to treatment compared to other patients; only 9.4% of people in treatment report the use of illicit drugs at any point, whilst it is estimated that at least half of the people living with HIV in Ukraine have a history of drug use[62]. At the same time, PWID might be more likely to experience treatment failure (due to social, psychological, and economical challenges) and thus they may be more likely to be part of our sample than other patients on ART.

The genetic sequences used in our analysis formed a separate distinct clade within the viral genetic diversity observed earlier in Ukraine and in the other FSU (Figure 3.1). One of the possible explanations to this phenomenon is that our samples mostly came from a different population (i.e. sexual versus parenteral transmission groups). Another explanation lies with a systematic difference attributed to a different sequencing technique used to produce genetic data obtained in our study from the AIDS Centres. We believe that none of these reasons could have had an effect on the analyses or offer an alternative explanation to our findings.

In 2003 UNAIDS highlighted the potential contribution of conflict to the spread of HIV and other sexually transmitted diseases[63], and the topic has been intensively discussed since then[64-67]. Our finding, that viral lineage imports are correlated with some risky sexual practices, suggest that regions that have received new HIV lineages also have appropriate conditions for those lineages to flourish. IBBS data represent only urban PWID behaviours, since the survey is conducted only in regional capitals. Nonetheless, the HIV epidemic in Ukraine is largely urbanised: in 2013 77% of all HIV cases were registered among the urban population[59]. Regional capitals are also human mobility hubs that accommodate the majority of IDP[68]. There are no available regional data on sexual behaviours in the general population, but the PWID behaviour data provide a useful insight. It could be argued that the dominance of sexual HIV transmission in Ukraine (the proportion of sexually-attributed HIV incidence is growing every year[14]) can be attributed to the behaviours prevalent in the PWID group. For example, one observational study showed that in 2013-2014 in Kyiv 10% of sexually-acquired HIV occurred in patients who reported sexual contact with a drug injector[69] and contacts like these tend to be underreported. This is a widespread scenario in Eastern European countries: HIV prevailed first in PWID and then shifted towards the general population through unsafe sex practices[62,70,71]. The war and internal movement of people to regions with more reported high-risk sexual behaviours among PWID might have a catalytic effect on this process and create a favourable environment for new HIV transmission. Local HIV outbreaks should be actively monitored in the years following the war.

Our findings suggest that certain public health strategies should be prioritised during and after the conflict. Ukrainian non-governmental organisations have already made a significant effort to include IDP-representatives of certain transmission groups into

specialised preventive services; in particular, patients of OST programs were actively linked to care in their new places of residence[19]. A pro-active integration into routine HIV testing of people who have relocated due to the war, or who frequently travel to the war zone, should become a component of HIV prevention. Additionally, equal and timely access to treatment should be granted to people directly affected by the conflict. Our results suggest that risky behaviours among PWID might still be a concern, despite the successful harm reduction programs, especially in the regions with high rate of virus lineage importation. Consistent use of clean syringes and condoms should be promoted not only for PWID, but also for their sexual partners and the general population when appropriate. Finally, the observed increase of K65R, which might undermine the potential of TDF-based regimens for Pre-Exposure Prophylaxis in Ukraine, underlines a clear need for TDF-based regimens to be combined with more potent last generation antiretrovirals in order to protect against the emergence of antiretroviral resistance.

3.7 Conclusions

Recent movement of HIV infected people might change existing network structures, disturbing the "firewall" effect and accelerating HIV flows within Ukraine. Combined with risky PWID behaviours in the regions with highest number of lineage imports, this might facilitate the generalisation of the Ukrainian epidemic. Enabling sustainable prevention services and treatment provision in locations where services have been physically disrupted because of the armed conflict is a priority[72]. However, we want to highlight the necessity of urgent preventive measures in other parts of the country as well. Since our analysis likely reflects the redistribution of existing infections, not the location of new infections, we suggest that there is a need to

continuously monitor the effect of the war on HIV incidence and prevalence in Ukraine at both national and regional levels.

3.8 Ethical approval and acknowledgements

HIV drug resistance testing was performed at the L. V. Gromashevskij Institute of Epidemiology and Infectious Diseases of National Academy of Sciences of Ukraine following all Ukrainian legal and ethical requirements. Fully anonymised genetic sequence data were transferred to Oxford through a Material Transfer Agreement for molecular evolution analyses. Fully anonymised aggregated behavioural data (IBBS) were obtained from the Alliance for Public Health. The IBBS protocol and questionnaires were examined by the Committee on Sociologist's Professional Ethics at the Sociological Association of Ukraine. Epidemiological component was examined by the Committee on Medical Ethics of the L. V. Gromashevskij Institute of Epidemiology and Infectious Diseases of National Academy of Sciences of Ukraine.

We would like to thank Louis du Plessis of the Department of Zoology of the University of Oxford for his assistance. T.I.V. is supported by the Clarendon Fund and Hertford College of the University of Oxford, S.R.F. is supported by the NIH NIDA (Grant number DP1DA034989), G.M. is supported by an MRC Clinician Scientist Fellowship (MR/K010565/1). IBBS data collection was administered by the Ukrainian Institute for Social Research after O. Yaremenko (UISR) under an order from the ICF "Alliance for Public Health".

3.9 References

1. ECDC. HIV/AIDS surveillance in Europe 2015. Stockholm: European Centre for Disease Prevention and Control/WHO Regional Office for Europe; 2016.
2. Kuzio T. Crime, politics and business in 1990s Ukraine. Communist and Post-Communist Studies 2014;47:195-210.
3. Galli C. Trends in Europe and North America: The statistical yearbook of the economic commission for Europe 1998/1999. Journal of Government Information 2000;27:650-3.
4. Poznyak VB, Pelipas VE, Vievski AN, Miroshnichenko L. Illicit drug use and its health consequences in Belarus, Russian Federation and Ukraine: Impact of transition. European Addiction Research 2002;8:184-9.
5. Shelley LI. Organized Crime and Corruption in Ukraine: Impediments to the Development of a Free Market Economy. Demokratizatsiya 1998;6:648-63.
6. Ministry of Health of Ukraine. HIV/AIDS Epidemic in Ukraine. 2016.
7. Hamers FF. HIV infection in Ukraine (1987-96). Rev Epidemiol Sante Publique 2000;48 Suppl 1:1S3-1S15.
8. Barnett T, Whiteside A, Khodakevich L, Kruglov Y, Steshenko V. The HIV/AIDS epidemic in Ukraine: its potential social and economic impact. Social Science & Medicine 2000;51:1387-403.
9. Booth RE, Kwiatkowski CE, Mikulich-Gilbertson SK, et al. Predictors of risky needle use following interventions with injection drug users in Ukraine. Drug and Alcohol Dependence 2006;82:S49-S55.
10. Dehne KL, Grund JPC, Khodakevich L, Kobyshcha Y. The HIV/AIDS epidemic among drug injectors in Eastern Europe: Patterns, trends and determinants. Journal of Drug Issues 1999;29:729-76.
11. WHO. Good practices in Europe: HIV prevention for People Who Inject Drugs implemented by the International HIV/AIDS Alliance in Ukraine. Copenhagen, Denmark: WHO Regional Office for Europe; 2014.
12. MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and Hepatitis C in people who inject drugs: A review of reviews to assess evidence of effectiveness. International Journal of Drug Policy 2014;25:34-52.
13. Dvoriak S, Douglas B, Sylla L, Altice F. Opioid Substitution Therapy With Buprenorphine in Ukraine—Program Description and Policy Implications. National Institute on Drug Abuse; 2007.
14. Vitek CR, Cakalo JI, Kruglov YV, et al. Slowing of the HIV Epidemic in Ukraine: Evidence from Case Reporting and Key Population Surveys, 2005-2012. Plos One 2014;9.
15. Pizzi M. War in Ukraine threatens to worsen HIV crisis. Aljazeera America2015.
16. Holt E. Health care collapsing amid fighting in east Ukraine. Lancet 2015;385:494.
17. Owczarzak J, Karelin M, Phillips SD. A view from the frontlines in Slavyansk, Ukraine: HIV prevention, drug treatment, and help for people who use drugs in a conflict zone. International Journal of Drug Policy 2015;26:6-7.
18. Filippovych S. Impact of armed conflicts and warfare on opioid substitution treatment in Ukraine: responding to emergency needs. Int J Drug Policy 2015;26:3-5.

19. Duchenko A, Deshko T, Braga M. Crisis management by HIV/AIDS non-governmental organisations in the post-Euromaidan Ukraine led to opening new horizons. Drugs and Alcohol Today 2017:17(3).
20. Alliance for Public Health. Annual Report. Kyiv, Ukraine: ICF "Alliance for Public Health"; 2015.
21. International Displacement Monitoring Centre. Ukraine: Country Information 2015. Accessed: 2017.
22. State Statistics Service of Ukraine. All-Ukrainian population census. 2016.
23. Ministry of Health of Ukraine. HIV-infection in Ukraine. Informational bulletin N45. Kyiv: Ukrainian Center for Control for Socially Dangerous Diseases of Ministry of Health of Ukraine; Institute of Epidemiology and Infectious Diseases named after L.V. Gromashevskiy; 2016.
24. Bozicevic I, Handanagic S, Cakalo JI, Rinder Stengaard A, Rutherford G. HIV Strategic Information in Non-European Union Countries in the World Health Organization European Region: Capacity Development Needs. JMIR Public Health Surveill 2017;3:e41.
25. Friedman SR, Kottiri BJ, Neaigus A, Curtis R, Vermund SH, Des Jarlais DC. Network-related mechanisms may help explain long-term HIV-1 seroprevalence levels that remain high but do not approach population-group saturation. American Journal of Epidemiology 2000;152:913-22.
26. Khan B, Dombrowski K, Saad M, McLean K, Friedman S. Network Firewall Dynamics and the Subsaturation Stabilization of HIV. Discrete Dyn Nat Soc 2013;2013:720818.
27. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 2016;33:1870-4.
28. Bennett DE, Camacho RJ, Otelea D, et al. Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. Plos One 2009;4.
29. Pineda-Pena AC, Faria NR, Imbrechts S, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol 2013;19:337-48.
30. Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res 2014;42:e144.
31. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947-8.
32. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312-3.
33. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.
34. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.
35. Yu G, Smith DK, Zhu H, Guan Y, Tsan-Yuk Lam T. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 2016;8.
36. Ministry of Health of Ukraine. HIV-infection in Ukraine. Informational bulletin N41. Kyiv: Ukrainian Center for Control for Socially Dangerous Diseases of Ministry of Health of Ukraine; Institute of Epidemiology and Infectious Diseases named after L.V. Gromashevskiy; 2014.
37. UNAIDS. Ukraine Global AIDS Response Progress Reporting. 2015.

38. Berleva G, Dumchev K, Kasianchuk M, Nikolko M, Saliuk T, Shvab I. Analytical report: "Estimation of the Size of Populations Most-at-Risk for HIV Infection in Ukraine" as of 2012 based on the results of 2011 survey. Kyiv: Alliance for Public Health; 2012.
39. Balakiryeva OM, Nondar TV, Sereda YV, Sazonova YO. Analytical report "Behaviour monitoring and HIV-prevalence among injecting drug users as a component of secondgeneration surveillance" (based on results of the 2011 biobehavioral survey). Kyiv: International HIV/AIDS Alliance; 2012.
40. UN High Commissioner for Refugees. Preliminary analysis and needs assessment of Internally displaced persons2014.
41. Nizova NM, Sherbinska AM, Hetman LI, et al. HIV infection in Ukraine. Informational Bulletin. Kyiv: Ukrainian Center to control socially dangerous diseases of Ministry of Health of Ukraine; 2015.
42. State Statistics Service of Ukraine. Statistics 2016.
43. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 2012;29:1969-73.
44. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. Plos Comput Biol 2009;5.
45. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology 2006;4:e88.
46. Swofford DL. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts.: Sinauer Associates; 2002.
47. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol 2008;56:391-412.
48. O'Brien JD, Minin VN, Suchard MA. Learning to count: robust estimates for labeled distances between molecular sequences. Mol Biol Evol 2009;26:801-14.
49. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56-61.
50. Raghwani J, Rambaut A, Holmes EC, et al. Endemic dengue associated with the co-circulation of multiple viral lineages and localized density-dependent transmission. PLoS Pathog 2011;7:e1002064.
51. Bielejec F, Lemey P, Baele G, Rambaut A, Suchard MA. Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. Syst Biol 2014;63:493-504.
52. Gelman A, Meng XL. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat Sci 1998;13:163-85.
53. Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst Biol 2011;60:150-60.
54. Jeffreys H. Theory of probability. Third edition. Oxford: Clarendon Press; 1961.
55. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res 2001;125:279-84.
56. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Philos Trans R Soc Lond B Biol Sci 2013;368:20120196.

57. Ruane PJ, Luber AD. K65R-associated virologic failure in HIV-infected patients receiving tenofovir-containing triple nucleoside/nucleotide reverse transcriptase inhibitor regimens. MedGenMed 2004;6:31.
58. del Rio C. Current concepts in antiretroviral therapy failure. Top HIV Med 2006;14:102-6.
59. Ministry of Health of Ukraine. Overall number of people reciving ART in Ukraine as of 01.04.2017.
60. Balakirieva O, Bondar T, Loktieva I, Sazonova Y, Sereda Y. Summary of the analytical report "Monitoring the behaviour and HIV-infection prevalence among people who inject drugs as a component of HIV second generation surveillance". 2014.
61. Ministry of Health of Ukraine. Report on monitoring and evaluation of effectiveness of national program agains HIV/AIDS for 2014-2018, as of 2015. 2016.
62. Graw F, Leitner T, Ribeiro RM. Agent-based and phylogenetic analyses reveal how HIV-1 moves between risk groups: Injecting drug users sustain the heterosexual epidemic in Latvia. Epidemics 2012;4:104-16.
63. UNAIDS Office on AIDS, Society, and Humanitarian Response. HIV/AIDS and Conflict. Copenhagen, Denmark 2003.
64. Mock NB, Duale S, Brown LF, et al. Conflict and HIV: A framework for risk assessment to prevent HIV in conflict-affected settings in Africa. Emerg Themes Epidemiol 2004;1:6.
65. Tripodi P, Patel P. HIV/AIDS, peacekeeping and conflict crises in Africa. Med Confl Surviv 2004;20:195-208.
66. Mills EJ, Singh S, Nelson BD, Nachega JB. The impact of conflict on HIV/AIDS in sub-Saharan Africa. International Journal of STD and AIDS 2006;17:713-7.
67. Betsi NA, Koudou BG, Cisse G, et al. Effect of an armed conflict on human resources and health systems in Cote d'Ivoire: prevention of and care for people with HIV/AIDS. AIDS Care 2006;18:356-65.
68. Smal V, Poznyak O. Internally displaced people: social and economical integration in host communities based on data from Vinnytsya, Zaporizzhya, and Ivano-Frankivsk regions. Kyiv 2016.
69. Simmons R, Malyuta R, Chentsova N, et al. HIV Testing and Diagnosis Rates in Kiev, Ukraine: April 2013-March 2014. PLoS One 2015;10:e0137062.
70. Toussova O, Shcherbakova I, Volkova G, Niccolai L, Heimer R, Kozlov A. Potential bridges of heterosexual HIV transmission from drug users to the general population in St. Petersburg, Russia: is it easy to be a young female? J Urban Health 2009;86 Suppl 1:121-30.
71. Des Jarlais DC, Feelemyer JP, Modi SN, et al. Transitions from Injection-Drug-Use-Concentrated to Self-Sustaining Heterosexual HIV Epidemics: Patterns in the International Data. Plos One 2012;7.
72. UNAIDS. Operational research: Systems of service to prevent and treat HIV/AIDS under the situation of the armed conflict in the east of Ukraine. Kyiv: UNAIDS; 2016.

4 The changing epidemiological profile of HIV-1 subtype B epidemic in Ukraine

Parts of this chapter were published in:

Tetyana I Vasylyeva, Mariia Liulchuk, Louis du Plessis, Esther Fearnhill, Victoriia Zadorozhna, Nataliia Babii, Alla Scherbinska, Vladimir Novitsky, Oliver G Pybus, Nuno R Faria. The Changing Profile of HIV-1 Subtype B Epidemic in Ukraine. AIDS Res Hum Retroviruses. 2018.

Summary of authorship:

TIV wrote this chapter, conducted phylogenetic analyses and the simulation study, and created all figures in this chapter. LdP provided technical assistance with the simulation study. ML and EF provided genetic data from the Ukrainian HIV database and the CASCADE database, respectively. OGP and NRF provided supervisory support. All others helped with the interpretations of the results.

4.1 Abstract

While HIV-1 subtype B has caused a large epidemic in the western world, its transmission in Ukraine remains poorly understood. Here I assessed the genetic diversity of HIV-1 subtype B viruses circulating in Ukraine, characterized transmission group structure, and estimated key evolutionary and epidemiological parameters. I also conducted a small simulation study to assess the performance of various phylodynamic models. Specifically, I analysed 120 HIV-1 subtype B *pol* sequences (including 46 newly-generated) sampled from patients residing in 11 regions of Ukraine between 2002-2017. Phylogenies were estimated using maximum likelihood and Bayesian phylogenetic methods. A Bayesian molecular clock coalescent analysis was used to estimate effective population size dynamics and to date the most recent common ancestors of identified clades. A phylodynamic birth-death model was used to estimate the effective reproductive number (R_e) of these clades. I identified two phylogenetically distinct predominantly Ukrainian (≥75%) clades of HIV-1 subtype B. I found no significant transmission group structure for either clade, suggesting frequent mixing among transmission groups. The estimated dates of origin of both subtype B clades were around early 1970s, similar to the introduction of HIV-1 subtype A into Ukraine. R_e for Clade 1 was estimated to be 1.42 (95% HPD 1.26-1.56) and 1.69 (95% HPD 1.49-1.84) for Clade 2. The simulation study showed that epidemic size, epidemic growth rate, and the sampling scheme had an effect on the performance of phylodynamic models, but the sampling proportion did not. The study results highlight the necessity for strengthening preventive and monitoring efforts to reduce the further spread of HIV-1 subtype B.

4.2 Introduction

Ukraine has one of the biggest HIV epidemics among European countries: in 2015 the country had the highest rate of new HIV infections in Europe (~30 cases per 100,000 people per year)[1]. In the 1990s, people who inject drugs (PWID) were the transmission group that accounted for the majority of new HIV infections in Ukraine[2,3]. However, since the late 2000s, transmission among heterosexual, non-drug injecting populations has prevailed[4]. Even though HIV-1 subtype B and CRF03_AB were first detected in the country in the 1990s, the majority of HIV-1 infections in Ukraine are caused by HIV-1 subtype A[5]. In 2015, 89% of HIV-1 sequences in the Ukrainian HIV Drug Resistance database were identified as HIV-1 subtype A (see Chapter 3). Similarly, 88% of Ukrainian sequences in the Los Alamos HIV sequence database are classified as HIV-1 subtype A[6]. To date, most molecular epidemiology studies of the Ukrainian epidemic have focused on HIV-1 subtype A in PWID populations (see Chapters 2, 3, and [7,8]).

HIV-1 subtype B, which is the most prevalent HIV subtype in most Western European countries[9], has a limited distribution and prevalence in Ukraine compared to HIV-1 subtype A. In 2012-2015, only 4% (19 out of 448) of sequences in the Ukrainian Drug Resistance database were identified as HIV-1 subtype B (see Chapter 3), as were <10% of Ukrainian sequences in the Los Alamos HIV database. However, extrapolation of these proportions (4-10%) to the estimated 220,000 HIV-infected people in the country[1] suggests that HIV-1 subtype B still accounts for a sizeable number of HIV infections in Ukraine (between 9,000-22,000 infections).

The genetic diversity of HIV-1 subtype B in Ukraine seems to be characterised by two different co-circulating lineages: one is diagnosed mostly in PWID and characterised by very low genetic diversity, whilst another is more genetically similar to subtype B

strains observed in Western European countries[5,10]. The first lineage was reported mainly in Mykolaiv, a port city in the south of Ukraine, but was also observed in some other countries of the former Soviet Union (FSU) and was subsequently called B_{FSU}[5,11]. Two studies from Russia and Azerbaijan have shown that local HIV-1 subtype B infections among PWID clustered with viral sequences previously diagnosed in patients in Mykolaiv and Kyiv[12,13]. A recent study demonstrated high levels of phylogenetic clustering (66%) among men who have sex with men (MSM) and heterosexual men infected with HIV-1 subtype B in Kyiv, the capital of Ukraine[14].

The epidemiology of HIV-1 subtype B transmission across different groups, such as MSM and PWID, and among geographic regions in Ukraine remains poorly understood. In this study, I describe the dynamics of the HIV-1 subtype B epidemic in 11 Ukrainian regions, between 2002 and 2017. I apply molecular epidemiological methods to estimate the dates of introduction of HIV-1 subtype B into Ukraine, and assess the transmission dynamics of the most prevalent HIV-1 subtype B lineages. I also assess and compare the performance of various phylodynamic models in estimating epidemiological parameters by means of a simulation study.

4.3 Methods

4.3.1 Data

4.3.1.1 Genetic data

Ukrainian HIV-1 subtype B genetic sequences were derived from three sources. Firstly, I report 46 new HIV-1 subtype B *pol* gene sequences (accession numbers MH577242-MH577287) from patients visiting Ukrainian AIDS Centres (the hospitals for HIV treatment and care) in 2012-2017, which were made available through the

Ukrainian HIV drug resistance database. These data were sequenced at the L. V. Gromashevskij Institute of Epidemiology and Infectious Diseases in Kyiv using ViroSeq® HIV-1 Genotyping System v2.0 Kit (Celera Corporation, Abbott). These 46 sequences represent 11 administrative regions of Ukraine: Mykolaiv (N=24), Kyiv (N=6), Odessa (N=3), Vinnytsya (N=3), Dnipro (N=2), Kharkiv (N=2), Sumy (N=2), Ivano-Frankivsk, Kherson, Lugansk and Zhytomyr (N=1 each). Second, I used 73 HIV-1 subtype B *pol* sequences collected as part of the CASCADE (Concerted Action on SeroConversion to AIDS and Death in Europe) cohort study in Kyiv in 2013-2015[14]. These data were collected anonymously in four infectious disease clinics that form part of the Kyiv AIDS Centre[15] and were sequenced at University College London Hospital laboratories on Sanger or Illumina Miseq platforms. Third, I accessed the Los Alamos HIV database[7] and retrieved all available HIV-1 subtype B sequences collected in Ukraine that spanned the same genomic region (HXB2 positions: 2568-3256) and which had available sampling dates (N=45) and transmission group information (N=9, out of 45 sequences with available sampling dates). Only HIV-1 subtype B sequences with <5% ambiguous nucleotides were retained. Duplicate sequences were removed using the ElimDupes tool from the Los Alamos website (https://www.hiv.lanl.gov/). I used the REGAv3[16] and COMET[17] subtyping tools to confirm correct HIV subtype assignment. CRF03_AB, previously found in Ukraine (see Chapter 3 and [5]), has a break-point in *pol* gene, which makes *pol* sequences appropriate for a correct HIV subtype identification.

After removing duplicates, the *Ukrainian dataset* contained 120 HIV-1 subtype B sequences with matching socio-demographic and epidemiological data (gender, transmission group, date of sampling, city of sampling). I then downloaded the 10 closest reference sequences to the sequences in the Ukrainian dataset, using the

BLAST[18] tool from the Los Alamos database website. After removing duplicates, these sequences and the Ukrainian dataset comprised the *combined dataset* (N=299 sequences). The reference sequences came from Azerbaijan, Argentina, Australia, Belgium, Brazil, Bulgaria, Belarus, Canada, China, Cyprus, Czech Republic, Germany, Denmark, Estonia, Finland, France, Great Britain, Gergia, Italy, Jamaica, Japan, Kenya, Latvia, Luxemburgh, Mexico, Panama, Poland, Russia, Serbia, Slovenia, South Korea, Spain, Sweden, Thailand, USA,Venezuela. The accession numbers for the reference sequences are provided in Appendix 4.

4.3.1.2 Alignment

I codon-aligned the sequences using Clustal Omega[19] and then manually edited the alignment using MEGA7.0[20]. I deleted 43 codon positions associated with drug resistance[21]. The final alignment consisted of 614 nucleotides that correspond to nucleotide positions 2568-3256 of the HXB2 HIV reference sequence (*pol* gene). Mean pairwise genetic distances were calculated in MEGA using a Maximum Composite Likelihood nucleotide substitution model.

4.3.1.3 Epidemiological data

The annual number of all newly registered HIV-infections (all subtypes) by transmission group was available from the Ministry of Health of Ukraine website[3]. Annual data were available starting from 1995.

4.3.1.3 Statistical analysis

To test for possible associations between socio-demographic and other characteristics, and the membership of sequences in the dominant circulating clades, I performed Fisher's exact tests. To account for multiple comparisons between different groups of the same variable I used a Bonferroni multiple test correction: only p-values $\leq$ (0.05/the number of categories in the variable) were considered significant.

4.3.2 Phylogenetic analysis

4.3.2.1 Maximum likelihood phylogenies

I estimated maximum likelihood (ML) phylogenetic trees for the Ukrainian and the combined alignment using RAxML[22]. I used a general time reversible (GTR) nucleotide substitution model with gamma distributed rate variation among sites. In the combined dataset I identified two clades with >50 sequences, at least 75% of which were from Ukraine. Both clades exhibited SH-like statistical support values >0.80[23], which are considered statistically significant[24]. I used RAxML to estimate ML trees for both clades, after removing sequences that were not sampled in Ukraine. I used TempEst[25] to evaluate root-to-tip divergence in the combined dataset and for the two Ukrainian clades. I used BaTS[26] to calculate the association index (AI) for the phylogenies of the two Ukrainian clades. The AI is a statistic that describes the association between a phylogeny and traits at its tips (transmission groups in this case). BaTS also provides statistical support (p-values) for the AI statistic by comparing the values obtained from the posterior tree distribution to those obtained by randomly re-assigning traits at the tips 100 times.

4.3.2.2 Time-calibrated phylogenies

I estimated evolutionary parameters and time-calibrated phylogenies using the BEASTv1.8.4 software package[27]. The combined dataset was used to estimate the time of the most recent common ancestor (TMRCA) of the two Ukrainian clades. To identify the most appropriate coalescent and molecular clock models I used path-sampling[28] and stepping stone[29] marginal likelihood estimators (MLE). I assessed the fit of two molecular clock models (strict and lognormal relaxed) and of three coalescent demographic models (constant population size, exponential growth, and Bayesian skyline). The Bayesian skyline model[30] with a lognormal relaxed clock[31]

resulted in the best fit (Table 4.1) and was used in further analyses. I used a GTR nucleotide substitution model with gamma distributed rate variation among sites. Markov Chain Monte Carlo (MCMC) analyses of the combined dataset were run for 200 million generations (with a 10% burn-in). I then ran the Bayesian skyline analysis for 150 million generations (with a 10% burn-in) under the same evolutionary model for the two Ukrainian clades separately, in order to estimate fluctuations in effective population size (N_e) for these two clades. Convergence of the MCMC sampler was inspected using Tracer[32].

Table 4.1. Model selection for the combined dataset

Molecular clock model	Coalescent prior	Path-sampling	Stepping-stone
Lognormal Relaxed	**Bayesian Skyline**	**-24270**	**-24290**
Lognormal Relaxed	Exponential growth	-24328	-24340
Lognormal Relaxed	Constant size	-24509	-24520
Strict	Bayesian Skyline	-24354	-24372
Strict	Exponential growth	-24418	-24431
Strict	Constant size	-24610	-24622

4.3.2.3 Estimates of the effective reproductive number

The effective reproductive number (R_e) of an epidemic is the average number of secondary infections attributable to an infected person. This value is often used to describe transmission dynamics in a population; $R_e > 1$ means the epidemic is growing, while $R_e < 1$ means the epidemic is declining. I used two phylodynamic birth-death models (BDM) implemented in BEASTv2.4.7[33] to estimate R_e for the Ukrainian epidemic. First, I applied a constant rate birth-death model to the two Ukrainian clades[34]. This model assumes R_e to be constant through time, resulting in an average R_e for each clade. I then applied a birth-death skyline (BDSKY) model[35] to all Ukrainian sequences with available sampling dates (N=164). This model assumes

a piecewise constant R_e over 10 equidistant time-intervals, between the origin of the tree and the date of the most recent sequence[35].

I used the following specifications for both BD models: (i) a GTR nucleotide substitution model with gamma distributed rate variation among sites; (ii) a uniform prior for the date of origin, which is the duration of the epidemic represented by the sequences in the analysis, with an upper bound of 100; (iii) sampling proportion=0 before the first sampling date (2002 for both clades, and 2001 for the full Ukrainian dataset), after that date I applied a beta distribution prior (α=1.0, β=10.0) to the sampling proportion parameter, which was constrained between 0.1% and 10%; (iv) a lognormal distribution prior for R_e (mean=1.0, standard deviation = 1.25); and (v) a normally distributed TMRCA prior, based on the values of the TMRCA estimated in the previous analysis using the combined dataset (see section 4.4.2). I also assumed that the infectious period, which is the period between contracting the virus and becoming uninfectious (being in treatment and virally suppressed, or dying), was, on average, 4 years for a patient in Ukraine; this parameter was fixed. To assess the robustness of my results to this assumption, I performed additional sensitivity analyses for this parameter for the constant rate BDM by considering four additional values for the infectious period: 1, 2, 6 and 8 years, which correspond to rates of becoming uninfectious of 1.0, 0.5, 0.167, and 0.125 per year. I ran the MCMC analyses for 150×10^6 generations (with a 10% burn-in) for the constant rate BDM, and for 400×10^6 generations (20% burn-in) for the BDSKY analysis.

4.3.3 Simulation study

4.3.3.1 Simulation study rationale and aims

In this chapter and in Chapter 2 I used Bayesian Skyline plots (BSP)[30] to estimate the changes of effective population size (N_e) through time. The Bayesian Skyline model

is a piecewise-constant demographic function, originally described as a method to investigate and inform the selection of an appropriate parametric model[30], but whose independent use is now common practice. In this thesis I have applied BSP to study two very different epidemics from the same region: a large Russian-Ukrainian subtype A epidemic and a much smaller Ukrainian subtype B epidemic. For HIV-1 subtype A from Russia and Ukraine, genetic data were available starting from 1993, and more data were sampled since then with a regular interval of about 2-3 years, with a more dense sampling in the last ten years. Interestingly, the BSPs reconstructed from *pol* versus *env* genetic sequences from Russia and Ukraine, presented in Chapter 2 (Fig. 2.3), showed different N_e trajectories through time, which I think might be explained by a much wider geographical and temporal representation of the *pol* sequences. For HIV-1 subtype B from Ukraine, on the contrary, genetic data were only available starting from 2001-2002, with more data available after 2010. To assess how the BSP performs when estimating changes in population dynamics for different epidemic scenarios, and under different sampling approaches, I conducted a simulation study.

Furthermore, analysing the Ukrainian subtype B epidemic, I applied a more novel approach, the Birth-Death Skyline (BDSKY) model, to estimate changes in reproductive number of the epidemic through time. The BDSKY is becoming more widely used in infectious disease epidemiology, but concerns have been raised that BDSKY estimates can be subject to bias if some of the model parameters are misspecified[36]. Since the sampling proportion parameter used in the BDSKY analysis in this chapter did not directly reflect the sampling approach used to obtain the Ukrainian sequences, I included the BDSKY model as one of the inference approaches tested in the simulation study, to investigate any possible bias.

Thus, I performed a simulation study to assess the performance of different phylodynamic models in reconstructing HIV population dynamics, under different epidemiological and study design scenarios. In particular, I aimed to compare the performance of a parametric coalescent-based model to that of the non-parametric Bayesian Skyline model. A parametric model assumes a specific function of population growth, described by a small number of parameters (e.g. the constant, exponential, logistic growth models)[37]. A combination of several parametric models can be used to describe more complex population growth scenarios, with multiple discrete changes in population dynamics. I chose a constant-exponential-constant (CEC) model, previously applied to phylodynamic analyses of Hepatitis C epidemics[38] (Fig. 4.1).

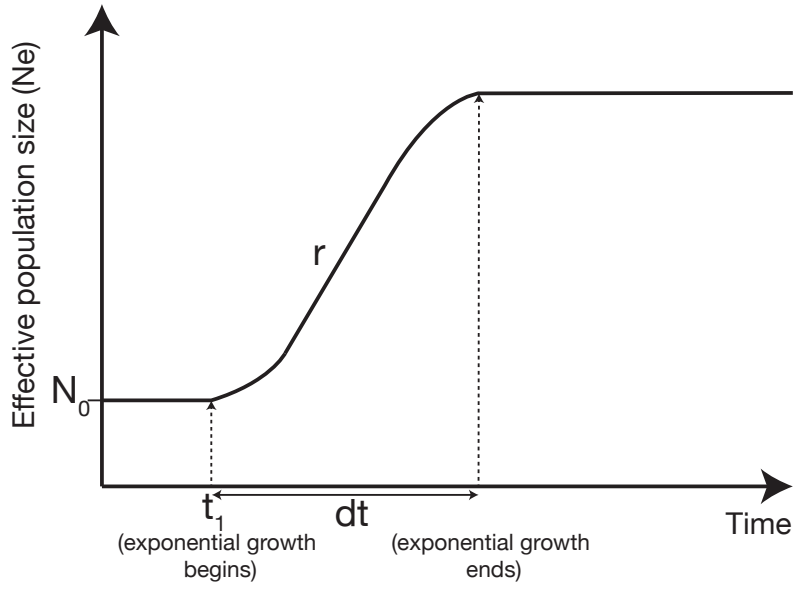


Figure 4.1. An illustration of a CEC model and the four parameters that describe it: population size at the initial stage (N_0), exponential growth rate (r), time of the start of the exponential growth (t_1), and the duration of the exponential growth phase (dt). Adapted from Pybus et al[38] (Licence to reproduce the material granted on December 19th 2018).

The CEC model is a demographic model that contains three phases, an initial period of low constant population size, followed by a period of exponential growth, and concluded by another period of constant population size. This model is fully described by four parameters: the population size at the initial stage (N_0), the exponential growth rate (r), the time of the start of the exponential growth (t_1), and the duration of the exponential growth phase (dt)[38].

Finally, I also applied the BDSKY model to the simulated data, to study the accuracy of R_e estimates obtained from BEAST under different epidemic scenarios.

4.3.3.2 Simulation scenarios

I used the Moments and Stochastic Trees from Event Reactions (MASTER) stochastic simulator[39], implemented as a package in BEASTv2.4.7[33], to simulate phylogenies and population dynamics under a logistic growth demographic model. MASTER simulates a population growth process using a Chemical (or Combinatoric) Master Equation (CME). MASTER simultaneously produces trajectories of population size and phylogenetic trees, representing the lineage histories of the simulated populations. I simulated a viral epidemic under a stochastic logistic growth model with a linear birth rate (λ) and a non-linear death rate (ϒ), defined by the following equations:

$$X \xrightarrow{\lambda} 2X \qquad (4.1)$$

$$2X \xrightarrow{\Upsilon} X \qquad (4.2)$$

where X is an infected individual. The birth reaction, at rate λ, is linear and constant over the course of the simulation. However, the death reaction, at rate ϒ, is density-dependent and increases as the population grows, resulting in the effective population

growth slowing down (since two individuals need to interact to result in a death event). Eventually, the birth rate equals the effective death rate and the population size stabilizes around a quasi-steady state. At this stage the population size fluctuates stochastically around an equilibrium value. The death rate (Υ) is composed of the sum of the removal rate (becoming non-infectious through death or treatment initiation, μ) and the sampling rate, Ψ. To approximate the CEC-like dynamic, I assumed two different death rates: a higher rate for the initial constant population size phase (Υ_1), and a much lower rate starting from the time after the start of the exponential growth phase t_1 (Υ_2) (Table 4.2).

Overall, I simulated six scenarios of epidemic growth over 40 years (Table 4.2). I stopped all the simulations at year 40, which resulted in stochastically different number of tips in the resulting phylogenies.

Table 4.2. Parameters of the epidemic growth scenarios

Scenario	λ	t_1	Υ_1	Υ_2	Ψ	Sampling scheme
1	1.0	10	0.1	0.00002	0.05%	Constant through time
2	1.0	10	0.1	0.00002	0.05%	Varying through time
3	0.5	5	0.1	0.0001	0.25%	Constant through time
4	0.5	5	0.1	0.0001	1.0%	Constant through time
5	0.5	5	0.1	0.0001	0.25%	Varying through time
6	0.5	5	0.1	0.0001	1.0%	Varying through time

I specified these scenarios to reflect variability in:

- The sampling proportion (proportion of sampled sequences out of the HIV-infected population)
- The sampling scheme (constant through time versus varying through time). I refer to the even sampling of a certain proportion of individuals through time, from year 0 to year 40, as “constant through time”. By “varying through time” I mean no sampling events occur before year 20, then a certain proportion of infections are

sampled between years 20 and 21, then sequences are sampled between years 21 and 30, and then the same proportion of individuals as between years 20 and 21 is sampled after year 30 through the end of the simulation (year 40)

- The size of the epidemic
- The epidemic growth rate through time

Scenarios 1 and 2 represent a rapid growth situation, similar to that observed for the subtype A epidemic in the 1990s (Table 4.2). Scenarios 3-6 represent a slower growth scenario and a 10-times smaller epidemic, which is similar to the Ukrainian subtype B epidemic. For all six scenarios I repeated the simulation 10 times.

4.3.3.3 BEAST analyses

I performed three sets of independent BEAST analyses of the phylogenies simulated under each of the scenarios. For all of the BEAST analyses, I fixed the tree to the phylogenies obtained from simulations. I did not simulate nucleotide sequences along the branches of the simulated phylogenies, and, thus no molecular clock or sequence evolution models were included in the analyses. The analyses therefore assume the tree is known without error. In all BEAST analyses, every 10,000th state was sampled. Convergence of the MCMC sampler was inspected using Tracer[32].

I used the following specifications for the phylodynamic analyses:

1. Bayesian Skyline model[30]. I ran MCMC analyses for all the scenarios for 100-300x10^6 states (with a 10% burn-in), using BEASTv1.8.4[27].

2. CEC model[38]. I ran MCMC analyses for all the scenarios for 5-20x10^6 states (with a 10% burn-in) using BEASTv1.8.4. I used normal prior distributions for the time of the start of the exponential growth t_1 parameter (mean=20 and standard deviation=20) and the duration of the exponential growth phase (mean=20 and standard deviation=10), a Laplace prior distribution for the exponential growth rate r

(mean=0 and scale=1), and a 1/x prior for the population size at the initial stage N_0 parameter.

3. BDSKY model[35]. I ran MCMC analyses for all the scenarios for 5-30x10^6 states (with a 10% burn-in) using BEASTv2.4.7[33]. I parameterised the model to estimate R_e at eight equidistant times, between time=0 and time=35. I used lognormal prior distributions for the becoming uninfectious rate parameter (mean=0 and standard deviation=1.0) and the R_e parameter (mean=0 and standard deviation=1.0), a uniform prior on the origin parameter, and a Beta prior distribution for the sampling proportion parameter (α=1.0 and β=1.0).

4.3.3.4 Interpreting the simulation study results

I used several approaches to compare the effective population size (N_e) trajectories estimated by the Bayesian Skyline and the CEC models in BEAST to the number of infected individuals simulated in MASTER (the "true value"):

1. I calculated the percentage of the duration of the true value trajectory covered by the 95% Highest Posterior Density (HPD) intervals estimated in BEAST, slicing the trajectory in 40 equidistant intervals (*estimate coverage*).

2. I calculated how far the mean posterior N_e value deviated from the true value for the same 40 equidistant intervals (*relative estimation bias*) using the following formula:

$$\sum_{t=0}^{N} \frac{|X_t - \widehat{X}|}{X_t} \qquad (4.3)$$

where X_t is the true value and $\hat{X}$ was the mean posterior N_e value.

3. I used a linear model to test the association between the mean posterior N_e value and the true values (*the scalar coefficient*), in a similar manner to a previously

described method[40,41]. For this purpose I fitted a linear regression model with the posterior mean estimates of N_e as the explanatory variable and the true value as the dependent variable, using the following formula:

$$N = \alpha + \beta N_e + \varepsilon \quad (4.4)$$

where N is the true value, α is the intercept, β is the scalar coefficient, N_e is the effective population size, and ε is the residual error.

All of the above calculations (1-3) were performed on $\log_{10}$ transformed N_e values.

For the R_e estimates obtained from the BDSKY analyses, I calculated the true values of R_e from the simulated population size trajectory and then calculated how often the true values of R_e were within the HPD intervals of R_e estimated from BEAST.

To calculate the true values of R_e I first rephrased the stochastic MASTER equations (4.1) and (4.2) to their analogue deterministic ordinary differential equation (ODE):

$$\frac{dX}{dt} = \lambda X - \gamma X^2 \quad (4.5)$$

which represents a population with carrying capacity $\frac{\lambda}{\gamma}$.

This system is identical to a standard Susceptible-Infected-Susceptible (SIS) epidemiological model with a fixed population size N, which results in the following set of ODEs:

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \quad (4.6)$$

$$\frac{dI}{dt} = \frac{\beta SI}{N} - \mu I \quad (4.7)$$

where S is the number of susceptible individuals, I is the number of infected individuals, β is the transmission rate, and μ is the removal rate. The carrying capacity for this model is $\frac{(\beta-\mu)N}{\beta}$.

We can then relate λ and ϒ to the SIS parameters as:

$$\frac{(\beta-\mu)N}{\beta} = \frac{\lambda}{\gamma} \quad (4.8)$$

Given that the basic reproductive number for this density-dependent SIS system is defined as:

$$R_0 = \frac{\beta}{\mu}\left(\frac{N-1}{N}\right) \quad (4.9)$$

and knowing equation (4.8), I calculate R_0 for any combination of λ and γ used in the MASTER simulations, for any given N.

I can also estimate the effective reproductive number at any time point, knowing the size of the infected population (i.e. the true value from the simulation curve) using the equation:

$$R_e(t) = \frac{\beta}{\mu}\left(\frac{N-I_t}{N}\right) \quad (4.10)$$

where I_t is the number of infected individuals at time t.

For each of the inference approaches, I first calculated the values of the *estimate coverage*, *relative estimation bias*, *scalar coefficient*, and the R_e coverage value for each of the 10 simulation replicates, and then calculated an average value for each of the six demographic scenarios.

4.4 Results

4.4.1 Data

Ukrainian HIV-1 subtype B sequences (N=120) were obtained from patients belonging to four transmission groups: PWID (N=26), MSM (N=32), HET (non-PWID heterosexuals) (N=54), and mother-to-child transmission (MTC, N=8). Most of the Ukrainian sequences (75%, N=90) were from male patients. Regarding their geographic distribution, 73 of the sequences (61%) came from Kyiv, 29 (24%) from Mykolaiv. The remaining regions (N=9) were represented by <=5 sequences each (Table 4.3). I compared the socio-demographic characteristics of the sequences appearing in one of the two clades, versus those that did not, using the Fisher's exact test.

The geographical distribution of the subtype B dataset used here is more representative than the Ukrainian HIV-1 subtype B sequences available from the Los Alamos database (Fig. 4.2, A), but is less extensive than that of the dominant HIV-1 subtype A in Ukraine (Fig. 4.2, B). Figure 4.2, C shows the ML phylogenetic tree estimated from the combined dataset. This phylogeny reveals the two clades that mostly contained Ukrainian sequences (≥75%): Clades 1 and 2, with SH-like support values of 0.99 and 0.86, respectively. Clade 1 contains 65 sequences (53 sequences from Ukraine, 4 from Finland, 3 from Germany, and 1 sequence from USA, UK, Kenya, Belgium, and Japan, each). Clade 2 contains 71 sequences (53 sequences from Ukraine, 4 from Russia, 4 from Georgia, 2 from Azerbaijan, Czech Republic, and Luxembourg each, and 1 from Belarus, Uzbekistan, Slovenia, and Italy each).

Table 4.3. Socio-demographic characteristics of the combined and Ukrainian datasets. Values in bold indicate statistically significant p-values.

Variable		Full dataset			In the Ukrainian clades		Not in the Ukrainian clades		Fisher's exact test
		N	%		N	%	N	%	p-value
Combined dataset									
Sampling region	Ukraine	120	40		106	78	14	9	**<0.001**
	Eastern Europe/ Central Asia	32	11		15	11	17	10	1
	World	147	49		15	11	132	81	**<0.001**
Total		299	100		136	100	163	100	-
Ukrainian dataset									
Gender	Male	90	75		81	76	9	64	0.32
	Female	30	25		25	24	5	36	0.32
Trans-mission group	HET	54	45		44	41	10	71	0.09
	PWID	26	22		24	23	2	14	0.734
	MTC	8	7		8	8	0	0	1
	MSM	32	27		30	28	2	14	0.667
Region	Kyiv	73	61		64	60	9	64	0.78
	Mykolaiv	29	24		27	25	2	14	0.514
	Vinnytsya	5	4		4	4	1	7	0.468
	Odessa	3	3		3	3	0	0	1
	Dnipro	2	2		1	1	1	7	0.221
	Kharkiv	2	2		2	2	0	0	1
	Sumy	2	2		2	2	0	0	1
	Kherson	1	1		0	0	1	7	0.117
	Ivano-Frankivsk	1	1		1	1	0	0	1
	Lugansk	1	1		1	1	0	0	1
	Zhytomyr	1	1		1	1	0	0	1
Data source	Ukrainian HIV drug resistance database	46	38		38	36	8	57	0.12
	Cascade	65	54		60	57	5	36	0.63
	Los Alamos database	9	8		8	7	1	7	1
Total		120	100		106	100	14	100	-

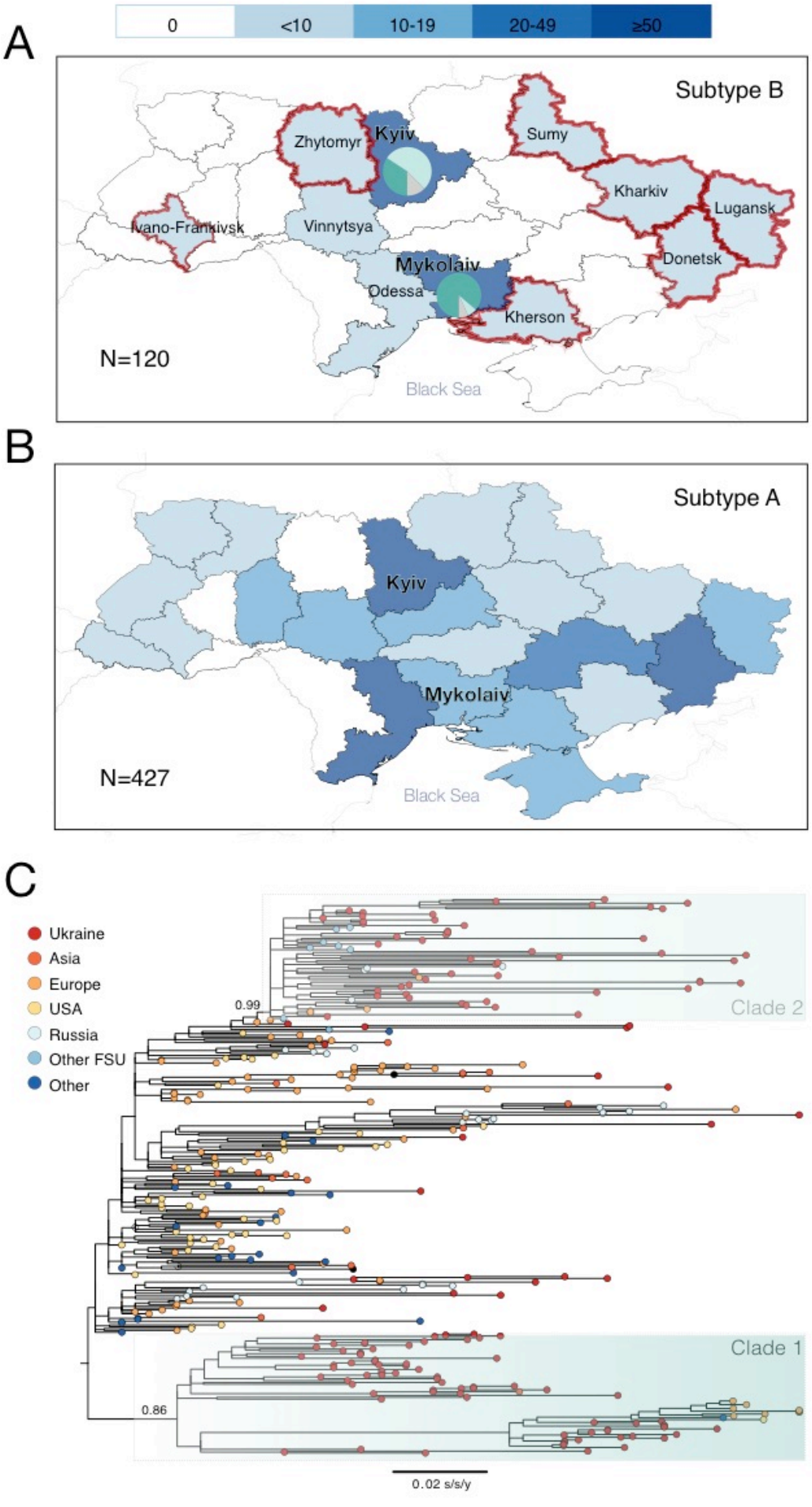

A
0
<10
10-19
20-49
≥50
Subtype B
Zhytomyr
Kyiv
Sumy
Kharkiv
Lugansk
Ivano-Frankivsk
Vinnytsya
Donetsk
Mykolaiv
Odessa
Kherson
N=120
Black Sea
B
Subtype A
Kyiv
Mykolaiv
N=427
Black Sea
C
Ukraine
Asia
Europe
USA
Russia
Other FSU
Other
0.99
Clade 2
Clade 1
0.86
0.02 s/s/y

Figure 4.2 (previous page). A) The number and geographical distribution of HIV-1 subtype B sequences in the combined dataset. B) The number and geographical distribution of HIV-1 subtype A sequences available from the Ukrainian Drug Resistance database in 2012-2015. Administrative units highlighted in red indicate newly sampled regions that were not represented in the Los Alamos HIV sequence database prior to this study. The pie charts for the Kyiv and Mykolaiv regions represent the proportion of sequences from these regions belonging to Clade 1 (light green), Clade 2 (dark green), or neither (grey). C) The maximum likelihood phylogeny of the combined dataset with highlighted Clades 1 and 2. The colours at the tips represent the sampling regions. The bar at the bottom represents genetic distance. The values at the nodes for Clade 1 and 2 represent SH-like statistical support.

A comparison of the socio-demographic characteristics of Clades 1 and 2 is presented in Table 4.4. Clade 1 contains more sequences from Mykolaiv (p-value <0.001). Sequences from Eastern Europe/Central Asia were more prevalent in Clade 2 (p-value <0.001). When Bonferroni correction was taken into account, there was no significant difference between the two clades in terms of transmission group, gender composition, or data source.

The AI analysis showed transmission group admixture for both clades: the AI values for Clades 1 and 2 are 4.32 (p-value 0.32) and 4.45 (p-value 0.60), respectively. The mean pairwise genetic distance in Clade 1 was 2.9% and 1.8% in Clade 2 (t-test p-value <0.001). Since Clade 2 had lower pairwise genetic diversity, and included sequences from other FSU countries, this clade can be identified as the B_{FSU} strain[5,11].

Table 4.4. Socio-demographic characteristics of the two Ukrainian clades. Values in bold indicate statistically significant p-values, accounting for Bonferroni correction for multiple comparisons.

Variable		In the Ukrainian clades			Clade 1		Clade 2		Fisher's exact test
		N	%		N	%	N	%	p-value
All sequences									
Sampling region	Ukraine	106	40		53	82	53	75	0.409
	Eastern Europe/Central Asia	15	11		0	0	15	21	**<0.001**
	World	15	49		12	19	3	4	**0.012**
Total		136	100		65	100	71	100	-
Only Ukrainian sequences									
Gender	Male	81	75		38	72	43	81	0.52
	Female	25	25		15	28	10	19	0.52
Trans-mission group	HET	44	45		21	40	23	43	0.846
	PWID	24	22		13	25	11	21	0.846
	MTC	8	7		7	13	1	2	0.06
	MSM	30	27		12	23	18	34	0.199
Region	Kyiv	64	61		25	47	39	74	0.012
	Mykolaiv	27	24		25	47	2	4	**<0.001**
	Vinnytsya	4	4		-	-	4	8	0.118
	Odessa	3	3		2	4	1	2	1
	Dnipro	1	2		-	-	1	2	1
	Kharkiv	2	2		-	-	2	4	0.495
	Sumy	2	2		-	-	2	4	0.495
	Kherson	0	1		-	-	-	-	1
	Ivano-Frankivsk	1	1		-	-	1	2	1
	Lugansk	1	1		1	2	-	-	1
	Zhytomyr	1	1		-	-	1	2	1
Data source	Ukrainian HIV drug resistance database	38	38		24	45	14	26	0.071
	Cascade	60	54		24	45	36	68	0.035
	Los Alamos database	8	8		5	9	3	6	0.716
Total		106	100		53	100	53	100	-

4.4.2 Molecular clock analysis

A regression of sampling time against root-to-tip genetic divergence showed strong temporal signal in the combined dataset (correlation coefficient=0.61), but weak signal when the two Ukrainian clades were analysed separately (correlation coefficients 0.3 and 0.27 for Clade 1 and Clade 2, respectively) (Fig. 4.3). Therefore I first ran a BEAST analysis with a lognormal relaxed clock on the combined dataset and estimated the TMRCAs of Clade 1 and 2 to be 1971 (95% Highest Posterior Density (HPD) 1956 – 1978) and 1973 (95% HPD 1965 – 1978), respectively. These dates were then used as priors for a TMRCA parameter for all subsequent Bayesian phylogenetic analyses.

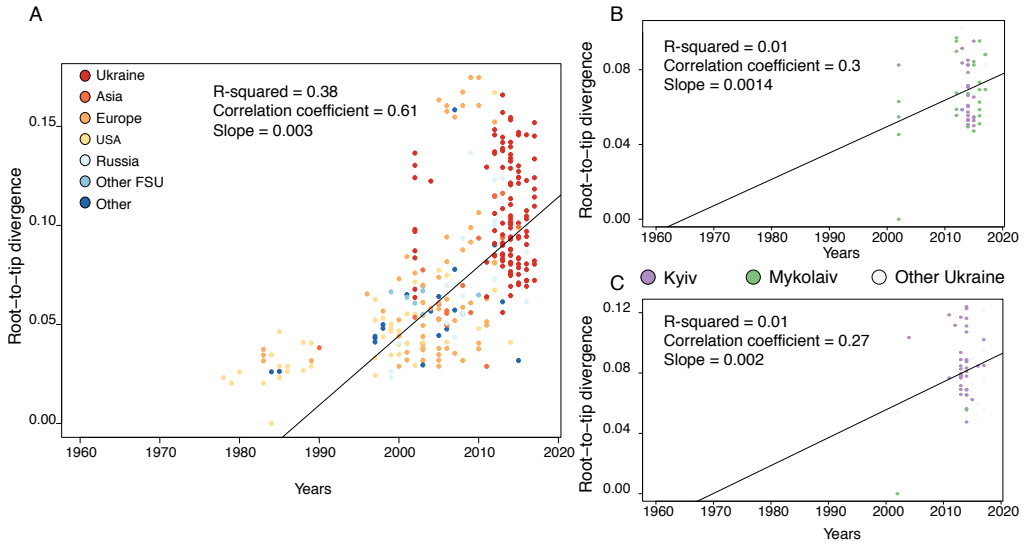
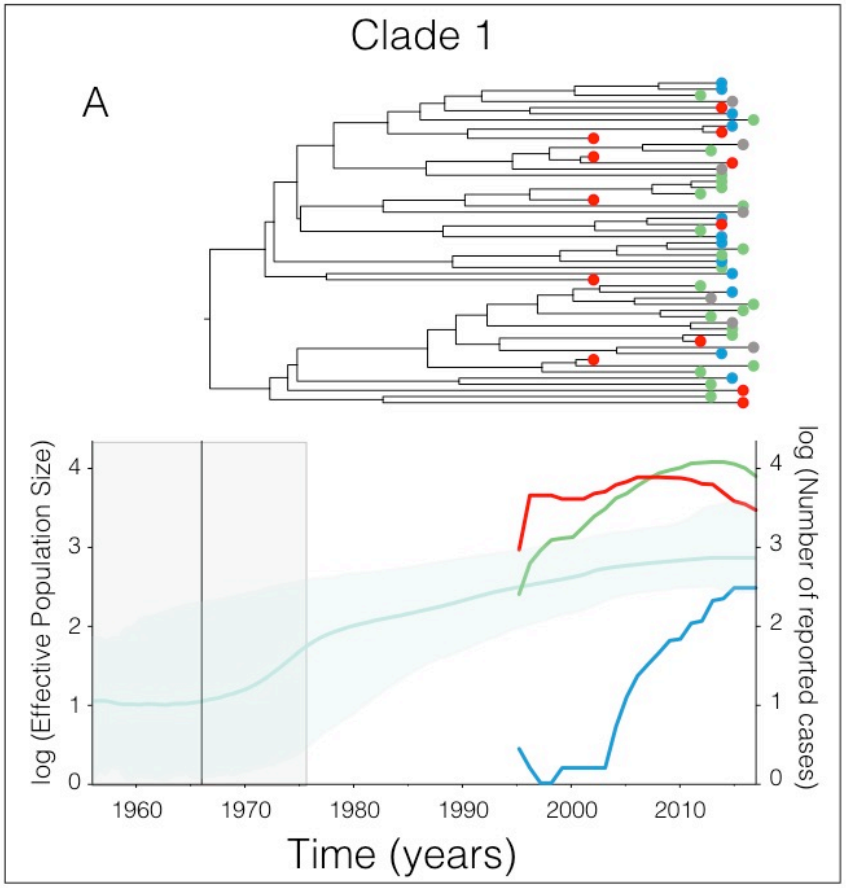


Figure 4.3. Root-to-tip divergence plots estimated using TempEst. a) Combined dataset; b) Clade1; c) Clade 2.

The molecular clock trees of the two clades are presented in Figures 4.4, panel A (top) and 4.4, panel B (top).


Clade 1
A
log (Effective Population Size)
log (Number of reported cases)
0
1
2
3
4
1960
1970
1980
1990
2000
2010
Time (years)


HET
MSM
PWID
MTC

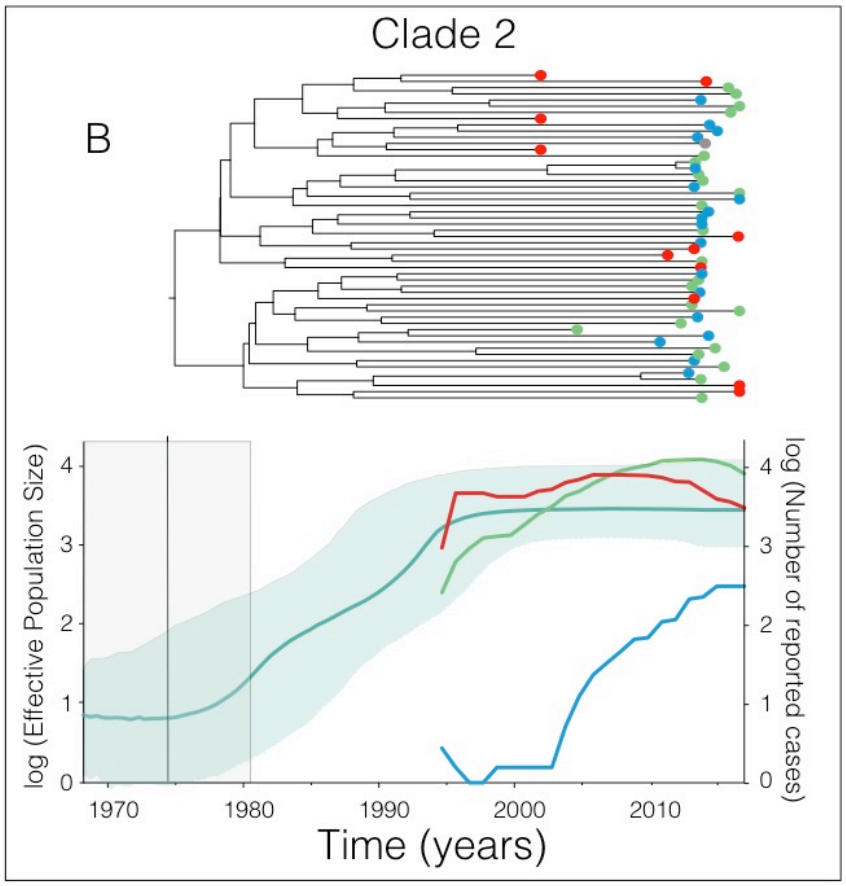

Clade 2
B
log (Effective Population Size)
log (Number of reported cases)
0
1
2
3
4
1970
1980
1990
2000
2010
Time (years)

Figure 4.4 (previous page). A) and B) In the top: Molecular clock trees for Clades 1 and 2, respectively. The colours at the tree tips represent patients' transmission group. In the bottom: Bayesian Skyline plots for Clades 1 and 2, respectively. The shaded green area highlights the 95% Highest Posterior Density (HPD) credible region. The black vertical lines and the shaded grey area represent the time to most recent common ancestor (tMRCA) and its 95% HPD credible. The coloured lines with the right y-axis represent the log Number of reported HIV cases per transmission group.

When I analysed the two clades separately, I obtained TMRCA estimates consistent with the analysis of the combined dataset (1967, 95% HPD 1956 – 1976, and 1975, 95% HPD 1968 – 1981, for Clade 1 and Clade 2, respectively) (Fig. 4.4, panel A (bottom) and panel B (bottom)), as expected given the use of a strong TMRCA prior. For both clades, the Bayesian Skyline plot (BSP) was correlated with the annual number of new infections in HET and PWID, but not MSM (Fig. 4.4, A and 4.4, B). The values of the HET and PWID curves were almost a magnitude higher than the N_e estimated by BSP, and the values of the MSM curve were magnitudes lower than the N_e estimates until the last 10 years. The N_e of Clade 1 grew steadily over two orders of magnitude until 2017. The N_e of Clade 2 grew approximately 2.5 orders of magnitude within the same period; however this clade appears to have experienced a more rapid growth between the 1980s and late 1990s. The overall number of reported HIV cases per year in HET and PWID, but not MSM, agree well with the estimated N_e curves for Clade 2 (Fig. 4.4, B).

I estimated R_e for clades 1 and 2 using the constant rate BDM to be 1.42 (95% HPD 1.26-1.56) and 1.69 (95% HPD 1.49-1.84), respectively, under the assumption that the infectious period is 4 years (Fig. 4.5).

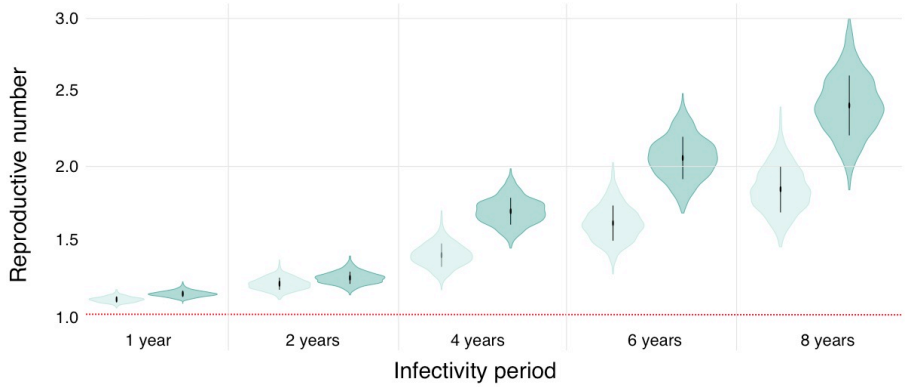


Figure 4.5. The estimates of R_e for each clade, assuming different durations of infectivity. The horizontal red dotted line represents the epidemiological threshold (R_e=1). The vertical lines within the violin plots reflect the second quartile of the R_e distribution.

When I performed additional sensitivity analyses for the R_e parameter, I found that, as expected, estimated values of R_e increase as the infectious period increases. Estimated R_e remained between 1 and 2 if the infectious period is assumed to be ≤4 years. If the infectious period is considered to be >4 years, then R_e gradually increases for both clades, with a particularly high R_e for Clade 2 (R_e=2.42, 95% HPD 2.04-2.8) if I assume an infectious period of 8 years. Importantly, the R_e of Clade 2 was consistently higher than that of Clade 1, regardless of the specific value of the infectious period used.

The results of the BDSKY analysis, applied to all Ukrainian sequences with available sampling dates (N=164), showed that R_e grew in the 1990s and reached a peak value of 3.36 (2.41 – 4.33) around 1998 (Fig. 4.6). It seems to have declined to a value of 0.13 (0.02-0.27) in 2001, and stayed below 1 until 2008, after which again grew to 2.74 (2.12 – 3.32). R_e dropped again after 2012 to around 1.5 (0.86 – 2.05) and remained stable in recent years.

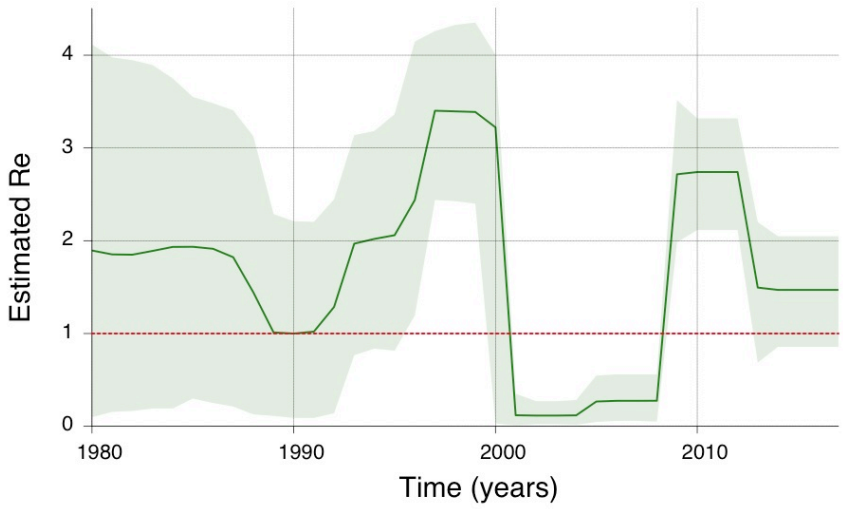


Figure 4.6. R_e estimates obtained using the BDSKY model applied to all Ukrainian sequences, over 10 equidistant time-intervals. The shaded area represents 95% Highest Posterior Density (HPD) of the R_e estimates. The horizontal red dotted line represents the epidemiological threshold (R_e=1).

Surprisingly, the results of the BDSKY analysis of the full Ukrainian dataset and the results of the Bayesian Skyline analysis of the two Ukrainian clades do not fully agree. While the BDSKY plot shows a large reduction in R_e after 2000, the BSPs of both clades show no decline in N_e throughout the epidemic (Figure 4.4).

4.4.3 Simulation study

4.4.3.1 Performance of the coalescent models

The simulated number of infections (the true value) and the N_e estimates and their 95% HPD intervals can be found in Figures 4.7-4.12 for scenarios 1-6, respectively. Only 5 of the 10 replicates of each scenario are shown.

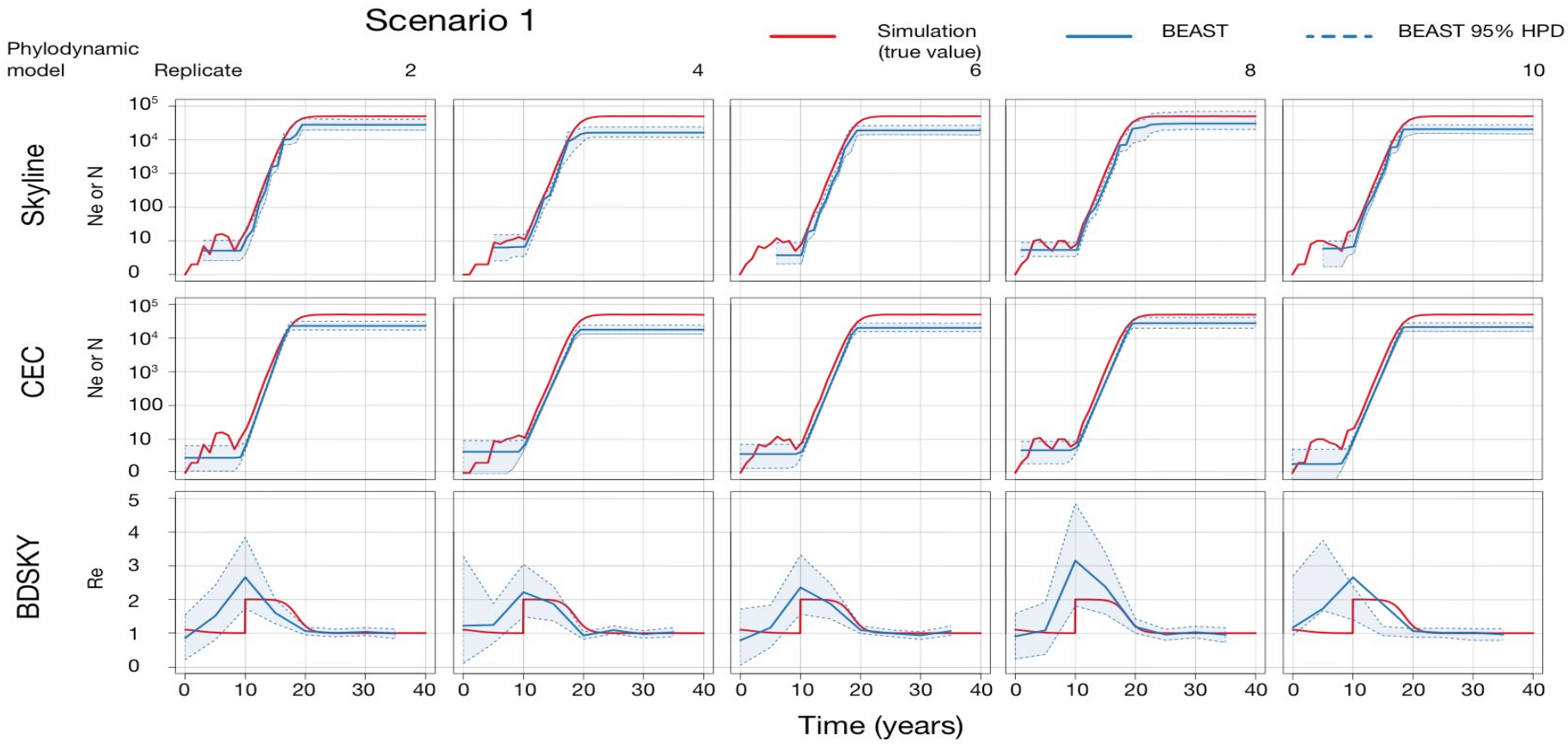


Figure 4.7. The simulated number of infections and R_e calculated from the simulated trajectories (the true value, in red) and the estimated values of N_e and R_e (in blue) for scenario 1 (y-axis). The inference approach is indicated in the left (Bayesian Skyline, CEC, and BDSKY). Every second replicate is shown.

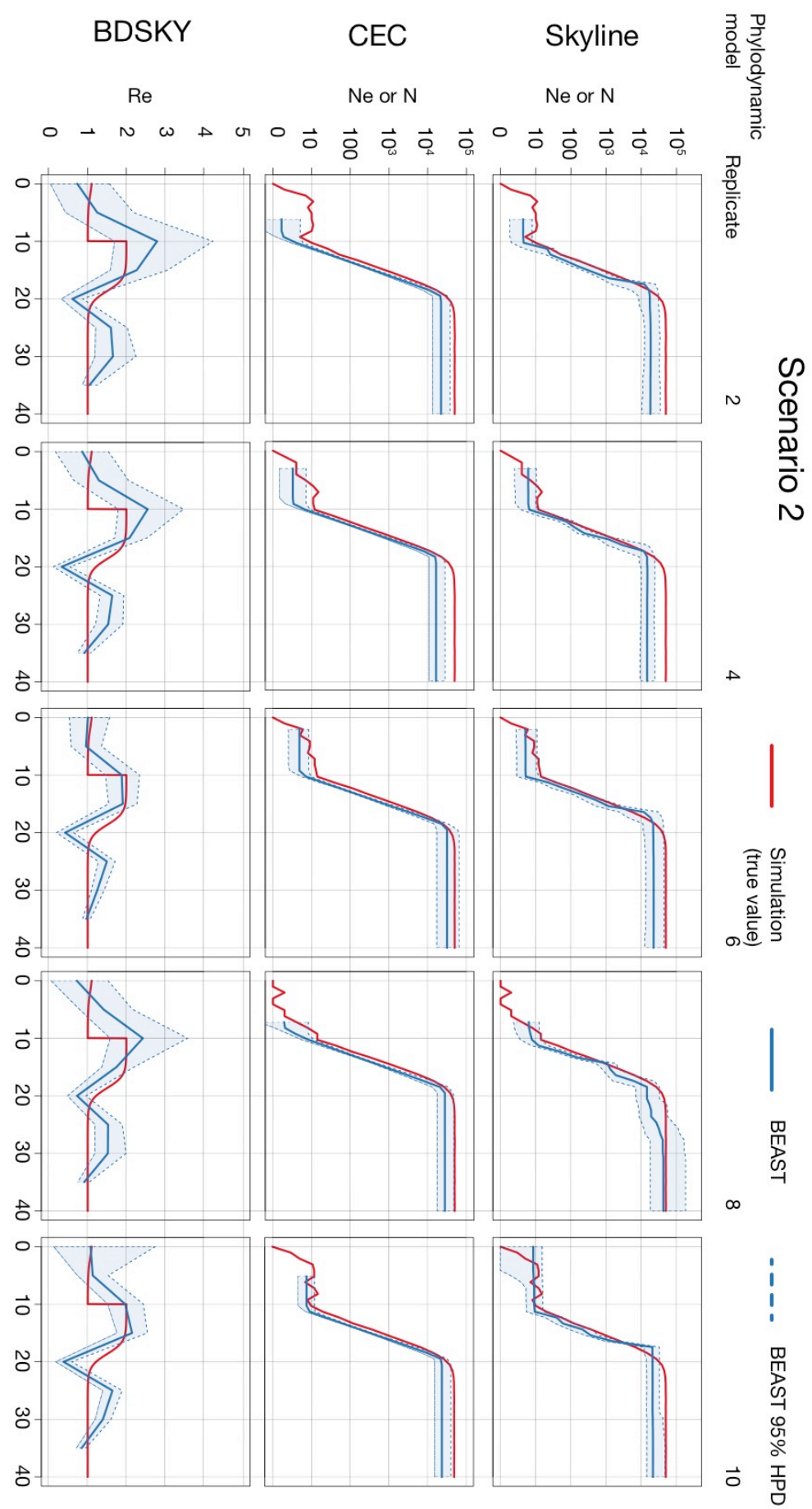


Figure 4.8. Details as per the legend of Figure 4.7, but for scenario 2.

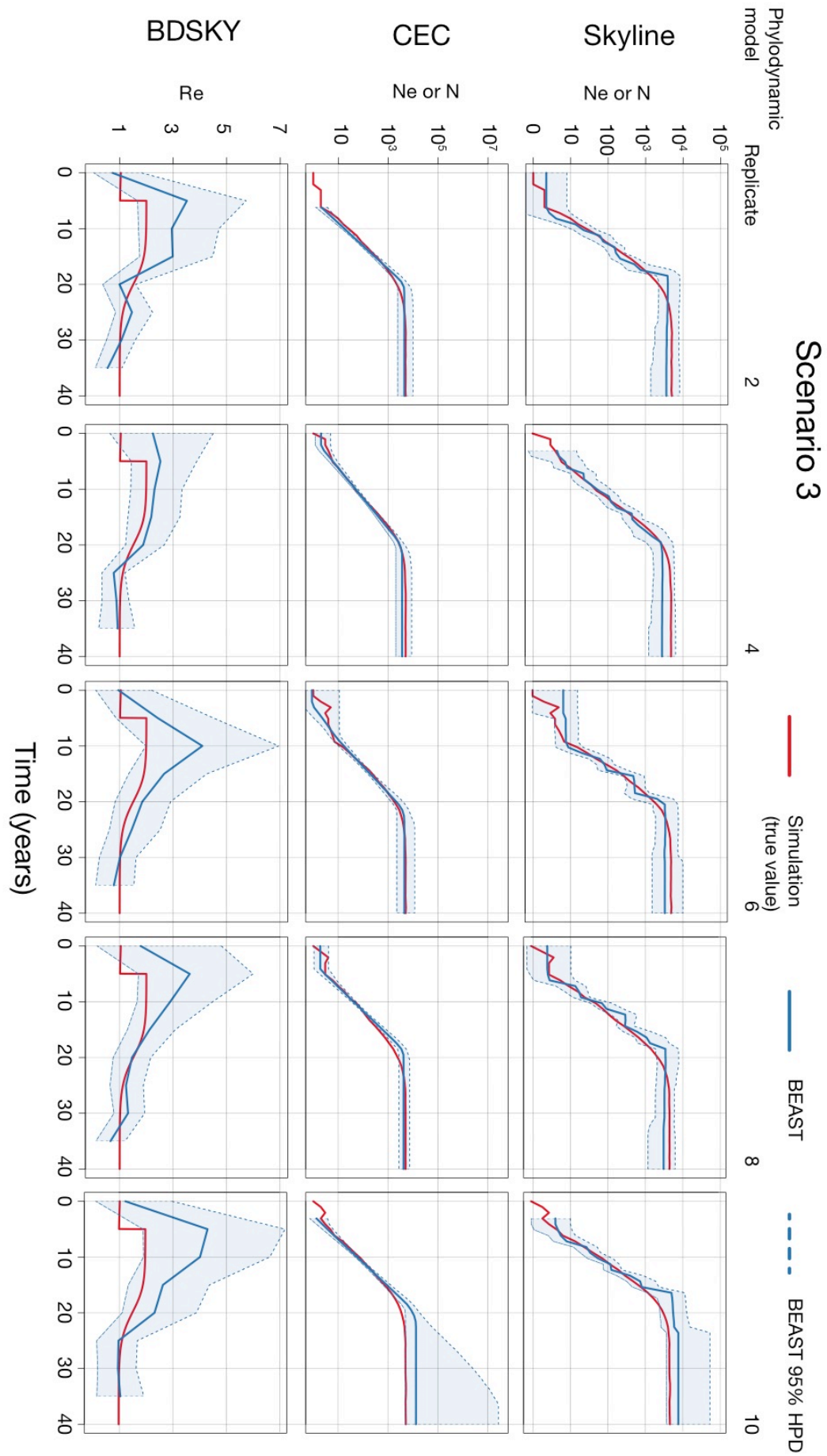


Figure 4.9. Details as per the legend of Figure 4.7, but for scenario 3.

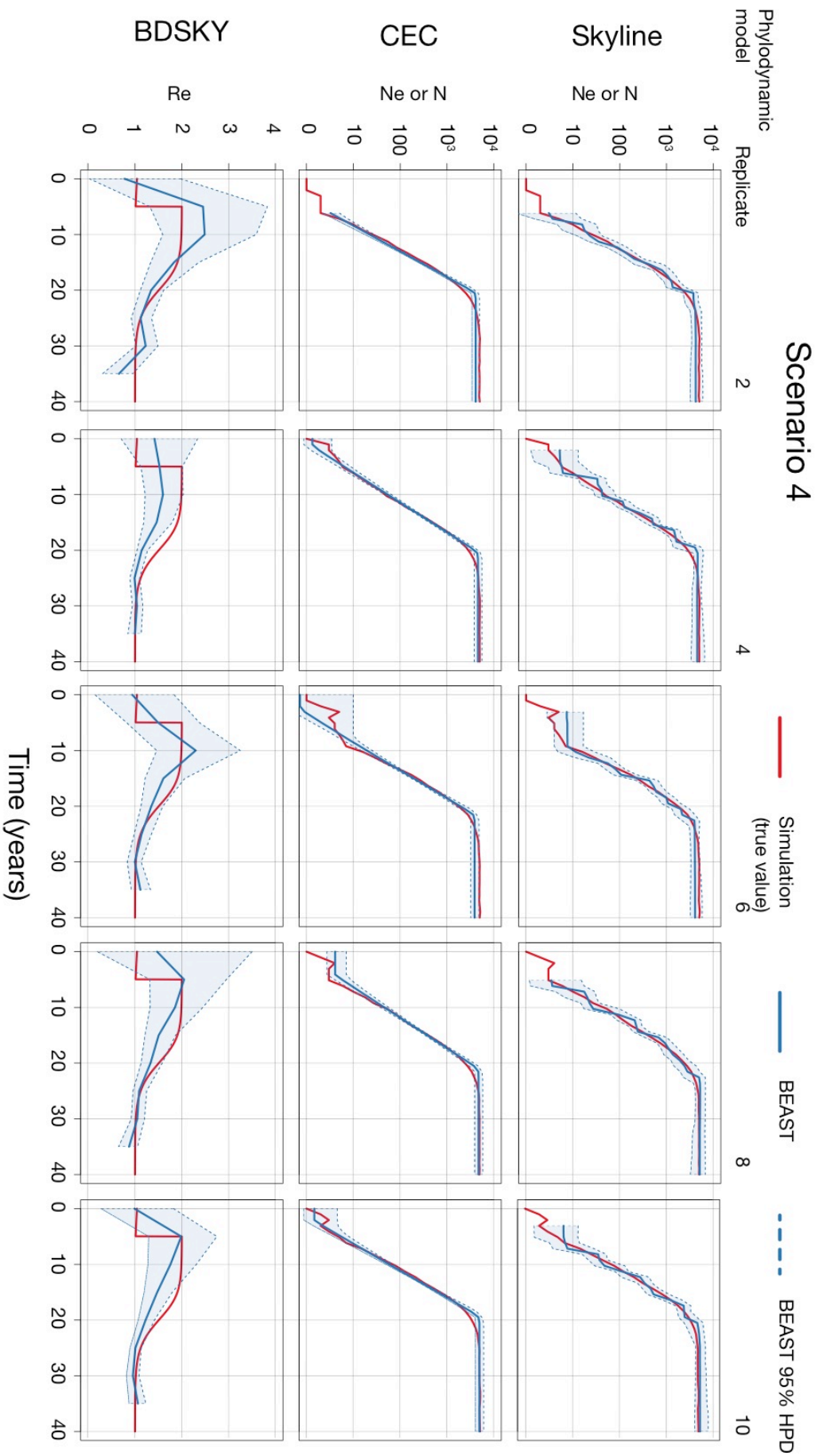


Figure 4.10. Details as per the legend of Figure 4.7, but for scenario 4.

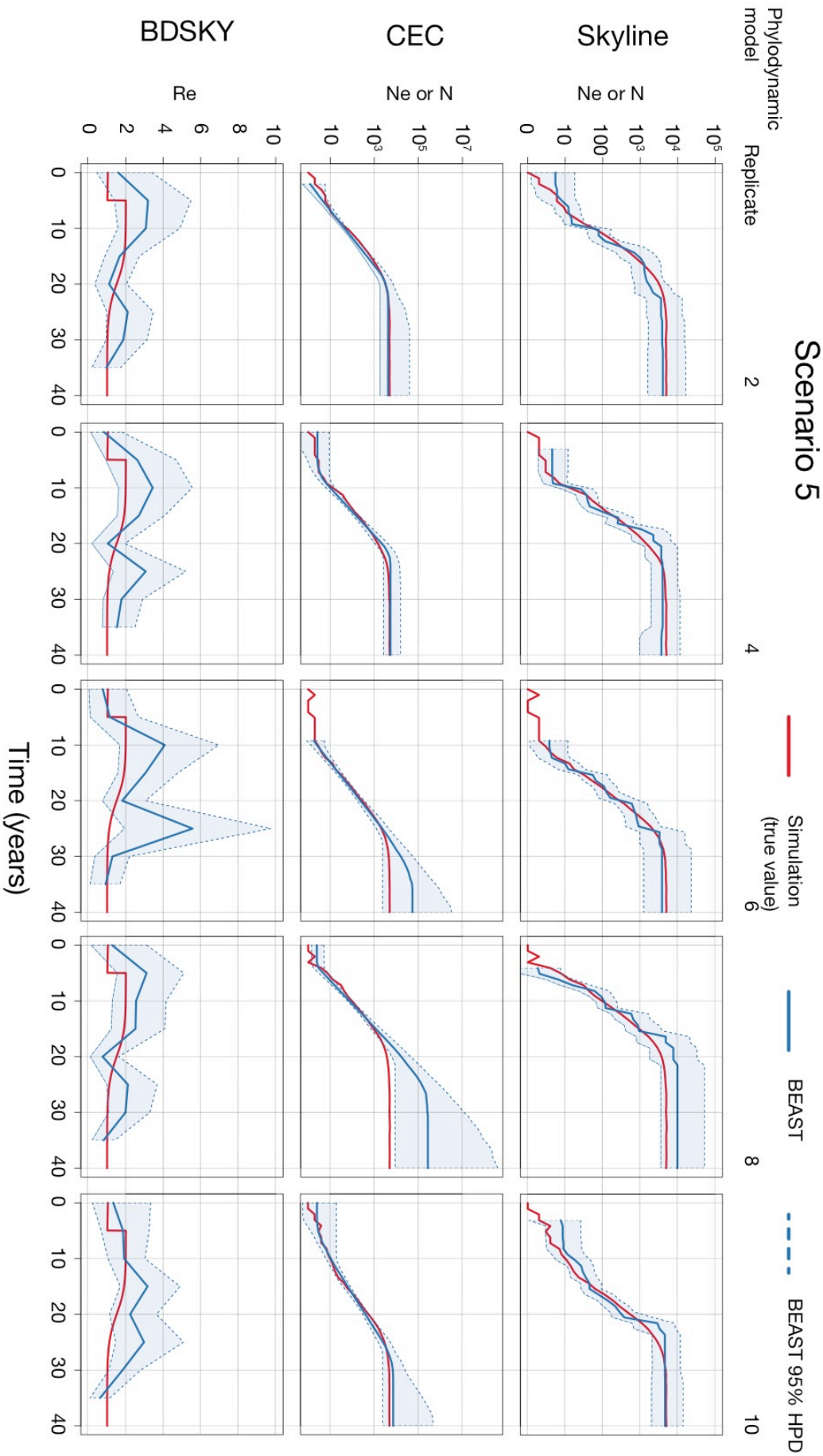


Figure 4.11. Details as per the legend of Figure 4.7, but for scenario 5.

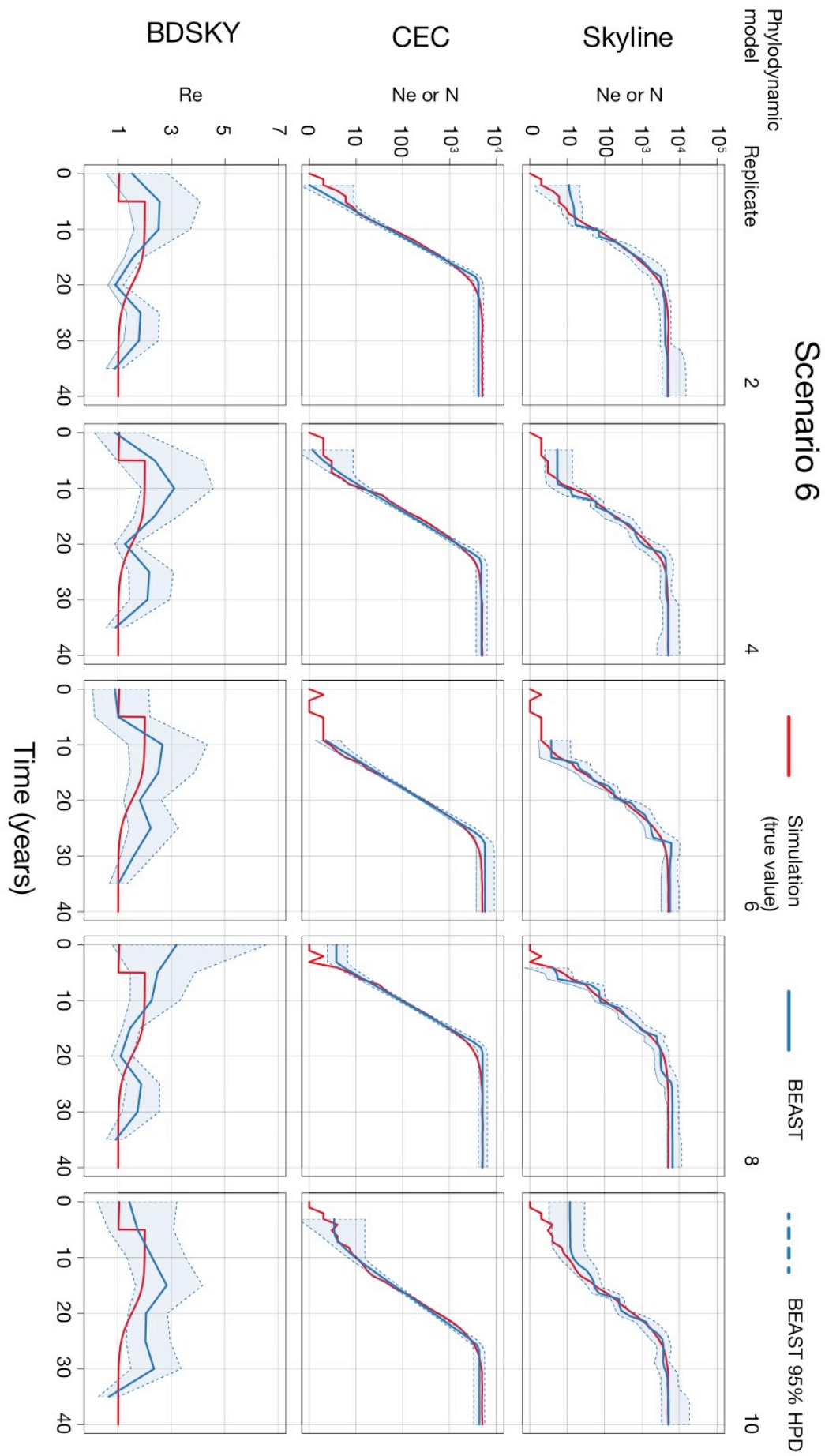


Figure 4.12. Details as per the legend of Figure 4.7, but for scenario 6.

First, I assessed the accuracy of the Bayesian Skyline and CEC coalescent models in estimating the trajectory of N_e through time by calculating *the estimation coverage* of the true number of infections. *The estimation coverage* equals the percentage of time that the 95% HPD intervals of the N_e estimate obtained from BEAST include the true population size value (Table 4.5).

Table 4.5. The estimation coverage of the true simulated number of infected individuals by the 95% HPD intervals estimated in BEAST

Scenario	Estimate coverage, %	
	Bayesian Skyline	CEC
1	26.3	12.6
2	48.9	42.8
3	91.9	82.0
4	92.3	64.6
5	97.6	89.5
6	94.7	85.8
Average	75.3	62.9

The true population size was retained in the 95% HPDs of the non-parametric Bayesian Skyline model for about 75% of the duration of epidemic trajectories. The HPD intervals included the true value for >50% of the duration of scenarios 3-6 (these scenarios represent slower growth and smaller epidemics). Scenarios number 5 and 6, which represent simulations with different sampling proportions (0.25% and 1%, respectively), but the same sampling scheme (varying sampling through time), performed best in recovering the true values (Table 4.5). The HPD intervals obtained using the CEC model on average contained the true value less often, resulting in average *estimation coverage* of 63%. As with the Skyline model, the HPD intervals estimated by the CEC model for scenarios 5 and 6 included the true values for a longer duration than other scenarios (*estimation coverage* = 89.5% and = 85.8% for scenarios 5 and 6, respectively). The true population size in Scenario 1 (this scenario

represents rapid growth and large epidemics) was less likely to be included in the HPD intervals estimated by both the Bayesian Skyline and CEC models (*estimation coverage* = 26.3% and = 12.6% for the Skyline and the CEC models, respectively).

I next calculated how far the mean N_e estimate deviated from the true value (*relative estimation bias*) (Table 4.6). The magnitudes of the biases of all scenarios are comparable to each other, for both models. Nonetheless, the smallest *relative bias* was observed for scenario 4, which represented a slower growing epidemic of smaller size, with 1% constant sampling through time, both for the Skyline and CEC models (3.20 ± 1.98 and 2.14 ± 0.89). The largest *relative bias* was observed for the scenario 6 for the Skyline analysis (5.28 ± 3.74) and for the scenario 1 for CEC (5.21 ± 0.95).

Table 4.6. The relative estimation bias.

Scenario	Bayesian Skyline	CEC
	Mean	Mean
1	4.42	5.21
2	3.91	4.44
3	3.62	3.14
4	3.20	2.14
5	4.72	4.25
6	5.28	2.72

I further used a linear model to fit the posterior mean estimates of N_e to the simulated number of infected individuals. The values of the gradient of the linear regression (*scalar coefficient,* β) and the coefficient of determination (R^2), which is a statistical measure of how close the data points are to the fitted regression line, are presented in Table 4.7. I report the values of R^2 to compare the fit of the linear model between different scenarios and inference approaches in this study, though these values cannot be compared to other studies, since this measure depends on the number of time

points that I chose here. The N_e values estimated by both models were very close to the true values; the β estimates were on average very close to 1 for all simulations.

Table 4.7. The scalar coefficient β and the coefficient of determination R^2 of the linear regression between the estimated N_e and the true values.

Scenario	Bayesian Skyline		CEC	
	β	Adjusted R^2	β	Adjusted R^2
1	1.04	0.994	1.00	0.996
2	1.02	0.991	0.97	0.996
3	1.01	0.983	0.93	0.989
4	1.04	0.989	1.00	0.993
5	1.02	0.982	0.88	0.979
6	1.05	0.986	0.99	0.995

4.4.3.2 Performance of the BDSKY model

To assess the performance of the BDSKY model in estimating the epidemic reproductive number parameter (R_e), I first calculated the true values of the R_e from the simulated trajectories of the number of infected individuals using equation (4.10) (Table 4.8).

Table 4.8. True values of the reproductive number (Re) estimated from the simulated infection trajectories

Year in simulation	R_e by scenario					
	1	2	3	4	5	6
0	1.10	1.10	1.04	1.04	1.04	1.04
5	1.02	1.02	2.00	2.00	2.00	2.00
10	2.00	2.00	1.99	1.99	1.99	1.99
15	1.97	1.97	1.91	1.91	1.91	1.91
20	1.21	1.22	1.51	1.51	1.51	1.51
25	1.00	1.00	1.12	1.12	1.12	1.12
30	1.00	1.00	1.02	1.02	1.02	1.02
35	1.00	1.00	1.00	1.00	1.00	1.00

I then checked if the true values of R_e from simulations were within the 95% HPD intervals of R_e estimated from BEAST2 at each of eight specified time points (Table 4.9).

Table 4.9. The proportion of time points, at which the true value of R_e was recovered by the 95% HPD intervals estimated in BEAST2, and the direction of the bias, if the true R_e was not within the HPD intervals

Scenario	R_e coverage, %	The direction of the bias
1	87%	Underestimated
2	57%	Overestimated
3	97%	No overall bias
4	80%	Underestimated
5	85%	Overestimated
6	66%	Overestimated

The true R_e was recovered most frequently by the BDSKY method for scenario 3, for which the true value of R_e was within the confidence intervals at 97% of time points. BDSKY estimates obtained for scenario 3 did not include the true R_e as often, for the true value was recovered 57% of the time.

4.5 Discussion

4.5.1. Transmission dynamics, and geographical and transmission group composition of HIV-1 subtype B clades circulating in Ukraine

HIV-1 subtype A has dominated the epidemic in Ukraine and other FSU countries since the mid-1990s[42]. The subtype A epidemic in Ukraine is well described, including its explosive growth among PWID in the 1990s and the recent spread after the initiation of war in the eastern regions of Ukraine (see Chapters 2 and 3, and [5]). In contrast, HIV-1 subtype B in Ukraine has been more neglected, primarily due to its

lower prevalence. Here I characterise the current molecular epidemiology of HIV-1 subtype B in Ukraine, including its transmission group composition and virus transmission dynamics. I also conducted a simulation study to evaluate the appropriateness and accuracy of phylodynamic models to answer epidemiological questions. The simulation scenarios were chosen so that they correspond approximately to the Ukrainian HIV-1 subtype A and B epidemics.

Previous reports have suggested that the HIV-1 subtype B epidemic in Ukraine started with the introduction of this subtype into the drug injecting community of Mykolaiv in 1990s[5,43]. In 2001-2002, HIV-1 subtype B was mostly prevalent in Mykolaiv (N=22), followed by Kyiv (N=18), and few cases reported in Odessa[5]. Two HIV-1 subtype B lineages were found circulating in Kyiv[5], which is consistent with my findings here. However, the geographic distribution of the HIV-1 subtype B sequences here is quite different: the majority of HIV-1 subtype B sequences come from Kyiv (61% of the Ukrainian dataset) or other regions (15%). This could be explained by a potential selection bias in the CASCADE data[14], sampled exclusively in Kyiv. Among the available 46 HIV-1 subtype B sequences in the Ukrainian HIV Drug Resistance database, 24 (52%) originate from Mykolaiv followed by small numbers of sequences from 9 other regions across Ukraine.

Overall this analysis indicates a lack of transmission group structure for the two clades of HIV-1 subtype B circulating in Ukraine, suggesting frequent viral mixing among transmission groups. Importantly, by analysing the largest dataset of subtype B from Ukraine to date, this chapter shows that Ukrainian HIV-1 subtype B, and B_{FSU} (Clade 2) in particular, is no longer concentrated in selected regions (i.e., Mykolaiv and Kyiv) or in specific transmission groups. Similar observations of the expansion of B_{FSU} from PWID to other transmission groups have been made in Russia, where HIV-

1 subtype B sequences isolated from MSM patients clustered with B_{FSU} sequences collected from PWID[44].

The results of the BSP analysis, which showed a continuous growth of the subtype B epidemic based on the analysis of the two Ukrainian clades, did not fully agree with the BDSKY results that showed a much higher than 1 epidemic reproductive number between 2000 and 2008. Given that the Bayesian Skyline has little power to estimate a decline in N_e if no coalescent events are present for a period of time[30], additional information might help to resolve which model describes the true epidemic dynamic. Here, I show that the estimated effective population size of both Ukrainian clades is congruent with the annual number of new HIV cases in both PWID and HET in Ukraine[3] (Fig. 4.4).

I also estimated the average reproductive number R_e for both clades. I found that Clade 2 (B_{FSU}) has a consistently higher R_e than Clade 1, regardless of the infectious period value used. The R_e estimates for both clades (1.42 and 1.69, assuming an infectious period = 4 years) were above one, indicating epidemic growth. The results of the BDSKY analysis suggested that the R_e of the Ukrainian subtype B epidemic was significantly below one for a brief period between 2000 and 2008. The wide HPD intervals of the R_e estimate before 1995 make interpretations of this parameter uncertain prior to that time. While the HPD intervals were narrower after 1995, the two peaks in R_e directly precede the time of the sampling events in the full Ukrainian dataset. Notably, I assumed that the sampling proportion remained constant after the time of the oldest sample (2001 in this analysis). This is a common assumption for the BDSKY model[45], but one which does not directly reflect the sampling of my sequence data. Thus, the results of the BDSKY analysis need to be interpreted with caution, as they may represent a sampling artefact, as also suggested by the findings of my

simulation study. In particular, the simulations showed that if a constant sampling through time was assumed by the BDSKY model, yet other sampling schemes were used to actually sample the infections shown in the tree, then estimates of R_e might be biased, particularly towards over-estimating the true value of R_e later in the epidemic (i.e. closer to the present). In most scenarios R_e was more precisely estimated during the first 20 years of the simulation, while estimates were less accurate for the last 20 years, when R_e had decreased from its peak and was around 1, particularly under scenarios 2, 5, and 6, all of which represented the "varying through time" sampling scheme (Figures 4.8, 4.11, and 4.12). Interestingly, if the true value of R_e was not covered by the HPD intervals, then it was likely to be underestimated by the BDSKY method for scenarios 1 and 4 (these scenarios represent constant sampling through time), and overestimated for scenarios 2, 5, and 6 (these scenarios represent varying sampling through time), while scenario 3 showed unbiased estimates. This exercise shows that the choice of an appropriate sampling model is crucial for obtaining accurate estimates under the BDSKY model.

The estimates of R_e for the whole subtype B epidemic, and for clades 1 and 2 separately, were much lower than R_e previously estimated for the HIV-1 subtype A epidemic in Ukraine in the 1990s (R_e=10 per drug injector, although R_e for subtype A is likely to be lower now, when the epidemic stopped growing) (see Chapter 2). One of the explanations for the difference in epidemic potential for the two HIV-1 subtypes is "first-comer advantage", which hypothesizes that the first successfully established "founder" strain can significantly slow down the spread of other strains in the region, by either inhibiting superinfection due to the stimulation of anti-HIV immune responses and the exhaustion of the number of target cells, or by reducing the number of susceptible non-infected individuals[46]. For example, analyses of HIV-1

subtypes A and D in Uganda suggest that subtype A would need to have 25% transmission advantage compared to the "first comer", subtype D, in order to establish an epidemic[46]. The TMRCA of Ukrainian HIV-1 subtype A was previously estimated to be around 1970 (95% HPD 1962-1976)[7], which is similar to the TMRCAs estimated here for HIV-1 subtype B Clade 2 (1973, 95%HPD 1965-1978) and Clade 1 (1971, 95%HPD 1956-1978). Thus, my data suggest the approximately simultaneous introductions of HIV-1 subtypes A and B into Ukraine. However, the smaller number of HIV-1 subtype B sequences in my analysis here, compared to the number of HIV-1 subtype A sequences available in Chapter 2, is associated with comparatively large confidence intervals for the estimates in this chapter. Further studies with more sequence data should provide a better understanding of the possible effects of first-comer advantage, if any, of HIV-1 subtypes in Ukraine.

One of the limitations of this study is that the data came from three different sources with different sampling techniques and different data collection methods for self-reported transmission group information. In particular, CASCADE data were collected anonymously, while data at the AIDS-Centers were collected by doctors in a non-anonymous manner. In my sample 75% of patients were male; interestingly, 67% of HET patients were also male. MSM are often stigmatised and as a consequence this transmission group tend to be underreported in many settings, including Western Europe[47-49]. I cannot exclude the possibility that some patients incorrectly self-reported as belonging to the non-MSM group in Ukraine. Similarly, PWID might not have disclosed their drug injecting practices to AIDS-Centre clinicians, particularly due to stigma against HIV and drug use in medical facilities in Ukraine[50,51]. As a result, underreporting of MSM and PWID status could overestimate the true proportion of HET in this study and therefore make the Ukrainian HIV-1 subtype B

epidemiological profile look different to that of predominantly MSM-driven subtype B epidemics in Russia[52] and some European countries[53]. Furthermore, my sample is likely to represent a very small proportion of existing HIV-1 subtype B infections in Ukraine (<1%). This is partly due to high rates of undiagnosed HIV infections in Ukraine, particularly in some transmission groups (such as MSM)[54]. As very few patients have access to drug resistance testing (see Chapter 3 for more details), viral genetic data from routine care are also very sparse.

I found that the epidemiological profile of HIV-1 subtype B epidemic in Ukraine has recently changed. Specifically, HIV-1 subtype B is no longer limited to the PWID sub-population and is spreading to multiple transmission groups throughout the country. As with HIV-1 subtype A in Ukraine, the subtype B epidemic appears to have generalised to HET, a much larger susceptible population, which could fuel the spread of HIV in Ukraine. The analyses here also suggest that HIV-1 subtype B in Ukraine is not concentrated in MSM, unlike the situation in Western Europe. Though the Ukrainian HIV-1 subtype B epidemic appears to be growing comparatively slowly, and is characterised by a lower R_e, its increasing range across locations and transmission groups highlights the need for better surveillance, linkage to care, anti-retroviral treatment and monitoring across the country.

4.5.2. Discussion of the simulation study results

I performed a simulation study under several different study design scenarios, revealing some important insights into the choice and interpretation of phylodynamic models applied in HIV epidemiology. Overall, the estimated N_e was very close to the simulated number of infections ("the true value"), which was reflected in the linear regression gradient being close to 1 for all scenarios analysed, using both coalescent models. I found that scenarios with the "varying through time" sampling scheme were

more likely to include the true value in the HPD intervals for N_e (i.e. *estimation coverage* of the true value was high for scenarios 5 and 6 under both coalescent models). This is partly expected by the assumptions of the coalescent approach, which requires a relatively small sample to have been sampled from a large population[55]. This result can also be explained by the fact that the analyses of scenarios 5 and 6 generated smaller phylogenies and therefore resulted in more uncertain estimates with broader HPD intervals, which were thus more likely to include the true value. This was also demonstrated in the analysis of *relative estimation bias*, which was lower for scenarios 3 and 4 (which represent a slower growing, smaller epidemics, with constant sampling through time). As such, the "constant through time" sampling scheme resulted in more confident estimates with narrower HPD intervals, and mean posterior N_e values closer to the true values, but the true values were also more likely to fall outside the HPD intervals. Notably, both scenarios that represented *rapid growth* epidemics (scenarios 1 and 2) established both a lower *coverage* and a higher *bias*, which arose from an underestimation of the true value of N_e.

Overall, I found that the coalescent models were able to recover accurate estimates of the changes in N_e, although the true value is often biased and not covered by the HPD intervals. The bias appears to stem from the coalescent models being overconfident in their estimates, especially on large trees, and not from a model misspecification. This bias likely results from applying coalescent models that assume a deterministically changing population size to simulations where the population size fluctuates stochastically[36,56,57]. In addition, I find that the accuracy and bias are more dependent on the size of the trees being simulated than the overall sampling proportion (although all of my scenarios had sampling proportions of < 1%, well within the range of acceptable range for coalescent approximation).

The sampling proportion appears to have only a small effect (compared to the size of the epidemic, the epidemic growth rate, and the sampling scheme) on the performance of the coalescent models; this finding was similar to all of the parameters that were used to compare the true value and N_e (the *coverage*, the *bias*, and the *linear regression gradient*). Crucially, BDSKY estimates of R_e lost accuracy when the sampling assumption of the model (constant sampling through time) was violated. This resulted in skylines with two peaks. The period of a low R_e between the peaks (Figures 4.8, 4.11, and 4.12) results from the model constraining the sampling rate to be constant during the period when no samples are being taken; the only way the model can compensate for the lack of samples is by inferring a very low R_e. When new samples are taken after t=30, the model overcompensates and overestimates R_e in the later stages of the simulation, resulting in the mean bias being positive, as opposed to negative, as in the other inference approaches.

The results of the simulation study allows me to assign a greater confidence to the Bayesian skyline plot estimates of N_e for the two clades of HIV-1 subtype B in Ukraine, showing a phase of exponential growth followed by a decrease in transmission. Furthermore, the simulations support my previously mentioned notion that the BDSKY results obtained for the Ukrainian subtype B epidemic should be interpreted cautiously: while the decline in R_e between 2000 and 2008 might be significant, the particularly high values of R_e immediately before 2000, and after 2008, may be an overestimation.

4.6 Ethical approval and acknowledgements

HIV drug resistance testing was performed at the L. V. Gromashevskij Institute of Epidemiology and Infectious Diseases of National Academy of Sciences of Ukraine following all Ukrainian legal and ethical requirements. Fully anonymised genetic sequence data were transferred to Oxford through a Material Transfer Agreement for molecular evolutionary analyses. Ethical approval for CASCADE was given by the ethics committee of the L. V. Gromashevskij Institute of Epidemiology and Infectious Diseases of National Academy of Sciences of Ukraine and the ethics committee of University College London (UCL).

CASCADE was supported by the European Union Seventh Framework Programme (FP7/2007–2013) under EuroCoord (grant agreement number 260694).

4.8 References

1. ECDC. HIV/AIDS surveillance in Europe 2015. Stockholm: European Centre for Disease Prevention and Control/WHO Regional Office for Europe; 2016.
2. Dehne KL, Grund JPC, Khodakevich L, Kobyshcha Y. The HIV/AIDS epidemic among drug injectors in Eastern Europe: Patterns, trends and determinants. J Drug Issues 1999;29:729-76.
3. Ministry of Health of Ukraine. Operational information on officially ergistered cases of HIV, AIDS, and the numbers of AIDS deathes, as of March 2017. Kyiv: Ukrainian Centres to Prevent and Fight AIDS 2017.
4. Vitek CR, Cakalo JI, Kruglov YV, et al. Slowing of the HIV Epidemic in Ukraine: Evidence from Case Reporting and Key Population Surveys, 2005-2012. Plos One 2014;9.
5. Saad MD, Shcherbinskaya AM, Nadai Y, et al. Molecular epidemiology of HIV Type 1 in Ukraine: birthplace of an epidemic. AIDS Res Hum Retroviruses 2006;22:709-14.
6. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.
7. Diez-Fuertes F, Cabello M, Thomson MM. Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. Infect Genet Evol 2015;33:197-205.
8. Novitsky VA, Montano MA, Essex M. Molecular epidemiology of an HIV-1 subtype A subcluster among injection drug users in the Southern Ukraine. AIDS Res Hum Retroviruses 1998;14:1079-85.
9. Abecasis AB, Wensing AM, Paraskevis D, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 2013;10:7.
10. Bobkov A KE, Selimova L, Bobkova M, et al. A Sudden Epidemic of HIV Type 1 among Injecting Drug Users in the Former Soviet Union: Identification of Subtype A, Subtype B, and Novel gagA/envB Recombinants. AIDS Res Hum Retroviruses 1998;14(8).
11. Bobkova M. Current status of HIV-1 diversity and drug resistance monitoring in the former USSR. AIDS Rev 2013;15:204-12.
12. Saad MD, Aliev Q, Botros BA, et al. Genetic forms of HIV Type 1 in the former Soviet Union dominate the epidemic in Azerbaijan. AIDS Res Hum Retroviruses 2006;22:796-800.
13. Kazennova E, Laga V, Lapovok I, et al. HIV-1 genetic variants in the Russian Far East. AIDS Res Hum Retroviruses 2014;30:742-52.
14. Fearnhill E, Gourlay A, Malyuta R, et al. A Phylogenetic Analysis of Human Immunodeficiency Virus Type 1 Sequences in Kiev: Findings Among Key Populations. Clin Infect Dis 2017;65:1127-35.
15. Simmons R, Malyuta R, Chentsova N, et al. HIV Testing and Diagnosis Rates in Kiev, Ukraine: April 2013-March 2014. Plos One 2015;10.
16. Pineda-Pena AC, Faria NR, Imbrechts S, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol 2013;19:337-48.
17. Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res 2014;42:e144.

18. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389-402.
19. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.
20. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 2016;33:1870-4.
21. Bennett DE, Camacho RJ, Otelea D, et al. Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. Plos One 2009;4.
22. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312-3.
23. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 1999;16:1114-6.
24. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 2010;59:307-21.
25. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.
26. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infect Genet Evol 2008;8:239-46.
27. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012;29:1969-73.
28. Gelman A, Meng XL. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. Stat Sci 1998;13:163-85.
29. Xie WG, Lewis PO, Fan Y, Kuo L, Chen MH. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. Syst Biol 2011;60:150-60.
30. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 2005;22:1185-92.
31. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology 2006;4:e88.
32. Rambaut A DA, Xie D, Baele G & Suchard MA. Tracer v1.7. 2018.
33. Bouckaert R, Heled J, Kuhnert D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Plos Comput Biol 2014;10.
34. Stadler T, Kouyos R, von Wyl V, et al. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol 2012;29:347-57.
35. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 2013;110:228-33.
36. Volz EM, Frost SDW. Sampling through time and phylodynamic inference with coalescent and birth-death models. J R Soc Interface 2014;11.
37. Pybus OG, Rambaut A. GENIE: estimating demographic history from molecular phylogenies. Bioinformatics 2002;18:1404-5.

38. Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A. The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol 2003;20:381-7.
39. Vaughan TG, Drummond AJ. A stochastic simulator of birth-death master equations with application to phylodynamics. Mol Biol Evol 2013;30:1480-93.
40. Bennett SN, Drummond AJ, Kapan DD, et al. Epidemic dynamics revealed in dengue evolution. Mol Biol Evol 2010;27:811-8.
41. Faria NR, Suchard MA, Abecasis A, et al. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. Infect Genet Evol 2012;12:453-60.
42. Aibekova L, Foley B, Hortelano G, et al. Molecular epidemiology of HIV-1 subtype A in former Soviet Union countries. Plos One 2018;13:e0191891.
43. Nabatov AA, Kravchenko ON, Lyulchuk MG, Shcherbinskaya AM, Lukashov VV. Simultaneous introduction of HIV type 1 subtype A and B viruses into injecting drug users in southern Ukraine at the beginning of the epidemic in the former Soviet Union. AIDS Res Hum Retroviruses 2002;18:891-5.
44. Kazennova E, Laga V, Gromov K, et al. Genetic Variants of HIV Type 1 in Men Who Have Sex with Men in Russia. Aids Res Hum Retrov 2017;33:1061-4.
45. Mir D, Graf T, Almeida SED, Pinto AR, Delatorre E, Bello G. Inferring population dynamics of HIV-1 subtype C epidemics in Eastern Africa and Southern Brazil applying different Bayesian phylodynamics approaches. Sci Rep-Uk 2018;8.
46. Ferdinandy B, Mones E, Vicsek T, Muller V. HIV competition dynamics over sexual networks: first comer advantage conserves founder effects. Plos Comput Biol 2015;11:e1004093.
47. Hue S, Brown AE, Ragonnet-Cronin M, et al. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. AIDS 2014;28:1967-75.
48. van de Laar TJ, Bezemer D, van Laethem K, et al. Phylogenetic evidence for underreporting of male-to-male sex among human immunodeficiency virus-infected donors in the Netherlands and Flanders. Transfusion 2017;57:1235-47.
49. Esbjornsson J, Mild M, Audelin A, et al. HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. Virus Evol 2016;2:vew010.
50. Kiriazova T, Postnov O, Bingham T, et al. Patient and provider perspectives inform an intervention to improve linkage to care for HIV patients in Ukraine. BMC Health Serv Res 2018;18:58.
51. Malta M, Ralil da Costa M, Bastos FI. The paradigm of universal access to HIV-treatment and human rights violation: how do we treat HIV-positive people who use drugs? Curr HIV/AIDS Rep 2014;11:52-62.
52. Karamov E, Epremyan K, Siniavin A, et al. HIV-1 genetic diversity in recently diagnosed infections in Moscow: predominance of A_{FSU}, frequent branching in clusters, and circulation of the Iberian subtype G variant. AIDS Res Hum Retroviruses 2018.
53. Paraskevis D, Pybus O, Magiorkinis G, et al. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. Retrovirology 2009;6:49.
54. Simmons R, Malyuta R, Chentsova N, et al. HIV Incidence Estimates Using the Limiting Antigen Avidity EIA Assay at Testing Sites in Kiev City, Ukraine: 2013-2014. Plos One 2016;11.
55. Kingman JFC. The coalescent. Stochastic Processes and Their Applications 1982;13:235-48.

56. Boskova V, Bonhoeffer S, Stadler T. Inference of epidemiological dynamics based on simulated phylogenies using birth-death and coalescent models. Plos Comput Biol 2014;10:e1003913.
57. Stadler T, Vaughan TG, Gavryushkin A, et al. How well can the exponential-growth coalescent approximate constant-rate birth-death population dynamics? P Roy Soc B-Biol Sci 2015;282.

5 Tracing the impact of public health interventions on HIV-1 epidemic in Portugal using genetic epidemiology

Contributions by collaborators: Louis du Plessis developed the hierarchical model for the BEAST2 analysis, Andrea C Pineda-Peña and Ana B Abecasis selected Portuguese sequences for the analysis, Ricardo J Camacho and Perpétua Gomes created and managed the Portuguese drug resistance database, from which the sequences for the analysis came. Nuno R Faria performed some preliminary analyses and supervised this work.

5.1 Abstract

Estimation of temporal changes in HIV transmission patterns can help to elucidate the impact of preventive strategies and public health policies. Portuguese HIV-1 subtype B and G *pol* genetic sequences were appended to global reference datasets to identify country-specific transmission clusters. I used Bayesian birth-death models (BDM) to estimate subtype-specific effective reproductive numbers (R_e). To quantify mixing among transmission groups I used phylogeographic discrete trait analyses (DTA). I identified five subtype B Portuguese clusters (N=26-79) and a large monophyletic subtype G Portuguese clade (N=236). I estimated that major shifts in HIV-1 transmission patterns occurred around 1999 (95%HPD 1998-2000) and 2000 (1998-2001) for subtypes B and G, respectively. For subtype B, R_e dropped from 1.91 (1.73-2.09) to 0.62 (0.52-0.72). For subtype G, R_e decreased from 1.49 (1.39-1.59) to 0.72 (0.63-0.8). The DTA suggests that people who inject drugs (PWID) and heterosexuals were at the source of most (>80%) virus lineage migrations for subtypes G and B, respectively. The estimated declines in R_e coincide with the introduction of highly active antiretroviral therapy and the scale-up of harm reduction for PWID in Portugal. Observed transmission events across transmission groups emphasize the importance of prevention efforts for bridging populations.

5.2 Introduction

HIV-1 subtype B dominates the HIV epidemic in Western European countries[1]. Since 1980s, the Portuguese HIV epidemic was concentrated in heterosexual (HET) populations, in which subtype B was the most common strain[2]. However, by 2001 nearly two thirds of reported AIDS cases in Portugal were linked to people who inject drugs (PWID)[3], and the country had the highest HIV incidence in PWID in the European Union[4,5]. Although the HIV Portuguese epidemic in PWID was initially associated with transmission of subtype B, subtype G and CRF14_BG infections grew rapidly in this transmission group[6]. Furthermore, there is some recent evidence of mixing between transmission groups[7], which may increase the magnitude of the epidemic over time[8]. Currently, the Portuguese HIV molecular profile is very diverse, with a variety of HIV-1 subtypes (A, C, D, H, J, F) and recombinant forms, as well as a higher proportion of HIV-2 infections compared to other European settings[1,2,6,7,9-11]. Besides the introduction of anti-retroviral therapy (ART) in 1996, several public health approaches have been used to reduce HIV incidence in Portugal. Since the early 1990s Portugal has been at the forefront of the implementation of harm reduction strategies[12]. Methadone opioid substitution treatment (OST) has been available to PWID since 1977, while buprenorphine OST became available in 1999. In 2004 OST has been extended to pharmacy-based provision. The first syringe-exchange program (SEP) was established in 1993. In 2001, Portugal effectively decriminalised drug possession, which reportedly brought a number of positive health outcomes for the PWID population, resulting in the reduction of the number of drug-related deaths and possibly a drop in HIV incidence[13,14].

The basic reproductive number (R_0) is defined as the number of secondary infections per primary infected individual in a completely susceptible population. In contrast to

R_0, the effective reproductive number (R_e) is defined as the average number of secondary infections caused by an infected individual at a particular time point during an epidemic. Both R_0 and R_e are used to describe epidemic growth, but R_e is a time-dependent quantity that changes over the course of an epidemic. Recent developments in phylodynamic models allow to combine genetic and epidemiological modelling to quantify R_0 and R_e in HIV populations (see Chapter 2 and [15-17]). Yet, R_0 and R_e remain to be estimated for the main subtypes and transmission groups associated with the Portuguese HIV epidemic.

Here I collate 680 HIV-1 subtype B and G *pol* genetic sequences with known transmission group information from the Portuguese HIV drug resistance database. I use several phylodynamic approaches to investigate temporal changes in the rate of epidemic spread of the predominant subtypes and transmission clusters, and describe the mixing and viral flow between different transmission groups. My findings highlight the effect of targeted interventions on the HIV-1 dynamics in Portugal.

5.3 Methods

5.3.1 Genetic data

HIV-1 subtype B and G *pol* genetic sequences were obtained from the Portuguese HIV database, which contains genotypes of more than 10,000 HIV-infected patients from the Portuguese mainland, and the Madeira and Azores archipelagos. The Portuguese HIV database includes *pol* nucleotide sequences with the following associated information: hospital of attendance, date of sample collection, and self-reported transmission group (for some sequences). For the purpose of this analysis I considered only sequences collected between 2001-2013 with available information regarding (self-reported) transmission group. Such data were retrieved from patients

who attended 18 hospitals located over the Portuguese mainland and the Madeira Archipelago. If two sequences were available for the same patient, only the sequence with the earlier sampling date was used. Reference HIV-1 subtype B sequences were obtained from the SPREAD database (SPREADdb) created by the European Society for Translational Antiviral research (ESAR) (http://www.esar-society.eu/). The SPREADdb contains annotated sequences generated from newly diagnosed patients during 2001-2013 from 26 European countries[18]. For each Portuguese subtype B sequence, I extracted the ten most similar sequences from the SPREADdb. For HIV-1 subtype G, I used all publicly available subtype G sequences from the HIV Los Alamos database (LANLdb)[19] as reference sequences. I first codon-aligned all the Portuguese sequences using Clustal W[20] and subsequently edited the alignment manually in MEGA7[21]. Since subtype G and CRF14_BG do not have breakpoints in *pol* nucleotide sequences, a stringent subtyping approach was used. All sequences were subtyped using Rega subtyping tool version 3[22] and COMET[23] and only sequences with concordant subtype assignments were considered for subsequent analyses. Results from the manual phylogenetic subtyping analyses previously performed[22] were also considered. Potential CRF14_BG sequences or BG recombinants were excluded from the analyses.

I then aligned all sequences (including the reference) using Clustal W software[20]. Initial alignments were 879 and 1020 nucleotides (nt) long for subtypes B and G, respectively. I manually edited the alignment using MEGA7 program[21] to delete codon positions associated with drug resistance (RT: 41, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 115, 116, 151, 179, 181, 184, 188, 190, 210, 215, 219, 225, 230; PR: 23, 24, 30, 32, 46, 47, 48, 50, 53, 54, 73, 76, 82, 83, 84, 85, 88, 90)[24]. Deleting sites associated with drug resistance can hide similarities between sequences caused

by some shared epidemiological characteristics (i.e. transmission group or shared donor for transmitted drug resistance). However, this is a usual practice to avoid convergent evolution effect, i.e. when sequences that are unrelated appear closely related on a phylogenetic tree because they were exposed to the same treatment. The final alignments were 804 and 945 nt long for subtypes B and G, respectively.

5.3.2 Initial phylogenetic and molecular clock analyses

I used the combined subtype B and G datasets (Portuguese, SPREADdb and LANLdb) to generate maximum likelihood (ML) phylogenies for each subtype in FastTree[25]. I used the Hasegawa-Kishino-Yano (HKY+G) nucleotide substitution model with gamma-distributed rate variation among sites. For subtype B, I used Cluster Picker[26] to define HIV-1 clusters as clades with >10 sequences, within-cluster genetic distance <6% and SH-like statistical support >90%. Given that most Portuguese subtype G sequences belonged to a single well-supported clade (SH-like support 90%), I analysed all sequences (N=236) as single phylogeny/cluster. For each cluster I constructed a ML tree using RAxML[27], and quantified temporal signal using TempEst[28]. Specifically, I regressed genetic divergence against sampling dates to estimate temporal structure[28]. I find that each subtype B cluster contained a moderate temporal signal as indicated by correlation coefficients, R (Table 5.1).

Table 5.1. Root-to-tip correlation analysis in TempEst

Tree	Correlation Coefficient (R)
Cluster 1	0.54
Cluster 2	0.22
Cluster 3	0.37
Cluster 4	0.35
Cluster 5	0.33
Subtype G	0.23
Subtype G (with reference outgroup)	0.4

For the subtype G, the molecular clock signal was weak (R=0.23). Thus, I used a normal prior on the time of the most recent common ancestor (TMRCA) of the subtype G clade. To increase temporal resolution of the dataset, I downloaded all publicly available subtype G sequences from LANLdb collected before 2001 (N=32, see Appendix 5 for accession numbers) and added them to the 236 Portuguese sequences (full dataset, N=268). I reconstructed the maximum likelihood (ML) tree for the full dataset using RAxML (Fig. 5.1). This allowed me to improve the root-to-tip correlation (R), estimated using TempEst, from R=0.23 to R=0.40 (Table 5.1).

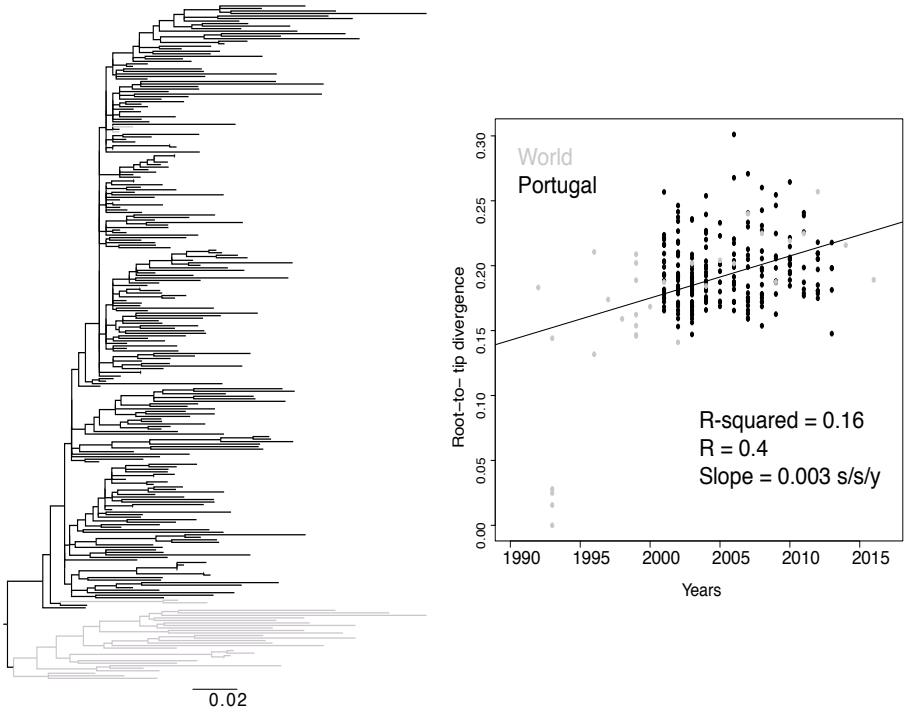


Figure 5.1. Maximum Likelihood phylogenetic tree of the Portuguese subtype G dataset (N=236) and all publicly available from the Los Alamos database subtype G sequences sampled prior to 2001 (N=32) (left). The bar indicates the equivalent of the genetic distance in substitutions/site/year. Results of the TempEst analysis for the tree (right). R – correlation coefficient.

I then used BEASTv1.8.4 to estimate the TMRCA from the full dataset under a strict molecular clock model, a Hasegawa-Kishino-Yano (HKY) nucleotide substitution model and a Skygrid coalescent tree prior[29]. The obtained TMRCA estimate for the subtype G Portuguese clade was 1977 (standard deviation of 2.1); this was used as a normal prior on subsequent subtype G analyses.

5.3.3 Reproductive number estimates

I used birth-death models (BDM) implemented in BEASTv2.4[30] to estimate time-varying rates of epidemic spread, measured as changes in the effective reproductive number, R_e. In this model, lineages are added to the tree upon infection, at rate λ (birth rate), and removed at rate δ (death rate), upon becoming uninfectious (either through treatment, isolation or death). I make the assumption that individuals become uninfectious upon sampling and are removed from the pool of transmitting individuals. Here, R_e is given by λ/δ and the average infectious period by $1/\delta$. I also estimate the sampling proportion and the origin of the epidemic (which is older than the TMRCA of the tree reconstructed from the sampled sequences). In addition, R_e, δ and the sampling proportion are allowed to instantaneously change at arbitrary time-points (thus these parameters are piecewise constant functions of time).

I used 3 different variations of the BDM in this analysis. In Model 1 (*constant rate birth-death*) R_e was assumed to be constant through time, resulting in an average R_e over the course of the epidemic[15]. Model 2 (*birth-death skyline, BDSKY*) assumes a piecewise constant R_e over 10 equidistant time-intervals between the TMRCA and the most recent sample[31]. Model 3 (*constant-shift-constant*) assumes a piecewise constant R_e, but with two-epochs, in analogy with a previous described coalescent model[32]. In addition to estimating the average R_e during an early and a late epoch, the model is

able to estimate the timing of the transition (shift) in the rate of the epidemic spread between the two epochs directly from sequence data.

R_e estimates were inferred separately for the two subtypes. For the subtype B clusters, I first used a hierarchical phylogenetic approach for the molecular clock, nucleotide substitution, and birth-death tree prior model parameters across all subtype B clusters[33]. Since the times of the most recent sampling dates and the TMRCAs will differ between clusters, I used cluster 3 to define the shift-times (times of the change of BDSKY intervals) of the birth-death skyline model, because the time of cluster 3's most recent sample (2008) is older than the time of the most recent sample in the other clusters. Thus, the time of the shifts was calculated on a scale between the TMRCA and the youngest tip in cluster 3.

For subtype B, I also performed BDSKY analyses (Model 2) with all five HIV-1 clusters analysed as a single phylogenetic tree to test the robustness of the hierarchical model approach. For subtype G, I ran additional analysis including PWID sequences only using Models 2 and 3. Specifically, I chose sequences that could be traced to the root of the phylogenetic tree and had a high posterior probability (>0.85%) for all internal nodes (N=99) as previously described[34].

I used strict molecular clock model and the HKY+G nucleotide substitution model[35]. The following assumptions were encoded as priors in the Bayesian analyses: (i) lognormal prior with mean 1.0 and standard deviation 1.25 was placed on the clock rate; (ii) lognormal prior with mean 1.0 and standard deviation 1.25 was used for the kappa parameter in HKY model; (iii) uniform prior between the TMRCA of the sampled tree and 100 years before the most recent sample was used for the origin of each cluster, (iv) lognormal distribution with mean 1.0 and standard deviation 1.25 was placed on R_e, (v) sampling proportion was set to 0 between the TMRCA and the

oldest sample, which was in 2001 for all clusters and both subtypes; after 2001 (between 2001 and the most recent sample in each cluster) a Beta distribution with parameters α=1.0 and β=10.0 was used for the sampling proportion, (vi) for the constant-shift-constant model a uniform prior between 1992 and 2007 was placed on the shift date, (viii) for subtype G I used a normally distributed TMRCA prior with mean 1976.9 and standard deviation 2.1 as described in section 5.3.2. Finally, the removal rate for all the BD models was fixed to 0.5 $years^{-1}$ (corresponding to an average infectious period of 2 years). In sensitivity analyses I varied the removal rate between 0.1 and 4.0 $years^{-1}$, resulting in average infectious periods between 3 months and 10 years. For all BEAST analyses mixing of the Markov Chain Monte Carlo (MCMC) was visually inspected in Tracer[36]; effective sample sizes were >200 for all parameters.

Additionally, for all the BDSKY analyses I chose 0.05 as a starting value for the sampling proportion parameter, and allowed this parameter to vary between 0 and 0.2. This value was chosen based on the number of subtype B and subtype G sequences in each cluster per year relative to the number of registered HIV-1 cases that year and the proportion of subtype B (≈40%) and subtype G (≈20%) cases[14]. The proportion of sequenced sequences never exceeded 9% by my estimates (Table 5.2); thus, 0-0.2 was a reasonable interval for this parameter.

Table 5.2 (next page). The number of sampled sequences per year per cluster, and the proportion of sampled sequences out of the estimated total number of patients diagnosed that year with the corresponding HIV-1 subtype

Cluster		Year												Overall
		2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	
Number of sequences	1	1	1	2	5	4	0	3	3	0	0	0	7	0	10
	2	8	16	15	11	2	7	6	4	2	4	1	3	0	14
	3	1	3	6	5	2	0	1	2	0	0	0	0	0	2
	4	3	9	9	7	2	3	0	2	1	2	0	1	1	7
	5	5	8	4	10	3	3	4	1	0	0	1	3	0	5
Total in clusters		18	37	36	38	13	13	14	12	3	6	2	14	1	38
Total in the analysis (subtype B)		37	74	93	46	15	34	26	27	21	16	17	30	8	119
Total in the analysis (subtype G)		24	32	45	23	9	17	19	18	11	13	8	12	5	67
Total number of patients diagnosed that year (all subtypes)		2842	2683	2485	2409	2225	2270	2173	2242	2041	1937	1685	1607	1464	10976
Number of patients diagnosed that year (subtype B = 40%)		1137	1073	994	964	890	908	869	897	816	775	674	643	586	4390
Number of patients diagnosed that year (subtype G = 40%)		568	537	497	482	445	454	435	448	408	387	337	321	293	2195
Proportion of sampled individuals with subtype B		3	7	9	5	2	4	3	3	3	2	3	5	1	3
Proportion of sampled individuals with subtype G		4	6	9	5	2	4	4	4	3	3	2	4	2	3

5.3.4 Discrete trait and association index analyses

To characterise the population structure according to transmission groups, I estimated a phylogenetically-based association index (AI) statistic, which quantifies the association between a phylogeny and trait values associated with each tree tip. Since the AI value typically depends on the number of tips in a phylogeny, I used several methods to estimate this parameter. First, I used BEAST to obtain AI normalised by the values obtained by tip randomisation (0 to 1, with 0=maximum structure and 1=panmixis), which allowed direct comparison of AI between clusters of different sizes; ii) I used BaTS[37] to obtain statistical support (p-values) for how the mean posterior estimate of AI from the empirical tree distribution is different from the AI obtained by randomly re-assigning traits at the tips 100 times; iii) I used HyPhy[38] to estimate Simmonds' AI statistic, which assesses compartmentalisation by weighting internal nodes based on their position on the tree, giving less value to the nodes near the root; it also allowed me to obtain statistical support (bootstrap values) for the observed AI[39]. I chose topology-based methods that incorporate phylogenetic uncertainty (like BaTS and Simmons AI) since these methods tend to perform best when studying shorter HIV-1 sequences like the ones used here[40].

To describe and quantify viral movement between different transmission groups, I performed a discrete trait diffusion analysis (DTA) using transmission group as a discrete trait[41] for clusters with significant transmission group structure. Initially, each sequence was annotated with one of the five known transmission groups, specifically: "MSM", "PWID", "HET", "MTC" (mother-to-child transmission), or "Other" (e.g. blood transfusion). To count the expected number of virus lineage movements across transmission groups I used a robust counting (Markov jumps) approach[42-44]. Based on preliminary analyses, I next considered three transmission groups (with the two

transmission groups that contributed the highest proportion of migration events in a given dataset, and with all other transmission groups included in the "Other"). For both analyses I calculated Bayes Factor (BF) support for each mixing pathway using a Bayesian stochastic search variable selection (BSSVS) procedure[44] and SpreaD3[45]. I assumed a BF>10 to be a strong evidence for a well-supported viral pathway between two transmission groups[44,46,47].

To increase computational speed, I ran BEAST analyses using the BEAGLE library[48] and an empirical distribution of 2,000 post-burnin (burn-in 10% and 20% for subtypes B and G, respectively) trees for each cluster in the analysis[34]. I used the following priors for this analysis: (i) an asymmetric continuous-time Markov chain prior was used for the transmission group clock rate[49]; (ii) a uniform prior was used for the transmission group frequencies; (iii) a gamma prior with shape parameter 1.0 was used for the transmission group rates; (iv) a uniform prior was used for the root frequencies.

5.4 Results

5.4.1 Genetic sequences

A total of 3462 HIV-1 subtype B and 2678 subtype G *pol* genetic sequences were available from the Portuguese HIV database. Of these, transmission group information was available for 444 subtype B and 236 subtype G sequences. Additionally, 2165 unique reference sequences were obtained from the SPREADdb for subtype B and 1008 unique reference sequences were obtained from the LANLdb for subtype G. Datasets that include both Portuguese and reference sequences will be referred to as "combined" datasets. The combined datasets therefore comprised 2609 and 1244 unique sequences for subtypes B and G, respectively (Fig. 5.2).

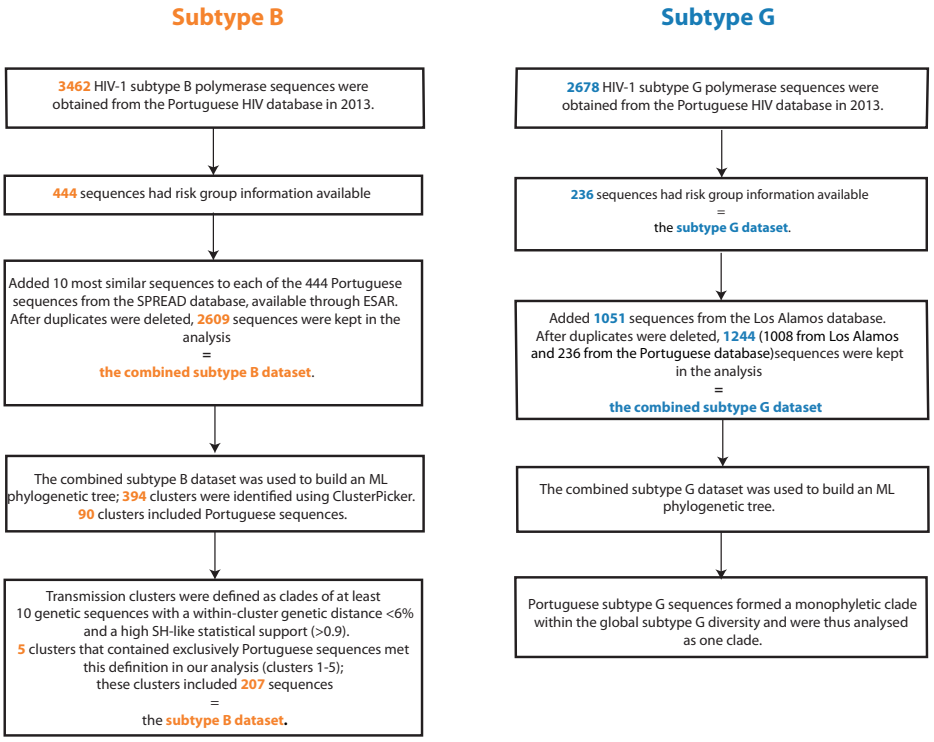


Figure 5.2. Diagram describing the process of selecting subtype B and G sequences for the analysis.

5.4.2 Phylogenetic cluster analysis

The ML phylogenetic trees estimated for the subtype B and G combined datasets are shown in Fig. 5.3. For subtype B, I identified 394 clusters, of which 90 clusters representing 334 patients included Portuguese sequences only. Five Portuguese transmission clusters contained >10 sequences each and were named as clusters 1 (N=26), 2 (N=79), 3 (N=20), 4 (N=40), and 5 (N=42) (Fig. 5.3). All of these clusters, except the largest (cluster 2), were also identified when a more conservative cluster definition was used (genetic distance threshold <4.5%).

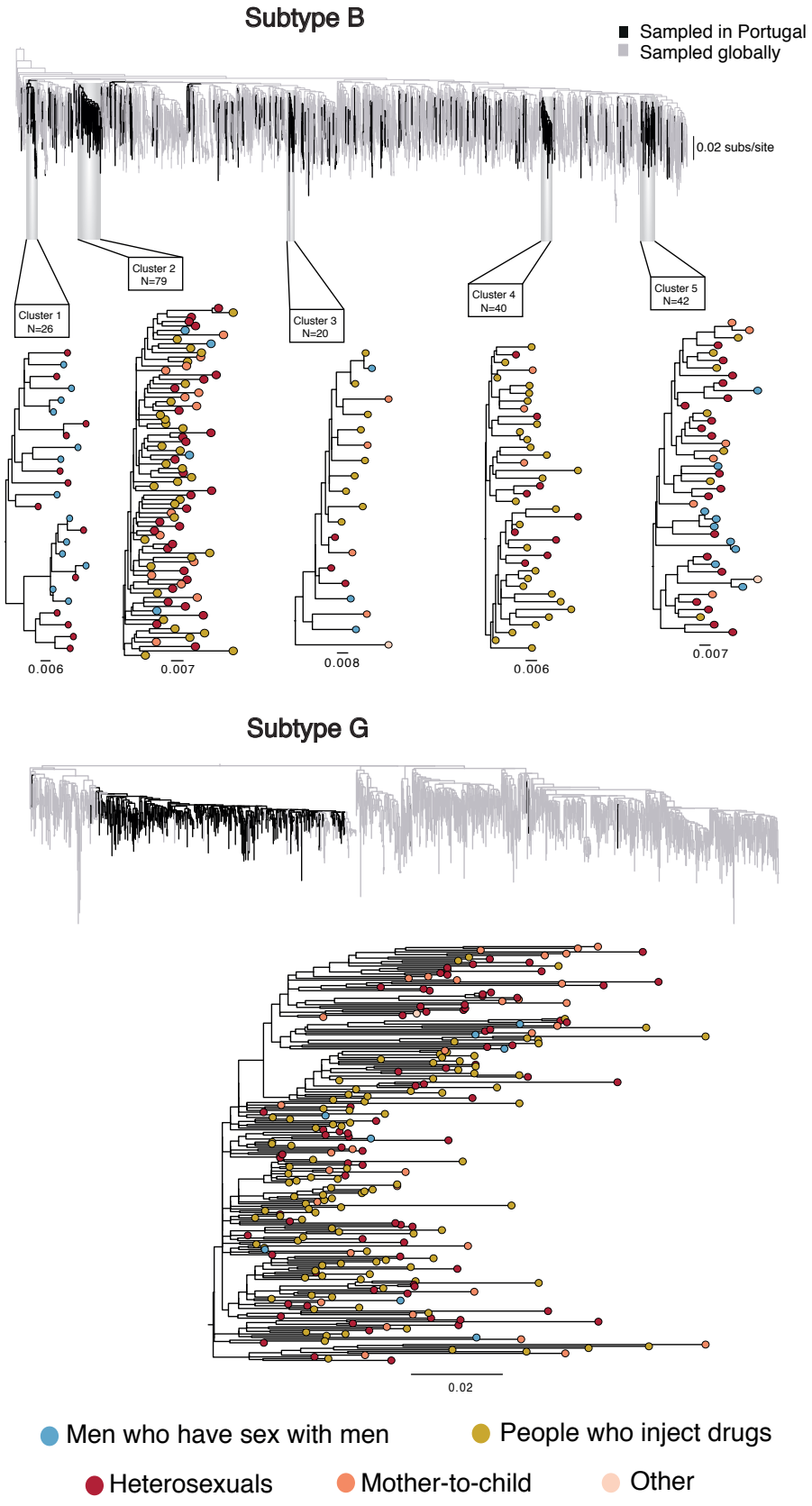


Figure 5.3. Top: Maximum Likelihood (ML) phylogenetic trees reconstructed from the combined subtype B (top) and subtype G (bottom) datasets. The black colour

indicates Portuguese sequences and is used to highlight their position on the global diversity trees. The highlighted clusters are the clades of at least ten sequences with a within-cluster genetic distance <6% and SH-like statistical support >90%. Bottom: ML phylogenetic trees of the identified phylogenetic clusters. The colours of the tips represent the transmission group of a patient from whom the sequence was sampled.

Clusters were located across the global subtype B diversity, indicating that subtype B has been introduced multiple times to Portugal. In contrast, most of the Portuguese subtype G sequences clustered together in a single clade within the global subtype G tree (SH-like support 0.90). Thus, subtype G sequences were subsequently analysed as a single epidemic clade. Information about the number of sequences per cluster, sampling intervals (years), and patient transmission group, for both subtypes, is provided in Table 5.3.

Table 5.3. Distribution of sequences in clusters for subtype B and in the subtype G clade

Subtype B cluster													Subtype G	
Trans-mission group	1		2		3		4		5		All clusters			
	N	%	N	%	N	%	N	%	N	%	N	%	N	%
MSM	13	50	4	5	3	15	-		9	21	29	14	8	3
PWID	-	-	32	41	9	45	28	70	7	17	76	37	110	47
HET	13	50	29	37	3	15	9	22.5	19	45	73	35	82	35
MTC	-	-	14	18	4	20	3	7.5	6	14	27	13	35	15
Other	-	-	-	-	1	5	-		1	2	2	1	1	0
Total	26	100	79	100	20	100	40	100	42	100	207	100	236	100
Sampl. years	2001-2012		2001-2012		2001-2008		2001-2013		2001-2012		2001-2013		2001-2013	

5.4.3 Estimates of the effective reproductive number using genetic data

To estimate temporal changes in the effective reproductive number I used a suite of birth-death models (BDM) that take into account shared ancestry and phylogenetic uncertainty. The constant-rate BDM (Model 1) suggested that the number of secondary infections was on average marginally above one over the course of the epidemic (Fig. 5.4, A and C). For example, assuming that an HIV-infected person remains infectious for a period of 2 years, after which this person is either virally suppressed or removed from a population (relocation or death), the median posterior estimate of R_e was 1.15 (95% Highest Posterior Density, HPD, interval: 1.09 to 1.2) for subtype B and 1.23 (95% HPD: 1.2 to 1.27) for subtype G (Fig. 5.4, A and C). A sensitivity analysis specifying various infectious periods (6 months to 5 years, removal rate 0.2-2.0) shows that the R_e estimates decrease with the shorter infectious periods but remain between 1 and 2 for all the specified values (Fig. 5.4, A and C).

When using a BDSKY model (Model 2), which allows for changes in the rate of transmission through time, I find that subtype B's R_e reached a peak around 2001 (R_e = 2.17, 95% HPD 1.53 to 2.82), after which R_e dropped below 1 after 2004 (95% HPD 2002.5-2008) (Fig. 5.4, B). For subtype G, the maximum value for R_e was estimated to exist in 1996 (R_e = 1.6, 95% HPD 0.98 to 2.2) (Fig. 5.4, D). Subtype G's mean R_e dropped below 1 from 2000 onwards (R_e = 0.51, 95% HPD 0.29 to 0.75). To investigate whether this pattern was observed in PWID populations, I repeated the subtype G analysis including only sequences from the PWID group. The reconstructed BDSKY from the PWID-only sequences showed similar patterns to the BDSKY reconstructed from the full dataset, with a rapid decline after 1996 (Fig. 5.5). Interestingly, for both subtypes, and for the PWID subsample, I detected a slight increasing trend in R_e during more recent years.

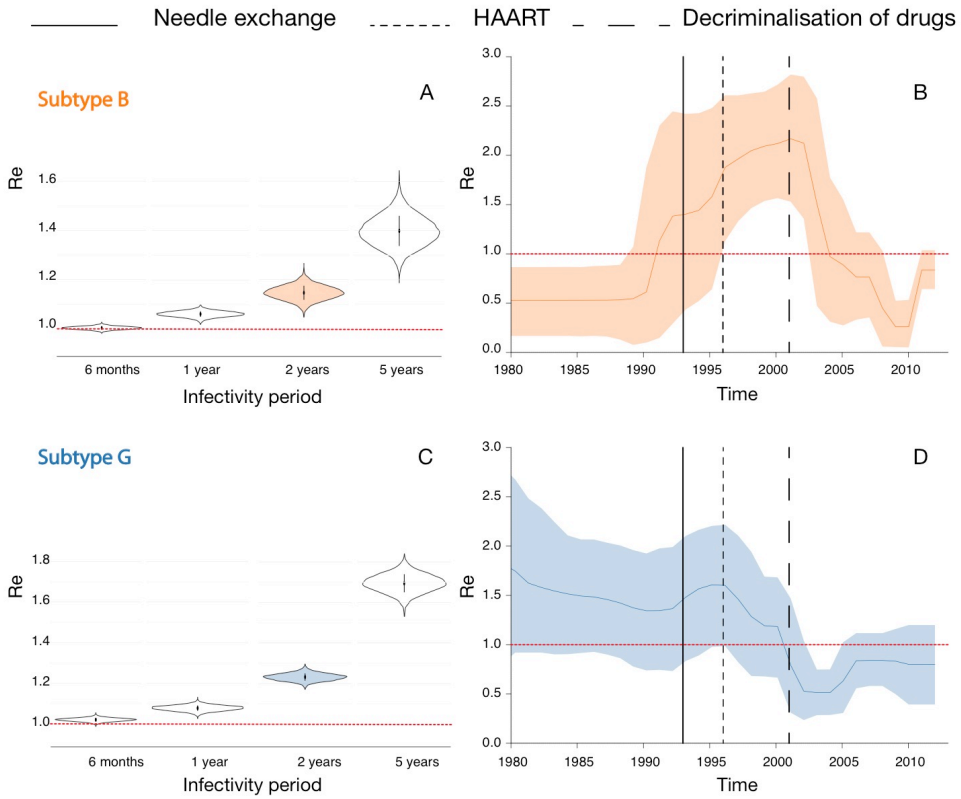


Figure 5.4. Panels A and B represent R_e estimated using the BD Model 1 (constant R_e) under different assumptions of infectious period (6 months – 5 years, becoming uninfectious rate 2.0 – 0.2) for subtypes B (panel A) and G (panel C). The vertical lines in the middle of the violin plots represent the interquartile range. The violin plots filled with colour represent the R_e values obtained by assuming the infectious period equals 2 years. The horizontal red dotted line represents the epidemiological threshold (R_e=1). Panels B and D represent R_e estimates obtained using BD Model 2 (BDSKY) over 10 equidistant time-intervals. Results for subtypes B (panel B) and G (panel D) were estimated by assuming the infectious period equals 2 years. The shaded area represents the 95% Highest Posterior Density (HPD) credible region. The horizontal red dotted line represents the epidemiological threshold (R_e=1). Vertical lines correspond to the time of introduction of major HIV preventive interventions (see text).

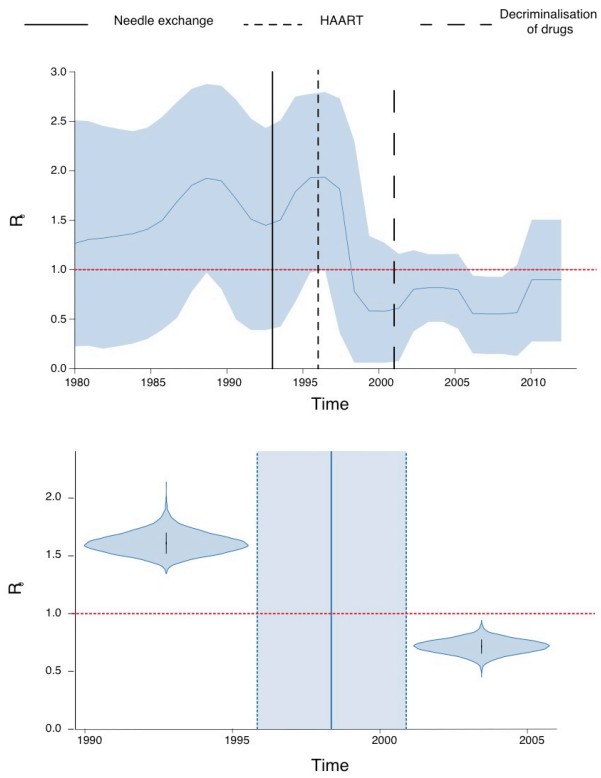


Figure 5.5. Subtype G R_e estimated using BD Models 2 (top) and 3 (bottom), using only PWID sequences (N=99). The shaded area represents the 95% Highest Posterior Density (HPD) credible region. The horizontal red dotted line represents the epidemiological threshold (R_e=1). Top: Vertical lines correspond to the time of introduction of major HIV preventive interventions. Bottom: The violin plots of R_e distribution before and after the time of estimated change in the rate of epidemic spread. The vertical lines in the middle of the violin plots represent the interquartile range.

Finally, the two-epoch BDM (Model 3), revealed initial phases of epidemic growth characterised by high R_e values, well above the epidemic threshold that later transitioned to a slower transmission phase. For subtype B, around the year 1999 (95% HPD 1998 to 2000) R_e decreased from a median estimate of 1.92 (95% HPD 1.75 to 2.1) to 0.62 (95% HPD 0.52 to 0.72) (Fig. 5.6). For subtype G, around the

year 2000 (95% HPD 1998 to 2001) R_e decreased from a median estimate of 1.49 (95% HPD 1.39 to 1.59) to 0.71 (95% HPD 0.63 to 0.8) (Fig 5.6). When performing analyses that include only sequences from PWID, I obtained trends similar to those obtained for the full subtype G dataset, but with a broader confidence interval, most likely because of the smaller number of sequences (Fig. 5.5): R_e transitioned from ~1.61 (95% HPD 1.44 to 1.79) to 0.71 (95% HPD 0.59 to 0.83). The time of the major shift in transmission dynamics was estimated to be 1998 (95% HPD 1996-2001), which is also similar to the estimate obtained using the full dataset. Notably, Model 3 indicates that the major transitions in subtypes B and G transmission dynamics occurred around the same period in time.

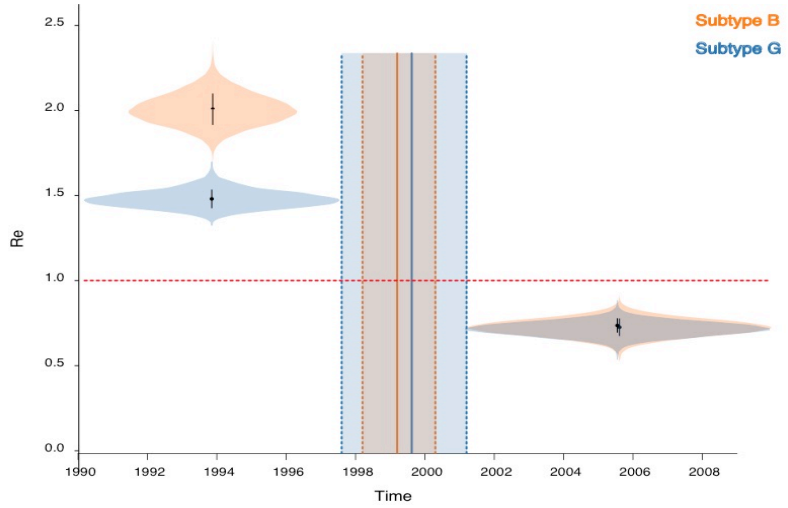


Figure 5.6. R_e estimates obtained using BD Model 3 (2 time-intervals) and the estimated time of change in the rate of epidemic spread for both subtypes. The violin plots of R_e distribution before and after the time of estimated change in viral spread rate are illustrated for both subtypes. The shaded area represents the 95% Highest Posterior Density (HPD)credible region. The horizontal red dotted line represents the epidemiological threshold (R_e=1).

Using the single-tree approach for subtype B, the median posterior R_e was estimated to be 1.2 (95% HPD interval 1.17 – 1.24). The BDSKY model showed that the R_e started declining after 1998 (95% HPD 1999 – 2000), and dropped significantly below 1 after 2005. Further, I estimated the time of the major shift in transmission dynamics to be in 1999 (95% HPD 1998 – 2000), with R_e being 2.00 (95% HPD 1.80 – 2.20) before and 0.57 (95% HPD 0.47-0.66) after the shift. All these estimates are similar to those obtained from the hierarchical model, supporting the robustness of my findings.

5.4.4 Inferring mixture of risk behavior

To quantify the level of phylogenetic clustering by transmission behavior I used three distinct approaches to quantify the association index (AI) (Table 5.4), and two different discretization schemes, assigning sequences to 5 or 3 different transmission groups. For the 5-transmission group discretization scheme, BaTS indicated no significant transmission group structure in subtype B clusters in contrast to the significant structure detected for subtype G (p-value <0.01). Simmonds' AI statistic suggests evidence of compartmentalization in subtype B cluster 5 (AI=0.77, bootstrap support 0.97) and in the Subtype G clade (AI=0.84, bootstrap support 0.98). Similar patterns for both subtypes were observed for the 3-transmission group discretization scheme.

I next performed discrete trait diffusion analyses (DTA) for subtype B cluster 5 and for subtype G. The DTA results for subtype B cluster 5 suggests that HET individuals disproportionally contributed to viral dissemination, as this group accounted for 82% of viral migration events (N=17, 95% HPD 15-19 out of a total N=21, 95% HPD 19-24), while representing only 45% of the sequences (Fig. 5.7, Table 5.3).

Table 5.4. Association indices (AI) values from the transmission group compartmentalisation analysis. Values in bold indicate statistically significant values (p-value <0.05 or bootstrap support >0.8); values in italics indicate statistically significant estimates at p-value <0.1 or bootstrap support >0.6. Clst – cluster; Btstr – bootstrap support.

		All transmission groups					3 largest groups				
		Re-scaled AI, BEAST	BaTS AI		Simmons AI		Re-scaled AI, BEAST	BaTS AI		Simmons AI	
Sub-type	Clst	Mean	Mean	p-value	Mean	Btstr	Mean	Mean	p-value	Mean	Btstr
B	1	0.80	1.33	0.21	*0.80*	*0.67*	-	-	-	-	-
	2	1.01	5.68	0.09	0.95	0.53	0.99	*5.28*	*0.06*	*0.93*	*0.65*
	3	0.95	1.65	0.28	0.93	0.51	0.97	1.51	0.46	1.00	0.31
	4	0.97	2.03	0.22	0.96	0.39	-	-	-	-	-
	5	0.80	3.01	0.12	**0.77**	**0.97**	0.74	*2.39*	*0.06*	**0.72**	**0.96**
G		**0.95**	**14.42**	**<0.01**	**0.84**	**0.98**	**0.94**	**13.86**	**<0.01**	**0.84**	**0.99**

I then used a Bayesian stochastic search variable selection approach (BSSVS) to identify the statistical support for virus movement among transmission groups[44]. I find that viral pathways with the highest support were: HET to MSM, HET to PWID, HET to MTC, and MSM to OTH (BF = 976, 136, 48, and 22, respectively) (Fig. 5.7). When only three transmission groups were considered (HET, MSM, or "Other"), the proportion of migration events from HET to other transmission groups was estimated at 90% (N=16, 95% HPD: 14-19, out of total N=18, 95% HPD: 15-21). Here, pathways with the highest BF support were HET to MSM and HET to OTH (BF=1105 and 550, respectively).

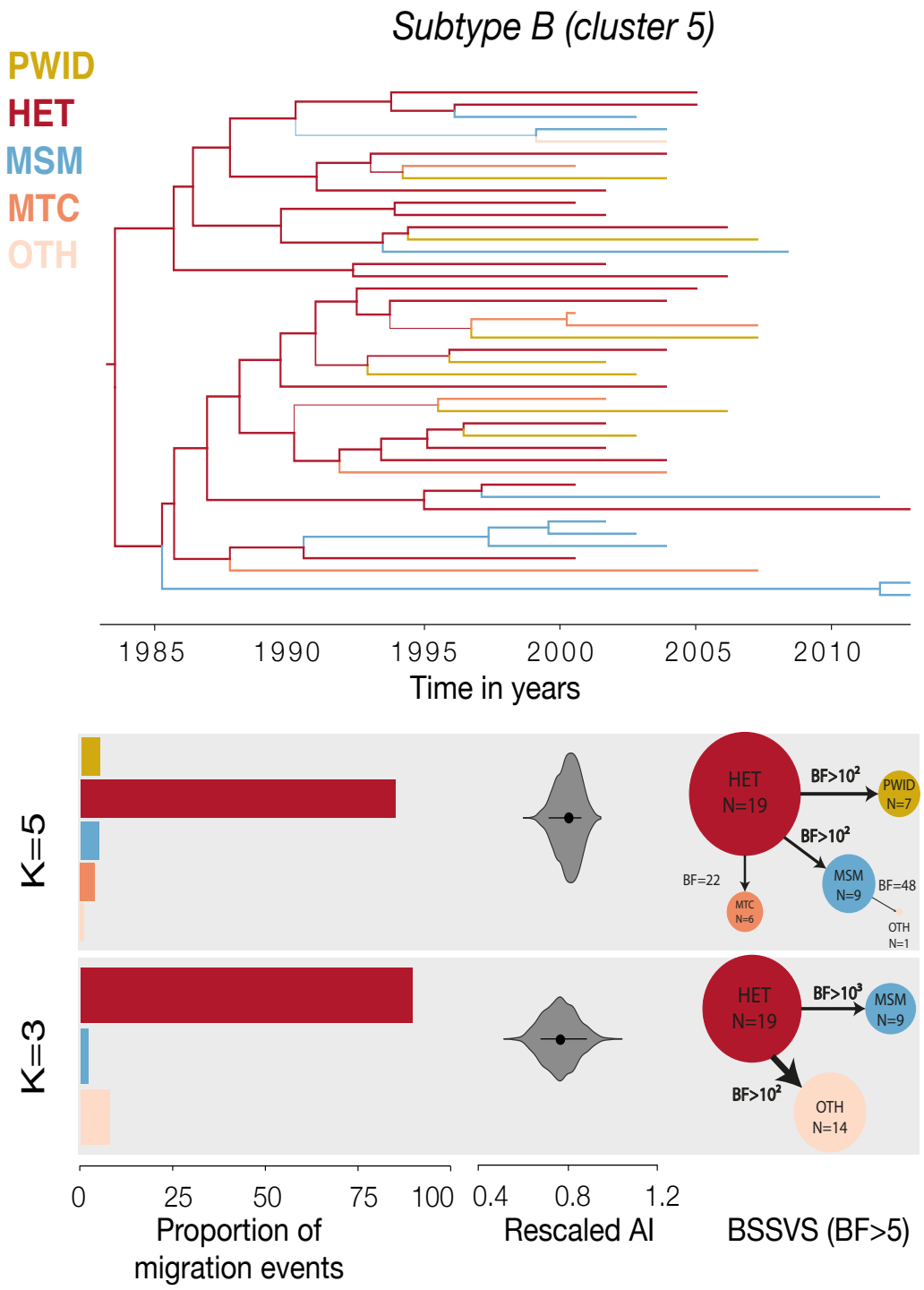


Figure 5.7. Results of the discrete trait analysis (DTA) for subtype B (cluster 5). Top: Molecular clock phylogenetic trees reconstructed for the subtype B cluster 5. The colour of the branches corresponds to the transmission group state; the width of the branches corresponds to the posterior probability support for the indicated transmission group state. Bottom: Results of the DTA analysis for subtype B

considering 5 and 3 transmission groups (k=5 and k=3). The colours represent the transmission group of a patient from whom the sequence was sampled. Bottom panels left to right: 1) Proportion of migration events attributed to a transmission group reconstructed using robust counts approach (Ns correspond to the number of sequences in the transmission group); 2) The rescaled Association index (AI); 3) Bayes Factor (BF) values for the significant virus migration pathways; the size of the circles corresponds to the number of sequences in the analysis; the width of the arrows corresponds to the number of migration events for the corresponding pathways.

Discrete diffusion analyses of subtype G suggests that PWID populations were the main contributor to viral migration across transmission groups, accounting for 81% of viral migration events (N=82, 95% HPD: 60-96, out of total N=101, 95% HPD: 90-110), while representing 47% of the sequences (Fig. 5.8, Table 5.3). HET accounted for 17% (N=18, 95% HPD: 5-32) of viral migrations in the same dataset while comprising 35% of the sequences. All other groups accounted for <1% of migration events. Pathways with the highest support were: PWID to HET, PWID to MTC, and HET to MTC, with BF = 29353, 29353, and 116, respectively.

When only PWID, HET, and "Other" were considered, the proportion of viral migration events remained almost exactly the same, with 82% of migration events attributed to PWID and 17% to HET (N=81 (60-96) for PWID and N=17, 95% HPD: 5-31) for HET out of total N=98, 95% HPD: 88-107). Again, pathways from PWID to HET, PWID to OTH, and HET to OTH gained the highest statistical support (BF = 6630, 6630, and 1656, respectively).

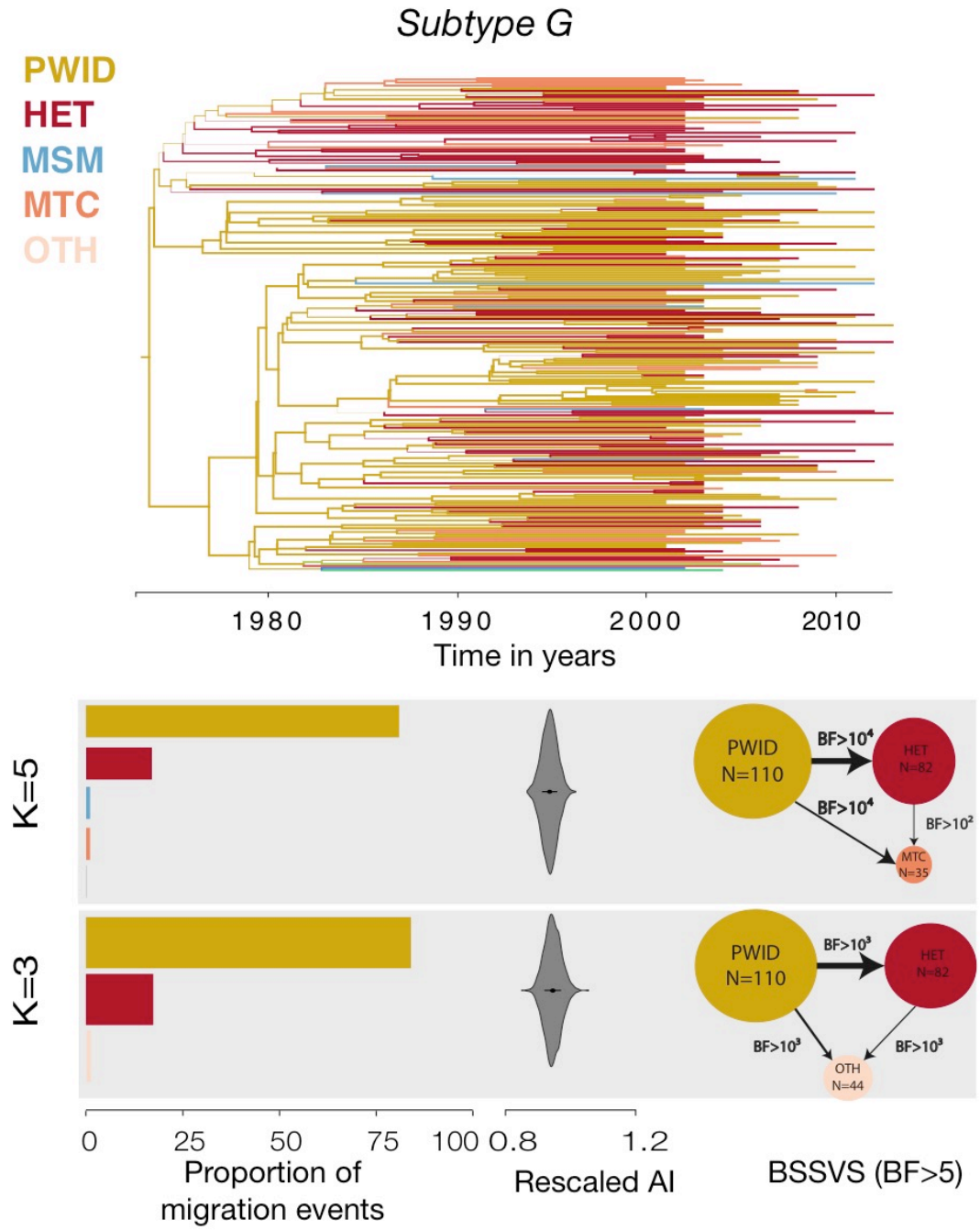


Figure 5.8. Details as per the legend of Figure 5.7, but for subtype G.

5.5 Discussion

I analysed HIV-1 genetic sequences from Portugal to better understand the dynamic patterns of subtype B and G transmission in the country. Remarkably, the analyses of genetic data reveal that the rate of epidemic spread dropped significantly between

1998-2001. Moreover, I found little evidence for phylogenetic clustering by risk behaviour, suggesting frequent mixing between transmission groups.

The estimated reproductive number of the Portuguese HIV-1 epidemic was on average above the epidemiological threshold of 1 throughout the history of subtype B and G epidemics in Portugal. Importantly, when I estimated changes in R_e over time, I was able to show that transmission dynamics of the subtype G clade and subtype B clusters have been well above 1 for more than 10 years. Overall, R_e started to decline around 2001 and 1996 for subtypes B and G, respectively. The timing of this decline coincides with several major public health interventions around the mid- to late 1990s, in particular the introduction of highly active antiretroviral treatment (HAART) in 1996. Interestingly, the reproductive number of subtype G, which is dominated by transmissions in PWID, started to drop long before that of subtype B. This might be explained by a combined effect of HAART and harm reduction programs introduced at the same time, as both needle exchange programs and OST have been shown to slow down existing PWID-driven epidemics and prevent others from establishing[50]. Furthermore, drug decriminalization in 2001 had a major effect on drug related mortality in Portugal, with the number of deaths decreasing from 80 to 16 between 2001-2012[51]. Finally, the timing of a major decline in viral transmission dynamics for both subtypes, with a particularly rapid decrease in R_e for the PWID-driven subtype G, compared to subtype B, coincided with the decriminalization of drugs (Fig. 5.3). This estimated decline in epidemic growth also coincides with a decline in the annual number of new HIV cases[14].

Crucially, the novel hierarchical approach implemented in the BEAST2 framework allowed me to combine the information from five different phylogenetic trees representing five subtype B clusters and estimate epidemiological parameters for this

subtype without forcing the clusters on one phylogenetic tree. Forcing all five clusters onto one tree would mean that the deep branches in the tree (the between-cluster branches connecting the five clusters) are saturated with substitutions, which in turn could lead to biases in the estimation of branch lengths for within-cluster branches. Furthermore, the approach I took here allowed me to reliably estimate the rate of epidemic spread even for the years prior to the sampling date of the oldest sequence in the analysis. This is an important advantage compared to methods that are based solely on the sequences' sampling proportion and might overestimate the R_e for the years before the first sampling date by attributing all the cases to patients diagnosed since that year[17].

I find strong evidence of frequent transmissions between different transmission groups, suggesting that interventions need to target epidemiologically linked transmission groups simultaneously. Mathematic modelling suggests that halting new infections in individuals that act as bridges across transmission groups may have the largest impact in further reducing HIV incidence over time[8]. Historically, subtype B has been mostly associated with sexual HIV transmission in Portugal, while subtype G and CRF14_BG seem to be more frequently reported in PWID[52]. My results strongly suggest that PWID disproportionately contributed to the subtype G epidemic, spreading the virus to HET populations. For subtype B, no transmission group structure was observed for most of the clusters, with only one subtype B cluster showing a particularly high viral flow from HET to MSM. Notably, in this analysis, sequences from MSM clustered closely with sequences from HET and PWID, similarly to the pattern previously reported from Minho province in the north of Portugal[7]. In other European countries and the United States, however, sequences sampled from MSM tend to form isolated clusters that are poorly mixed with other

transmission groups[53-57]. Because I rely on self-reported transmission group information, it is possible that the observed virus lineage movement between HET and MSM results, at least in part, from the misclassification of HET infections and underreporting of MSM contacts, as previously shown in the UK and Nordic countries[58,59].

Portugal has a well-developed surveillance system that enables timely identification of new infections, large-scale ART provision, and routine antiretroviral drug resistance testing, which forms the basis of large database of the protease and reverse transcriptase (PR/RT) genetic region sequences. Unfortunately, the majority of these sequences are unlinked to additional epidemiological information such as the transmission group, stage of infection, etc. This informational gap limits the use of this genetic data for public health purposes. Analysis of between and within transmission group transmissions is critical to design optimal strategies that can directly impact the progress of an HIV epidemic in a country: a recent study showed that by timely targeting clusters of HIV infections, 60% of new infections among MSM could be prevented[54]. Improving the linkage between different patient records would allow conducting studies similar to this on a much larger scale that could directly affect in real-time public health strategies in the country.

In conclusion, this chapter shows how genetic data can help us to address challenging epidemiological questions and support evidence for public health decision-making. Further improvements both in statistical approaches and in surveillance in health care settings can facilitate and encourage the application of phylogenetics to eliminate new infections among distinct transmission groups.

5.6 Acknowledgements

I thank the European Society for translational Antiretroviral Research for providing us with genetic sequence data for this study. I also thank Anne Cori of the Imperial College London and Christophe Fraser of the University of Oxford for early discussions. This study was partially supported by European Funds through grant 'Bio-Molecular and Epidemiological Surveillance of HIV Transmitted Drug Resistance, Hepatitis Co-Infections and Ongoing Transmission Patterns in Europe - BEST HOPE - (project funded through HIVERA: Harmonizing Integrating Vitalizing European Research on HIV/Aids, grant 249697)'; by L''Oréal Portugal Medals of Honor for Women in Science 2012 (financed through L''Oréal Portugal, Comissão Nacional da Unesco and Fundação para a Ciência e Tecnologia (FCT - http://www.fct.pt)); by FCT for funds to GHTM-UID/Multi/04413/2013; by the MigrantHIV project (financed by FCT: PTDC/DTP-EPI/7066/2014); by Gilead Génese HIVLatePresenters.

5.7 References

1. Abecasis AB, Wensing AM, Paraskevis D, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 2013;10:7.
2. Esteves A, Parreira R, Venenno T, et al. Molecular epidemiology of HIV type 1 infection in Portugal: high prevalence of non-B subtypes. AIDS Res Hum Retroviruses 2002;18:313-25.
3. Global AIDS surveillance. Part II. Wkly Epidemiol Rec 2001;76:390-6.
4. Amaral JA, Pereira EP, Paixao MT. Data and projections of HIV/AIDS cases in Portugal: An unstoppable epidemic? J Appl Stat 2005;32:127-40.
5. European Centre for Disease Prevention and Control WROfE. HIV/AIDS surveillance in Europe: Surveillance report. Stockholm: ECDC; 2015.
6. Esteves A, Parreira R, Piedade J, et al. Spreading of HIV-1 subtype G and envB/gagG recombinant strains among injecting drug users in Lisbon, Portugal. Aids Res Hum Retrov 2003;19:511-7.
7. Carvalho A, Costa P, Triunfante V, et al. Analysis of a local HIV-1 epidemic in Portugal highlights established transmission of non-B and non-G subtypes. J Clin Microbiol 2015;53:1506-14.
8. Gupta S, Anderson RM, May RM. Networks of sexual contacts: implications for the pattern of spread of HIV. AIDS 1989;3:807-17.
9. Thomson MM, Najera R. Increasing HIV-1 genetic diversity in Europe. J Infect Dis 2007;196:1120-4.
10. Bartolo I, Abecasis AB, Borrego P, et al. Origin and epidemiological history of HIV-1 CRF14_BG. PLoS One 2011;6:e24130.
11. Palma AC, Araujo F, Duque V, et al. Molecular epidemiology and prevalence of drug resistance-associated mutations in newly diagnosed HIV-1 patients in Portugal. Infect Genet Evol 2007;7:391-8.
12. MacArthur GJ, van Velzen E, Palmateer N, et al. Interventions to prevent HIV and Hepatitis C in people who inject drugs: A review of reviews to assess evidence of effectiveness. Int J Drug Policy 2014;25:34-52.
13. Loo MV, Van Beusekom I, Kahan JP. Decriminalization of drug use in Portugal: The development of a policy. Ann Am Acad Polit Ss 2002;582:49-63.
14. Departamento de Doenças Infeciosas do INSA. Infeção VIH SIDA: a situação em Portugal a 31 de dezembro de 2014. Lisboa: Instituto Nacional de Saúde Doutor Ricardo Jorge; 2015.
15. Stadler T, Kouyos R, von Wyl V, et al. Estimating the basic reproductive number from viral sequence data. Mol Biol Evol 2012;29:347-57.
16. Leventhal GE, Gunthard HF, Bonhoeffer S, Stadler T. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Mol Biol Evol 2014;31:6-17.
17. Bezemer D, Cori A, Ratmann O, et al. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLoS Med 2015;12:e1001898; discussion e.
18. Hofstra LM, Sauvageot N, Albert J, et al. Transmission of HIV Drug Resistance and the Predicted Effect on Current First-line Regimens in Europe. Clin Infect Dis 2016;62:655-63.
19. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.

20. Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7:539.
21. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 2016;33:1870-4.
22. Pineda-Pena AC, Faria NR, Imbrechts S, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol 2013;19:337-48.
23. Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res 2014;42:e144.
24. Bennett DE, Camacho RJ, Otelea D, et al. Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. Plos One 2009;4.
25. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009;26:1641-50.
26. Ragonnet-Cronin M, Hodcroft E, Hue S, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics 2013;14:317.
27. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312-3.
28. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2016;2:vew007.
29. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol 2013;30:713-24.
30. Bouckaert R, Heled J, Kuhnert D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 2014;10:e1003537.
31. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 2013;110:228-33.
32. Shapiro B, Drummond AJ, Rambaut A, et al. Rise and fall of the Beringian steppe bison. Science 2004;306:1561-5.
33. Edo-Matas D, Lemey P, Tom JA, et al. Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. Mol Biol Evol 2011;28:1605-16.
34. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56-61.
35. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160-74.
36. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Tracer v1.7. 2018.
37. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infect Genet Evol 2008;8:239-46.
38. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics 2005;21:676-9.

39. Wang TH, Donaldson YK, Brettle RP, Bell JE, Simmonds P. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. J Virol 2001;75:11686-99.
40. Zarate S, Pond SL, Shapshak P, Frost SD. Comparative study of methods for detecting sequence compartmentalization in human immunodeficiency virus type 1. J Virol 2007;81:6643-51.
41. Bezemer D, Faria NR, Hassan A, et al. HIV Type 1 Transmission Networks Among Men Having Sex with Men and Heterosexuals in Kenya. Aids Res Hum Retrov 2014;30:118-26.
42. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol 2008;56:391-412.
43. O'Brien JD, Minin VN, Suchard MA. Learning to Count: Robust Estimates for Labeled Distances between Molecular Sequences. Mol Biol Evol 2009;26:801-14.
44. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol 2009;5:e1000520.
45. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Molecular Biology and Evolution 2016;33:2167-9.
46. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc 1995;90:773-95.
47. Jeffreys H. Theory of probability. 3d ed. Oxford,: Clarendon Press; 1961.
48. Ayres DL, Darling A, Zwickl DJ, et al. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol 2012;61:170-3.
49. Edwards CJ, Suchard MA, Lemey P, et al. Ancient hybridization and an Irish origin for the modern polar bear matriline. Curr Biol 2011;21:1251-8.
50. Fernandes RM, Cary M, Duarte G, et al. Effectiveness of needle and syringe Programmes in people who inject drugs - An overview of systematic reviews. BMC Public Health 2017;17:309.
51. Drug decriminalisation in Portugal: setting the record straight. Transford: Getting drugs under control2014.
52. Esteves A, Parreira R, Piedade J, et al. Spreading of HIV-1 subtype G and envB/gagG recombinant strains among injecting drug users in Lisbon, Portugal. AIDS Res Hum Retroviruses 2003;19:511-7.
53. Patino-Galindo JA, Torres-Puente M, Bracho MA, et al. Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain. PLoS One 2017;12:e0171062.
54. Chaillon A, Essat A, Frange P, et al. Spatiotemporal dynamics of HIV-1 transmission in France (1999-2014) and impact of targeted prevention strategies. Retrovirology 2017;14:15.
55. Parczewski M, Leszczyszyn-Pynka M, Witak-Jedra M, et al. Expanding HIV-1 subtype B transmission networks among men who have sex with men in Poland. PLoS One 2017;12:e0172473.
56. Chalmet K, Staelens D, Blot S, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. BMC Infect Dis 2010;10:262.
57. Chan PA, Hogan JW, Huang A, et al. Phylogenetic Investigation of a Statewide HIV-1 Epidemic Reveals Ongoing and Active Transmission Networks Among Men Who Have Sex With Men. J Acquir Immune Defic Syndr 2015;70:428-35.

58. Hue S, Brown AE, Ragonnet-Cronin M, et al. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. AIDS 2014;28:1967-75.
59. Esbjornsson J, Mild M, Audelin A, et al. HIV-1 transmission between MSM and heterosexuals, and increasing proportions of circulating recombinant forms in the Nordic Countries. Virus Evol 2016;2:vew010.

6 Conclusions

Parts of the "Future directions" section were published in:

Tetyana I Vasylyeva, Samuel R Friedman, Dimitrios Paraskevis, Gkikas Magiorkinis. Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. Infect Genet Evol 2016,46:248-255.

Summary of authorship:

TIV wrote the paper and created all figures. SRF and DP helped to conceptualize the paper, and edited the original draft. GM provided supervisory support.

6.1 Characterizing the HIV epidemic in Eastern Europe

Eastern Europe has the largest HIV epidemic in Europe, which is also the biggest epidemic ever recorded in people who inject drugs (PWID)[1,2]. Several studies have been conducted to understand the behavioral patterns that led to the explosive epidemic observed in Russia and Ukraine in the 1990s[3-5], and to elucidate the factors preventing some epidemics in Eastern European countries from declining[6]. This thesis adds new information about past and current HIV transmission patterns in PWID populations, including those from Ukraine and Russia.

In Chapter 2 I investigated possible explanations for the uncontrolled rapid spread of HIV in Ukraine and Russia in the 1990s. Several hypotheses have been put forward to explain why HIV spread so rapidly in Ukrainian and Russian PWID populations. Apart from risky injecting practices, such as the use of someone else's used needles, drug preparation methods were also investigated, some of which included the use of HIV-infected blood as a buffer[5]. The molecular epidemiology framework presented in Chapter 2 showed that such drug preparation methods were unlikely to have played a role in the spread of the virus in PWID populations. According to this framework, such practices, which would involve a small number of transmitters, would correspond to a very long average infection generation time (or epidemic serial interval), a scenario that is both biologically and clinically implausible. Instead, my findings show that the PWID-driven HIV epidemic in Russia and Ukraine in the 1990s was characterised by a large proportion of transmitters (70%), a very short generation time, T (< 1 month), and a high basic reproductive number, R_0 (averaging 7 new infections per infected individual). This information sheds light on the initial stages of the Eastern European HIV epidemic, and can help to guide public health measures targeting the reduction of HIV transmission in this transmission group.

I continued my investigation of the Eastern European HIV epidemic by characterising recent changes in the virus' spatiotemporal distribution in Ukraine. In Chapter 3 I showed that, starting from 2014 (the year that military conflict began in Ukraine), Donetsk and Lugansk, the two regions in eastern Ukraine heavily affected by HIV (and where the armed conflict has occurred) became the dominant exporters of viral lineages in the country. While it is known that the recent military crisis in eastern Ukraine has disrupted HIV prevention and treatment services[7-9], my analysis provides evidence of the impact of armed conflict on HIV lineage movement. Large-scale human migration, which included the movement of around 1.7 million people to populous cities in the southern and central regions of the country[10], resulted in frequent HIV lineage flow to those regions. I show also that despite significant efforts from non-governmental organizations to maintain the provision of HIV prevention services[11], risky injecting and sexual practices are still prevalent among PWID in regions that were identified as the main recipients of exported virus lineages, creating a perfect storm for HIV re-emergence in these locations.

The analysis presented in Chapter 3 represents a step forward in understanding the effects on HIV of the military conflict in Ukraine in particular, and of so-called "Big Events"[12] in general. The spread of HIV and sexually transmitted diseases during wars and armed conflicts has been studied before, for instance in Africa[13]. "Big Events" such as armed conflicts and political and economical crises create ideal conditions for HIV spread, not only by inducing changes in sexual behaviour, but also through the establishment and growth of drug injecting networks[14]. Because of their disruptive nature, which affects both the delivery of public health services and the ability to carry out research in the area, the impact of "Big Events" in infectious disease transmission is often under-researched, particularly in developing countries[4].

Studying populations such as residents of war-torn regions and PWID can be extremely challenging using traditional epidemiology alone. It is partly because it can be dangerous for both researchers, and often participants themselves, for example if their activity, such as drug use, is illegal. Another reason is that members of these groups can be difficult to locate or to engage as research participants. By combining state-of-the-art genetic and epidemiological analyses, I argue that it is possible to investigate the impact of military conflicts, human migration, and public health HIV reduction campaigns; hence the framework of molecular epidemiology is particularly useful for hard-to-reach populations. Given the current refugee crises in Europe, Africa, and elsewhere[15,16], the development of new frameworks that are based on combined analyses of genetic, social and epidemiological data has the potential to improve public health efforts to curb the HIV/AIDS pandemic, especially in resource-limited settings.

Whilst Chapters 2 and 3 describe different aspects of the HIV-1 subtype A epidemic in Ukraine, Chapter 4 investigates spatial and temporal changes in the molecular epidemiological patterns of HIV-1 subtype B in Ukraine. Subtype B is the most prevalent subtype in Europe and the second most prevalent subtype in Ukraine (see Chapter 3 and [17]). However, little information existed about its distribution in different regions in Ukraine. According to the data available from the Los Alamos HIV sequence database, subtype B is far less prevalent in Ukraine than the dominant subtype A; as of August 2018, subtypes A and B comprise 86% and 10%, respectively, of all HIV sequences sampled from Ukraine[18]. The analysis presented in Chapter 4 is the largest (N=164 sequences) and the most geographically representative (sequences from 11 regions of Ukraine) study of HIV-1 subtype B in the countries of the former Soviet Union. My analysis shows that subtype B was

introduced in Ukraine in the early 1970s, which is a date close to that previously estimated for the introduction of subtype A to the country[19]. Thus, I argue that evolutionary or viral genetic factors are unlikely to explain the predominant spread of subtype A in post-Soviet countries, compared to the more limited epidemiological success of subtype B. Instead this difference is likely linked to social and cultural dissimilarities between the geographic regions where each HIV subtype was first introduced. Furthermore, I show that, in agreement with previous reports[20,21], there are two co-circulating subtype B lineages in Ukraine. I expand on previous reports by showing that both lineages seem to have expanded geographically and can now be found in a wider range of transmission groups. Studies of HIV subtype B in Ukraine have been overshadowed by those of the much more prevalent subtype A. The results presented in Chapter 4 suggest that it is important to monitor changes in HIV diversity in Ukraine, particularly the prevalence of subtype B and other circulating viral lineages, such as circulating recombinant strain CRF03_AB, a poorly understood A/B recombinant discovered in 1998-99 in Russia and Ukraine[22,23].

In summary, this thesis advances our knowledge of the HIV epidemic in Eastern Europe, with implications for other settings where there are emerging epidemics in PWID. My work represents the most comprehensive molecular epidemiology investigation of the Ukrainian HIV epidemic to date. The approaches undertaken here can guide analyses of similar HIV epidemics in other countries, especially those in countries of the former Soviet Union with substantial HIV-positive PWID populations.

6.2 Using phylodynamics to assess prevention strategies for hard-to-reach populations

The application of traditional epidemiological methods to PWID is often complicated by the fact that these populations are often hard to reach and to follow-up. By combining molecular epidemiology and simple compartmental modeling, the results presented in this thesis can be used to suggest which preventive strategies should be selected to control outbreaks in PWID populations. In particular, I show that HIV is transmitted very rapidly in this group and transmission often happens before an HIV-infected injector can be diagnosed and offered treatment. Even when in care, adherence to anti-retroviral treatment (ART) is often much lower for PWID than for other groups, due to a variety of social and medical reasons associated with the nature of drug use[24]. Consequently, two strategies appear to be most effective in this group: providing antiretroviral treatment to recently-infected individuals, and scaling-up primary preventive measures (e.g. harm reduction interventions, including needle exchange programs and methadone treatment; see section 1.3.1)[25]. While treatment of recently infected people is certainly desirable, it is often unachievable before an infected drug injector transmits the virus to a susceptible peer, as, by my estimate, the infection generation time for PWID is only one month on average. Harm reduction can be more effective and easier to implement, since it targets even those who are unaware of their infection status.

In Chapter 5 I investigated temporal fluctuations in HIV transmission dynamics in a largely PWID-dominated HIV epidemic. Here I applied a new phylodynamic approach to estimate major shifts in the epidemic effective reproductive number (R_e) over time and I applied this method to the epidemics of subtype G and subtype B in Portugal. My results show that the timing of a transition from $R_e>1$ to $R_e<1$ coincided

with the introduction of highly active anti-retroviral treatment (HAART) and the scale-up of harm reduction interventions in Portugal, such as needle exchange programs. While it is hard to disentangle the effect of the two different preventive measures, I find that the R_e of subtype G in Portugal started to decline five years earlier than the decline observed for subtype B (which was largely driven by sexual transmission). This suggests a beneficial effect of needle exchange programs for PWID.

These molecular epidemiological findings contribute to the evidence for the effectiveness of harm reduction approaches[26]. Given that conducting randomized controlled trials to assess the effectiveness of harm reduction is impractical, and also unethical[26], methods that allow us to assess the impact of specific campaigns targeted at reducing virus transmission are valuable. This thesis outlines feasible new ways to assess HIV prevention strategies in PWID using molecular epidemiology and phylodynamic methods.

6.3 Accurate choice of model parameters for phylodynamic inference

With the expanding availability of viral genetic data and the increasing popularity of molecular epidemiology analysis, it becomes imperative to choose optimal evolutionary models for hypothesis testing[27,28]. While some coalescent population models, such as the Bayesian Skyline[29], are more common than others, it is important for meaningful inference to have a solid understanding of the underlying assumptions and statistical performance of different models.

The simulation analysis described in Chapter 4 showed that a non-parametric coalescent approach (the Bayesian Skyline model) was slightly more accurate in

estimating the true population size than a parametric model (the constant-exponential-constant model, CEC), even though the simulation was designed to specifically generate CEC-like dynamics. The Bayesian Skyline model was initially developed as a model selection tool in order to identify an appropriate parametric model without conducting the analysis with all possible parametric models[29]. In recent years, however, it has been mostly used to estimate viral population dynamics directly, and Bayesian Skyline model analysis is often not followed by a fitting of a specific parametric model. My results show that for the epidemic scenarios investigated, the two models generally converge to the same estimates and perform similarly with regard to parameter estimation, such as a sampling proportion and the sampling scheme.

At the same time, my analysis of Ukrainian data, and the simulation study in Chapter 4, show that the choice of an appropriate sampling model is crucial for obtaining reliable estimates under the birth-death skyline (BDSKY) model[30]. I found that if one sampling scheme is assumed by the BDSKY model, but another sampling scheme was used to sample the sequences used in the analysis, then the reproductive number (R_e) estimated might not be accurate. More often than not, sampling of genetic sequences for epidemiological analysis is not done randomly, or in a structured manner, but by convenience. I argue that, to enable accurate and compatible results, we should strive whenever possible, to reflect this convenience sampling in prior information and the choice of sampling model when parameterizing phylodynamic inference models.

The identification of phylogenetic clusters can be used to aid the design of optimal intervention strategies to curb HIV infection. For example, analyses of phylogenetic transmission clusters have been used to estimate the contribution of recently infected

people to forward HIV transmission in a population[31] and to show that HIV-drug resistant variants are transmissible and can establish infection in HIV-uninfected people[32]. Moreover, phylodynamic analysis of HIV transmission clusters has become a common practice and is often used to inform strategies for HIV prevention and treatment[31,33]. However, cluster-based phylogenetic analyses are often challenging to interpret[34] and are frequently less informative about trends in HIV epidemic transmission at a country-wide level, compared to subtype-specific phylogenetic analyses[35]. There is a lack of a standard definition of a transmission cluster, which makes comparisons and interpretation of different studies challenging[34,36].

Other challenges in interpreting epidemiological and evolutionary analysis results of transmission clusters arise from the fact that phylogenetic clustering can be driven by non-epidemiological factors, such as a sampling proportion or a sampling scheme, and can be observed in homogenous populations[36], and that sampling density has an effect on phylogenetic clustering[37]. A recent simulation study has shown that analyzing individual transmission clusters may lead to erroneous estimates of population dynamics[38]. To circumvent this, in Chapter 5 I used a new hierarchical model implementation that permits the drawing information of from different phylogenetic clusters without strictly constraining them to follow the same evolutionary process[39]. Another advantage of the model presented in Chapter 5 is its faster computation speed in comparison with approaches considering a single, large phylogeny that assumes the same evolutionary process across different transmission clusters within that tree.

Finally, in Chapter 3 I applied another novel approach to answer questions that could not be addressed by traditional molecular epidemiology methods. While an initial phylogeographic analysis showed that viral lineage movement in Ukraine had a

westward direction, it was not clear whether this pattern occurred during a specific time period and whether it was initiated by the war, or if the east of Ukraine was the main region exporting virus lineages before 2014. In Chapter 3 I showed by model selection analysis that a phylogeographic "epoch model", which permitted a change in the rates of viral lineage migration before and after 2014, outperformed a model that assumed no change in HIV among-location transition rates during this period. The statistical support for the two-epoch model indicated an impact of war-related displacement in the spread of HIV in Ukraine.

6.4 Future directions

6.4.1 Combining social network approach and phylodynamics

Phylogenetic trees reconstructed from genetic sequences contain valuable information about the evolutionary history of the virus that can be used to infer possible scenarios of transmission among hosts. Genetic data has the advantage of containing unbiased biological information, compared, for example, to contact tracing data that are often subject to self-report bias. Reconstructing phylogenetic trees informs us if two sequences evolved from the same genetic strain: if the two sequences are clustered together on a phylogenetic tree, we can say that they have an ancestor in common. Further assigning sampling dates to sequences and implementing molecular clock analyses provides us with an estimate of the timing of the putative transmission events[40]. However, these estimates are a subject to bias, because there is a discrepancy between the timing of phylogenetic tree branching, which refers to the moment when two sampled lineages coalesce, and the timing of among-host transmission events. The strain that is transmitted might have evolved within the donor individual long before the transmission event happened[41,42], and, thus, coalescent events on the tree

can overestimate the time of transmission events. Further, inferring transmission networks from phylogenetic trees can be problematic in many epidemiological settings. For example, in densely sampled HIV outbreaks phylogenetic trees are likely to have low confidence support. This is especially true when the timescale of the transmission events is fast relative to the HIV evolution, which makes it difficult to infer transmission events based on these phylogenetic trees. Another example is when a reconstructed phylogenetic tree has a star-shape shape. Recently, a Bayesian model that takes into account within-host genetic diversity attempted to resolve this issue[43]. In that study, the model was able to reconstruct some parts of the transmission network, including correctly defining the source of infection in a hypothetical population, even though transmission pathways inferred from genetic data were still more ambiguous than those inferred from the detailed epidemiological data. An alternative Bayesian model, which also accounts for within-host virus evolution, and also takes into account the non-random host population structure of the epidemic (instead of assuming population panmixis), was suggested for similar settings where dense sampling is available[44]. It suggests that in a densely sampled outbreak, a well-resolved phylogenetic tree contains transmission pathways and by sampling different sub-sets of the tree and calculating posterior probabilities for the nodes, it is possible to reconstruct a reliable transmission network. The limitation of the later approach is that all of the cases have to be sampled, which is unrealistic for most viral infections, including HIV, mostly because HIV can be asymptomatic for a long time.

It is long well known that HIV spreads non-randomly through networks of closely connected people who engage in injecting or sexual practices (see section 1.3.1). As a consequence, social network analysis has been used extensively in HIV epidemiology as an approach to recruit participants, monitor and predict behavior patterns, and to

model future and past disease spread[45]. Social network-based studies have revealed that the risk of acquiring HIV is associated not only with individuals' behaviors, but also with the individual's position in the network and the behaviors of the individual's peers (sexual/injecting partners)[46]. The network structure might facilitate or limit the spread of the virus (as well as the dissemination of safe behavior messages) within groups. Thus, the social network approach has become an important epidemiological tool in HIV prevention and treatment[45].

As with phylogenetic analysis, social network data from people who share a risk behavior includes information about possible transmission routes. However, using network information alone to resolve transmission pathways is not straightforward: first, self-reported behavioural data can contain biased information (especially with regard to socially unaccepted or illegal activities)[47]; second, network studies will almost never be able to recruit all network members, resulting in incomplete network information[48]. Finally, behavioural data do not necessarily provide evidence about the transmission history of an infectious agent, rather, they contain information about shared risks, which may or may not lead to an actual infection event.

Consequently, combining genetic and social network data in a complex socio-molecular approach might reduce the limitations of each individual approach, and thus improve the way we infer transmission pathways across populations. First, recognizing the network structure of a population can help to advance phylodynamic methodologies. Many phylodynamic methods are based on coalescence models (see Section 1.4.2.1) which assume that host populations are panmictic[49] - that is, that every host has an equal probability to contact and transmit the infection to another individual. This is an unrealistic assumption for real-life epidemics. On the contrary, a network population structure of hosts implies preferential or assortative mixing by a

number of partners (e.g. "highly active people tend to have highly active partners"), often within social groups (e.g. ethnicities) or across sexual groups (on a populational level, men tend to have more sexual contacts with women than men with men or women with women)[50]. This leads to heterogeneity in the number of secondary infections per transmission group (effective reproductive number, R_e), such that some groups will contribute more to new HIV transmissions than others[50]. Moreover, some individuals within these groups, often referred to as "superspreaders", can account for a larger proportion of transmission events in a population, due to their risk behaviours, position in a social network, or for biological reasons, such as high viral load resulting in greater infectiousness[51,52]. Some researchers have found that the variance in the number of secondary infections in a population has an effect on reconstructed pathogen phylogenies, For example, the heterogeneity in R_e among infections has been proposed to result in phylogenies with more smaller-sized clusters and shorter mean branch lengths, compared to phylogenies reconstructed from a homogeneous population[53]. Introducing more complex phylodynamic models that take into account the network structure of a population may allow more accurate estimates of HIV transmission chains, identification of transmission clusters, and estimates of epidemiological parameters for specific clades or transmission clusters (see section 1.3.3).

Social network information can be used to help resolve ambiguous or equally plausible transmission pathways reconstructed from phylogenetic trees, if genetic and social data are obtained from the same individuals. One of the main differences between social and the genetic data is that social network data may, theoretically, include an overall contact network (as illustrated in Fig. 6.1), including non-infected individuals and individuals who were never recruited, but were reported to be part of

the network by other network members. Knowing additional information about other members of the network, whose viral genetic information was not collected, may help us to choose statistically among the large number of multiple plausible transmission pathways inferred from a phylogenetic tree.

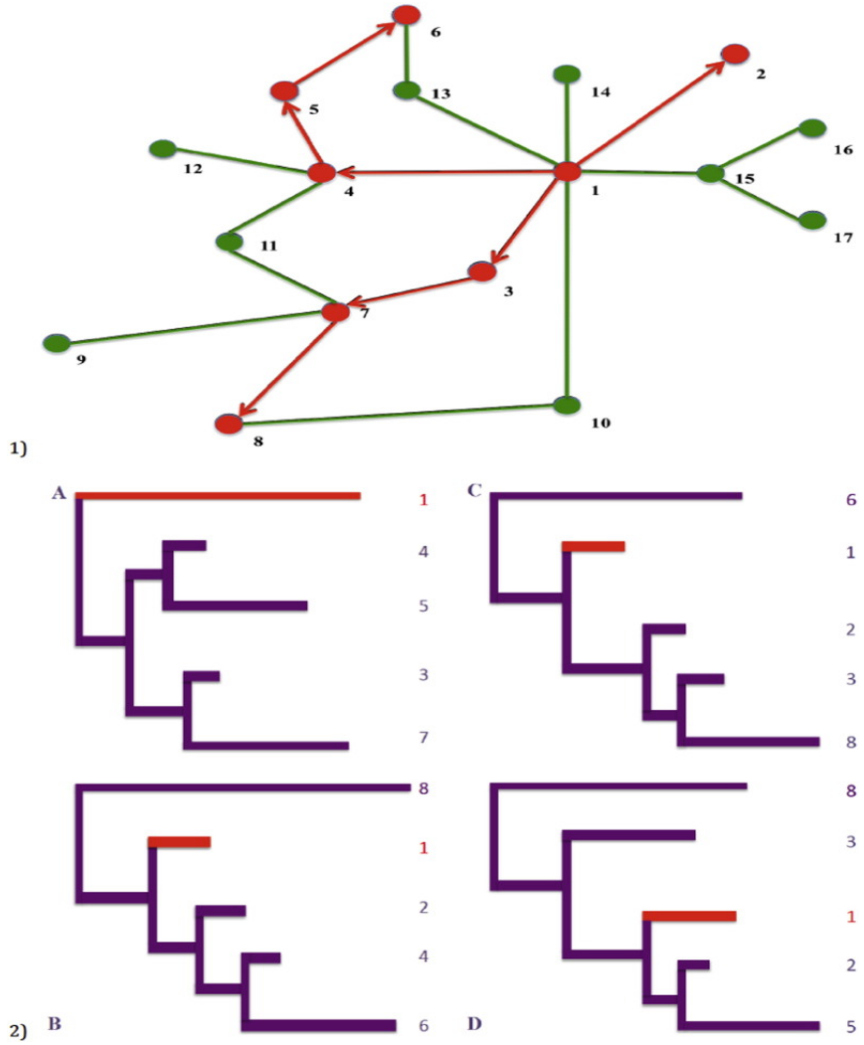


Figure 6.1. 1) A hypothetical contact network of PWID (each with a given number) and known transmission pathways. Red circles represent HIV-infected individuals and green circles represent HIV-free individuals. Red arrows indicate the direction of infection (who infected whom) and green lines indicate an injecting partnership that did not lead to transmission. 2) Phylogenetic trees reconstructed from a subset of PWID from the network above. Panel A represents a phylogenetic tree inferred from

samples collected from individuals 1, 3, 4, 5, 7 (who have higher degree of centrality in the contact network). Panels B, C, and D represent phylogenetic trees reconstructed from samples collected from other sets of individuals who have lower degrees of centrality. Red branches indicate the "seed" of the epidemic within the network.

Further, hypothetically, sampling individuals who have a more central position in a network means that estimated phylogenetic trees are more likely to contain branches that represent *direct* transmission events (i.e. without intermediate, unsampled infections) (Fig. 6.1). Conversely, sampling individuals with less central position in a network may result in phylogenetic trees in which there is a less direct association between tree structure and transmission events among the sampled individuals, making attempts to reconstruct the true transmission pathways more challenging.

The use of estimated times and directions of transmission events provided by social network data and molecular clock analysis has the potential to provide insights into dynamic changes in the structure of a transmission network. Typically, contact data collected in social network studies describe the most recent connections among individuals. In most social network studies, participants are questioned only about a short period of time (i.e. the last few months), because relationships among individuals may change rapidly and, additionally, recall bias is higher with respect to older connections[54]. Phylogenetic trees, on the contrary, can infer past events that occurred when viral lineages were transmitted both among and within infected individuals[44]. For HIV, these events may have occurred many years ago. Thus, combining the two data sources can provide complementary insights about changes in social structure of the population of hosts and, consequently, transmission networks.

When trying to reconstruct possible transmission chains from phylogenetic data, some researchers have relied on genetic distances as a way to identify probable transmission links[55]. Others have used molecular clock analysis to define epidemiological linkage, by estimating the most recent common ancestor of any two nodes within a defined period of time (e.g. less than 3 years)[40]. Some attempts to apply the socio-molecular clock approach to study viral transmission have compared social distances (e.g. the smallest number of individuals connecting two nodes in a contact network) with viral genetic distances between individuals[56,57]. Thus far, there are no commonly accepted or "conventional" methods to jointly analyse genetic and social network data. Several approaches have been suggested, including the simulation of phylogenies along social networks, node-to-node comparisons between real-life social network and viral genetic data, and using social network data to inform phylogenetic analysis, and vice versa[50]. Applying these approaches to real-life data in various epidemiological settings might help to choose amongst this multiplicity of methods and establish a standard approach for the use in infectious disease epidemiology.

There are several possible issues preventing the wider application of social network data to halt HIV transmission. First, the costs associated with collecting network data is high; second, in many settings HIV is still associated with stigma that demotivates participants to participate in studies and/or refer their partners; finally, network surveys often include sensitive questions about sexual and injecting practices, which renders these studies ethically challenging and potentially raises safety issues for field researchers[58-60]. Furthermore, the use of a combination of social and genetic data raises a great concern as it might cause serious ethical matter. If social network information in combination with phylodynamic analysis will indeed help to reconstruct transmission networks and identify the source of infection in a pair or a

group, then this can have serious, potentially criminal, implications for individuals, particularly in the settings where negligent or deliberate transmission is a crime, or where injecting drugs or for men having sex with men is illegal.

Encouragingly, obstacles encountered in previous years, such as computational complexity of network analysis and expensive and time-consuming sequencing, are becoming greatly relaxed due to increased computational power and rapid portable sequencing technologies. The wider use of social and molecular approaches raises further challenging questions. How should we relate social data to molecular data? How reliable are HIV transmission pathways estimated from individual data sources, compared with reconstructions resulting from combined approaches? Can combining the two methods at local geographic scales assist in case finding and significantly improve public health interventions to reduce HIV transmission at country-wide levels? More research is needed to address these questions.

6.4.2 Further use of molecular epidemiology to improve HIV prevention and treatment outcomes in Ukraine

While we now have a much better understanding of the Ukrainian HIV epidemic, there are still questions that are unanswered. As the war is expected to disrupt further the pre-war epidemiological stability, we need to continue to study viral spread and monitor outbreak occurrences. In particular, using the Minion device[61] to generate sequence data in the field in mobile clinics, from people who do not have access to the Ukrainian AIDS clinics, could help to direct efforts to assess the effect of treatment disruption on the levels of drug resistance in the region. Such studies could be important from both public health and clinical perspectives.

In Chapter 3 I suggested that the prevalence of the K65R mutation has increased since the initiation of war. K65R is a signature mutation for tenofovir desoproxil fumarate

(TDF) resistance associated with early virological failure[62]. (TDF)-based regimens are often prescribed in the general HIV-infected population and TDF is the main drug prescribed during pre-exposure prophylaxis (PrEP). PrEP is an effective public health intervention, which was introduced in Ukraine in early 2018, but its effect can be altered by the prevalence of anti-drug resistance mutations associated with common PrEP drugs. Using mathematical models, we could attempt to predict the effect of the existing drug resistance prevalence on the potential for PREP success in Ukraine. However, currently only data from treatment-experienced patients are available in Ukraine. Generating data from treatment-naïve patients, including those in war-torn regions, would help to build accurate predictive models and find ways to maximise the efficiency of PrEP in Ukraine and neighbouring countries.

Furthermore, based on the findings in Chapter 5, I argue that birth-death models have a broader potential for the study of the effect of interventions and preventive measures. One potential further application for phylogenetic birth-death models could be to describe changes in R_e of a local epidemic that might be attributable to a particular intervention. Studying HIV transmission dynamics in locations affected and not affected by the intervention might provide additional evidence for the efficacy of the intervention. We might conceive this research being undertaken within a country, to ensure that all other factors remained the same throughout the time of the intervention implementation, or undertaken in neighbouring countries, that adopt different HIV prevention agendas. An example would be the widespread opioid substitution treatment in Ukraine, versus its complete ban in neighbouring Russia. Conclusions based on molecular epidemiology analyses can have valuable policy implications, if carried out in a timely and reliable manner.

6.5 References

1. UNAIDS. The Gap Report. Geneva, Switzerland: Joint United Nations Programme on HIV/AIDS; 2014.
2. ECDC. HIV/AIDS surveillance in Europe 2015. Stockholm: European Centre for Disease Prevention and Control/WHO Regional Office for Europe; 2016.
3. Dehne KL, Grund JPC, Khodakevich L, Kobyshcha Y. The HIV/AIDS epidemic among drug injectors in Eastern Europe: Patterns, trends and determinants. J Drug Issues 1999;29:729-76.
4. Rhodes T, Simic M. Transition and the HIV risk environment. Brit Med J 2005;331:220-3.
5. Artyuh OR, Balakireva OM, Bochkova LV, et al. Analytical report. Monitoring of behavior of injecting drug users as a component of second generation epidemiological surveillance. Kyiv: International HIV/AIDS Alliance in Ukraine; 2005.
6. Csete J, Human Rights Watch Organization. Lessons not learned : human rights abuses and HIV/AIDS in the Russian Federation. New York: Human Rights Watch; 2004.
7. Owczarzak J, Karelin M, Phillips SD. A view from the frontlines in Slavyansk, Ukraine: HIV prevention, drug treatment, and help for people who use drugs in a conflict zone. Int J Drug Policy 2015;26:6-7.
8. Pizzi M. War in Ukraine threatens to worsen HIV crisis. Aljazeera America2015.
9. Holt E. Health care collapsing amid fighting in east Ukraine. Lancet 2015;385:494.
10. Interbal Displacement Monitoring Centre. Ukraine: Country Information 2015. Accessed in 2017.
11. Duchenko A, Deshko T, Braga M. Crisis management by HIV/AIDS non-governmental organisations in the post-Euromaidan Ukraine led to opening new horizons. Drugs and Alcohol Today 2017:17(3).
12. Friedman SR, Rossi D, Braine N. Theorizing "Big Events" as a potential risk environment for drug use, drug-related harm and HIV epidemic outbreaks. Int J Drug Policy 2009;20:283-91.
13. Mock NB, Duale S, Brown LF, et al. Conflict and HIV: A framework for risk assessment to prevent HIV in conflict-affected settings in Africa. Emerg Themes Epidemiol 2004;1:6.
14. Nikolopoulos GK, Sypsa V, Bonovas S, et al. Big Events in Greece and HIV Infection Among People Who Inject Drugs. Subst Use Misuse 2015;50:825-38.
15. Lori JR, Boyle JS. Forced migration: health and human rights issues among refugee populations. Nurs Outlook 2015;63:68-76.
16. United Nations High Commissioner for Refugees. Global trends report: world at war. Geneva: United Nations Refugee Agency; 2016.
17. Abecasis AB, Wensing AM, Paraskevis D, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 2013;10:7.
18. Kuiken C, Korber B, Shafer RW. HIV sequence databases. AIDS Rev 2003;5:52-61.

19. Diez-Fuertes F, Cabello M, Thomson MM. Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. Infect Genet Evol 2015;33:197-205.
20. Bobkov A KE, Selimova L, Bobkova M, et al. A Sudden Epidemic of HIV Type 1 among Injecting Drug Users in the Former Soviet Union: Identification of Subtype A, Subtype B, and Novel gagA/envB Recombinants. AIDS Res Hum Retroviruses 1998;14.
21. Saad MD, Shcherbinskaya AM, Nadai Y, et al. Molecular epidemiology of HIV Type 1 in Ukraine: birthplace of an epidemic. AIDS Res Hum Retroviruses 2006;22:709-14.
22. Liitsola K, Tashkinova I, Laukkanen T, et al. HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. Aids 1998;12:1907-19.
23. Liitsola K, Holm K, Bobkov A, et al. An AB recombinant and its parental HIV type 1 strains in the area of the former Soviet Union: low requirements for sequence identity in recombination. UNAIDS Virus Isolation Network. AIDS Res Hum Retroviruses 2000;16:1047-53.
24. Feelemyer J, Des Jarlais D, Arasteh K, Uuskula A. Adherence to Antiretroviral Medications Among Persons Who Inject Drugs in Transitional, Low and Middle Income Countries: An International Systematic Review. Aids Behav 2015;19:575-83.
25. Des Jarlais DC. Harm Reduction - a Framework for Incorporating Science into Drug Policy. American journal of public health 1995;85:10-2.
26. Wodak A, McLeod L. The role of harm reduction in controlling HIV among injecting drug users. Aids 2008;22 Suppl 2:S81-92.
27. Worobey M, Gemmel M, Teuwen DE, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 2008;455:661-4.
28. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. Molecular Biology and Evolution 2012;29:2157-67.
29. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 2005;22:1185-92.
30. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci U S A 2013;110:228-33.
31. Brenner B, Wainberg MA, Roger M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. Aids 2013;27:1045-57.
32. Kaye M, Chibo D, Birch C. Phylogenetic investigation of transmission pathways of drug-resistant HIV-1 utilizing pol sequences derived from resistance genotyping. Journal of acquired immune deficiency syndromes 2008;49:9-16.
33. Grabowski MK, Herbeck JT, Poon AFY. Genetic Cluster Analysis for HIV Prevention. Curr HIV/AIDS Rep 2018;15:182-9.
34. Hassan AS, Pybus OG, Sanders EJ, Albert J, Esbjornsson J. Defining HIV-1 transmission clusters based on sequence data. Aids 2017;31:1211-22.
35. Faria NR, Suchard MA, Abecasis A, et al. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. Infect Genet Evol 2012;12:453-60.

36. Frost SD, Pillay D. Understanding drivers of phylogenetic clustering in molecular epidemiological studies of HIV. The Journal of infectious diseases 2015;211:856-8.
37. Novitsky V, Moyo S, Lei Q, DeGruttola V, Essex M. Impact of sampling density on the extent of HIV clustering. AIDS Res Hum Retroviruses 2014;30:1226-35.
38. Dearlove BL, Xiang F, Frost SDW. Biased phylodynamic inferences from analysing clusters of viral sequences. Virus Evol 2017;3:vex020.
39. Edo-Matas D, Lemey P, Tom JA, et al. Impact of CCR5delta32 Host Genetic Background and Disease Progression on HIV-1 Intrahost Evolutionary Processes: Efficient Hypothesis Testing through Hierarchical Phylogenetic Models. Molecular Biology and Evolution 2011;28:1605-16.
40. Leigh Brown AJ, Lycett SJ, Weinert L, et al. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J Infect Dis 2011;204:1463-9.
41. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055-62.
42. Romero-Severson E, Skar H, Bulla I, Albert J, Leitner T. Timing and order of transmission events is not directly reflected in a pathogen phylogeny. Molecular biology and evolution 2014;31:2472-82.
43. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. Molecular biology and evolution 2014;31:1869-79.
44. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. PLoS computational biology 2015;11:e1004613.
45. Latkin CA, Davey-Rothwell MA, Knowlton AR, Alexander KA, Williams CT, Boodram B. Social network approaches to recruitment, HIV prevention, medical care, and medication adherence. Journal of acquired immune deficiency syndromes 2013;63 Suppl 1:S54-8.
46. Christley RM, Pinchbeck GL, Bowers RG, et al. Infection in social networks: using network analysis to identify high-risk individuals. American journal of epidemiology 2005;162:1024-31.
47. Johnson T, Fendrich M. Modeling sources of self report bias in a survey of drug use epidemiology. Ann Epidemiol 2005;15:381-9.
48. Janicik GA, Larrick RP. Social network schemas and the learning of incomplete networks. J Pers Soc Psychol 2005;88:348-64.
49. Kingman JFC. The coalescent. Stochastic Processes and their Applications 1982;13:235-48.
50. Goodreau SM. Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation. Genetics 2006;172:2033-45.
51. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.
52. Stein RA. Super-spreaders in infectious diseases. Int J Infect Dis 2011;15:e510-3.
53. Robinson K, Fyson N, Cohen T, Fraser C, Colijn C. How the Dynamics and Structure of Sexual Contact Networks Shape Pathogen Phylogenies. Plos Comput Biol 2013;9.

54. Bell DC, Belli-McQueen B, Haider A. Partner Naming and Forgetting: Recall of Network Members. Social networks 2007;29:279-99.
55. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 2008;5:e50.
56. Kostaki EG, Nikolopoulos GK, Pavlitina E, et al. Molecular Analysis of Human Immunodeficiency Virus Type 1 (HIV-1)-Infected Individuals in a Network-Based Intervention (Transmission Reduction Intervention Project): Phylogenetics Identify HIV-1-Infected Individuals With Social Links. The Journal of infectious diseases 2018;218:707-15.
57. Wertheim JO, Kosakovsky Pond SL, Forgione LA, et al. Social and Genetic Networks of HIV-1 Transmission in New York City. PLoS Pathog 2017;13:e1006000.
58. Nikolopoulos GK, Pavlitina E, Muth SQ, et al. A network intervention that locates and intervenes with recently HIV-infected persons: The Transmission Reduction Intervention Project (TRIP). Sci Rep-Uk 2016;6.
59. Smyrnov P, Williams LD, Korobchuk A, et al. Risk network approaches to locating undiagnosed HIV cases in Odessa, Ukraine. J Int Aids Soc 2018;21.
60. Vasylyeva TI, Friedman SR, Smyrnov P, Bondarenko K. A new approach to prevent HIV transmission: Project Protect intervention for recently infected individuals. AIDS Care 2015;27:223-8.
61. Laver T, Harrison J, O'Neill PA, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif 2015;3:1-8.
62. Ruane PJ, Luber AD. K65R-associated virologic failure in HIV-infected patients receiving tenofovir-containing triple nucleoside/nucleotide reverse transcriptase inhibitor regimens. MedGenMed 2004;6:31.

Appendix 1

(Chapter 2)

Dataset A

97CDKTB48, FJ822055, FJ822064, FJ822065, EU345595, EU345596, FJ822066, FJ822067, FJ822068, FJ822069, FJ822070, FJ822071, FJ822072, FJ822073, FJ822056, FJ822074, FJ822076, FJ822077, FJ822078, FJ822080, FJ822082, FJ822083, FJ822057, FJ822084, FJ822085, FJ822086, FJ822087, FJ822090, FJ822091, FJ822092, FJ822093, FJ822058, FJ822094, FJ822095, FJ822096, FJ822097, FJ822098, FJ822100, FJ822101, FJ822102, FJ822059, FJ822104, FJ822105, FJ822106, FJ822107, FJ822108, FJ822109, EU345598, FJ822110, FJ822111, FJ822113, FJ822060, FJ822115, FJ822116, FJ822117, FJ822118, FJ822119, FJ822120, FJ822121, FJ822122, FJ822123, FJ822061, FJ822124, FJ822125, FJ822126, FJ822127, FJ822129, FJ822130, FJ822131, FJ822132, FJ822133, FJ822062, FJ822134, EU345599, EU345600, EU345601, FJ822063, EU345602, EU345603, EU345604, EU345606, EU345841, EU345842, EU345844, GQ867609, GQ867624, GQ867605, GQ867628, GQ867630, GQ867635, GQ867618, GQ867616, GQ867632, GQ867633, GQ867620, GQ867627, GQ867608, GQ867629, GQ867631, GQ867606, GQ867607, GQ867614, GQ867612, GQ867604, GQ867636, GQ867617, GQ867626, GQ867619, GQ867615, GQ867625, GQ867623, GQ867621, GQ867597, GQ867610, GQ867611, GQ867599, GQ867634, GQ867622, GQ867603, GQ867602, GQ867601, GQ867598, GQ867600, EU345614, EU345615, EU345616, EU345617, EU345618, EU345619, EU345620, EU345621, EU345622, EU345623, EU345624, EU345625, EU345626, EU345627, EU345628, EU345629, EU345630, EU345631, EU345632, EU345633, EU345634, EU345635, EU345636, EU345637, EU345638, EU345639, EU345640, EU345641, EU345642, EU345643, EU345644, EU345645, EU345646, EU345647, EU345648, EU345649, EU345650, EU345651, EU345652, EU345653, EU345654, EU345655, EU345656, EU345657, EU345658, EU345659, EU345660, EU345661, EU345662, EU345663, EU345664, EU345665, EU345666, EU345667, EU345668, EU345669, EU345670, EU345671, EU345673, EU345674, EU345675, EU345676, EU345677, EU345678, EU345679, EU345680, EU345681, EU345682, EU345683, EU345684, EU345685, EU345686, EU345687, EU345688, EU345689, EU345690, EU345691, EU345692, EU345693, EU345694, EU345700, EU345701, EU345847, EU345702, EU345703, EU345704, EU345705, EU345706, EU345848, EU345707, EU345708, EU345709, EU345711, EU345713, EU345714, EU345715, EU345716, EU345718, EU345719, EU345721, EU345722, EU345852, EU345723, EU345725, EU345726, EU345727, EU345728, EU345729, EU345730, EU345731, EU345732, EU345733, EU345734, EU345735, EU345736, EU345737, EU345738, EU345739, EU345740, EU345741, EU345743, EU345744, EU345745, EU345746, EU345747, EU345748, EU345749, EU345750, EU345751, EU345752, EU345753, EU345754, EU345755, EU345756, EU345758, EU345759, EU345760, EU345761, EU345762, EU345763, EU345764, EU345765, EU345766, EU345767, EU345770, EU345771, EU345773, FJ009713, FJ009714, FJ009715, FJ009716, FJ009717, FJ009718, FJ009719, FJ009720, FJ009721, FJ009722, FJ009723, FJ009724, FJ009725, FJ009726, FJ009727, FJ009728, FJ009729, FJ009730, FJ009731, FJ009732, FJ009733, FJ009734, FJ009735, FJ009736, FJ009737, FJ009738, FJ009739, FJ009740, FJ009741, FJ009742, FJ009743, FJ009744, FJ009745, FJ009746, FJ009747, FJ009748, FJ009749, FJ009750, FJ009751, FJ009752, FJ009753, FJ009754, FJ009755, FJ009756, FJ009757, FJ009759, FJ009760, FJ009762, FJ009763, FJ009764, FJ009765, FJ009766, FJ009767, FJ009768, FJ009770, FJ009772, FJ009773, FJ009758, EU345780, EU345781, EU345784, EU345785, EU345854, EU345786, EU345855, EU345787, EU345858, EU345788, EU345859, EU345789, FJ009775, FJ009777, FJ009778, EU345791, EU345866, EU345867, EU345868, EU345869, EU345870, EU345792, EU345871, EU345872, EU345873, EU345874, EU345875, EU345876, EU345794, EU345877, EU345878, EU345879, EU345880, EU345795, EU345881, EU345882, EU345796, EU345883, EU345797, EU345884, EU345885, EU345886, EU345887, EU345888, EU345889, EU345890, EU345891, EU345892, EU345893, EU345894, EU345799, EU345800, EU345895, EU345896, EU345801, EU345897, EU345898, EU345899, EU345802, EU345900, EU345901, EU345803, EU345902, EU345903, EU345904, EU345905, EU345921, EU345806, EU345924, EU345925, EU345926, EF545110, EU345936, EU345940, EU345943, FJ009783, FJ009784, FJ009786, FJ009789, FJ009790, FJ009791, FJ009792, FJ009793, FJ009794, FJ009795, FJ009796, FJ009797, FJ009798, FJ009799, FJ009800, FJ009801, FJ009802, FJ009803, FJ009804, FJ009805, FJ009806, FJ009807, FJ009808, FJ009809, FJ009811, FJ009812, FJ009813, EU345807, EU345808,

EU345809, EU345810, EU345811, EU345812, EU345813, EU345814, EU345815, EU345816, EU345817, EU345818, EU345819, EU345820, EU345821, EU345822, EU345823, EU345824, EU345825, EU345826, EU345827, EU345828, EU345829, EU345830, SP811s2004, 00051s2000, SP812s2004, 02KZYUZ300408, 02KZYUZ300412, 02KZYUZ300413, 02KZYUZ300414, 02KZYUZ300415, 02KZYUZ300418, 02KZYUZ300420, 02KZYUZ300421, 02KZYUZ300425, 02KZYUZ300426, 02KZKAR300435, 02KZKAR300438, 02KZKAR300439, 02KZKAR300441, 02KZKAR300442, 02KZKAR300443, 02KZKAR300444, 02KZKAR300445, 02KZKAR300446, 02KZKAR300449, 02KZKAR300451, 02KZKAR300453, 02KZKAR300454, 02KZKAR300455, 02KZKAR300457, 02KZKAR300459, 02KZKAR300460, 02KZKAR300462, 02KZKAR300464, 02KZKAR300465, 02KZKAR300466, 02KZKAR300467, 02KZPAV300480, 02KZPAV300484, 02KZPAV300487, 02KZPAV300493, 02KZPAV300497, 02KZPAV300502, 03KZALM300210, 03KZ300215, 03KZALM300216, 03KZALM300219, 03KZALM300220, 03KZALM300230, 03KZMAN300232, 03KZ300235, 03KZ300243, 03KZVOS300245, 03KZVOS300246, 03KZVOS300248, 03KZVOS300249, 03KZVOS300254, 03KZVOS300255, 03KZVOS300256, 03KZVOS300259, 03KZVOS300260, 03KZVOS300261, 03KZVOS300262, 03KZVOS300263, 03KZVOS300264, 03KZVOS300265, 03KZVOS300267, 03KZYUZ300273, 03KZYUZ300278, 03KZYUZ300279, 03KZYUZ300280, 03KZYUZ300281, 03KZYUZ300282, 03KZYUZ300283, 03KZYUZ300284, 03KZYUZ300285, 03KZYUZ300287, 03KZYUZ300288, 03KZYUZ300290, 03KZYUZ300291, 03KZYUZ300292, 03KZYUZ300293, 03KZYUZ300296, 03KZYUZ300300, 03KZYUZ300302, 01UADN110, 01UADN113, 01UADN114, 01UADN116, 01UADN119, 01UADN121, 01UADN127, 01UADN128, 01UADN129, 01UADN136, 01UADN137, 01UADN139, 01UADN141, 01UADN145, 01UADN147, 01UADN149, 01UADN180, 01UADN216, 01UADN220, 01UADN222, 01UADN224, 01UADN228, 01UAKV099, 01UAKV108, 01UAKV172, 01UAKV178, 01UAKV248, 01UAKV254, 01UAKV269, 01UAOD008, 01UAOD010, 01UAOD025, 01UAOD027, 01UAOD028, 01UAOD035, 01UAOD040, 01UAOD041, 01UAOD049, 01UAOD051, 01UAOD069, 01UAOD070, 01UAOD072, 01UAOD089, 01UAOD095, 01UAOD096, 01UAOD098, 01UAOD102, 01UAOD103, 01UAPol279, 01UAPol286, 01UAPol288, 01UAPol289, 01UAPol290, 01UAPol291, 01UAPol293, 01UAPol295, 01UAPol296, 01UAPol297, 01UAPol298, 01UAPol299, 01UAPol300, 01UAPol302, 01UAPol303, 01UAPol304, 01UAPol310, 01UAPol311, 01UAPol313, 01UAPol314, 02UACR001, 02UACR002, 02UACR003, 02UACR004, 02UACR005, 02UACR006, 02UACR007, 02UACR008, 02UACR009, 02UACR010, 02UACR011, 02UACR012, 02UACR014, 02UACR015, 02UACR016, 02UACR018, 02UACR020, 02UACR021, 02UAKV005, 02UAKV107, 02UAKV114, 02UAKV115, 02UAKV117, 02UAKV121, 02UAKV131, 02UAKV145, 02UAKV149, 02UAKV150, 02UANK022, 02UAKV178, 02UAKV187, 02UAKV180, 02UAKV179, 02UAKV176, 02UAKV156, 02UAKV152, 02UAKV175, 02UAKV182, 01UAOD10, 01UAOD35, 01UAOD89, 02UZ816, 02UZ817, 02UZ818, 02UZ824, 02UZ825, 02UZ827, 02UZ831, 02UZ833, 02UZ834, 02UZ835, 02UZ837, 02UZ838, 02UZ840, 02UZ841, 02UZ842, 02UZ843, 02UZ844, 02UZ847, 02UZ848, 03UZ001, 03UZ002, 03UZ003, 03UZ004, 03UZ005, 03UZ006, 03UZ007, 03UZ010, 03UZ012, 03UZ013, 03UZ015, 03UZ016, 03UZ018, 03UZ019, 03UZ020, 03UZ021, 03UZ022, 03UZ023, 03UZ029, 03UZ030, 03UZ033, 03UZ034, 03UZ036, 03UZ038, 03UZ039, 03UZ041, 03UZ043, 03UZ044, 03UZ046, 03UZ047, 03UZ049, 03UZ051, 03UZ057, 03UZ068, 03UZ069, 03UZ072, 03UZ073, 03UZ076, 02UZ652, 02UZ654, 02UZ659, 02UZ660, 02UZ661, 02UZ663, 02UZ664, 02UZ665, 02UZ667, 02UZ668, 02UZ669, 02UZ670, 02UZ671, 02UZ672, 02UZ676, 02UZ679, 02UZ680, 02UZ687, 02UZ691, 02UZ693, 02UZ695, 02UZ696, 02UZ698, 02UZ702, 02UZ703, 02UZ707, 02UZ771, 02UZ708, 02UZ713, 02UZ723, 02UZ724, 02UZ725, 02UZ726, 02UZ727, 02UZ729, 02UZ730, 02UZ732, 02UZ733, 02UZ734, 02UZ737, 02UZ740, 02UZ741, 02UZ742, 02UZ744, 02UZ745, 02UZ747, 02UZ750, 02UZ752, 02UZ754, 02UZ755, 02UZ756, 02UZ757, 02UZ761, 02UZ762, 02UZ773, 02UZ774, 02UZ775, 02UZ776, 02UZ778, 02UZ784, 02UZ801, 02UZ803, 02UZ807, 02UZ814, GQ477451, GQ477450, GQ477449, GQ477448, GQ477447, GQ477446, GQ477445, GQ477444, GQ477443, GQ477442, GQ477441, EF158043, EF158042, EF158041, EF158040, 01AZ001, 01AZ003, 01AZ004, 01AZ006, 01AZ009, 01AZ014, 01AZ025, 01AZ026, 01AZ030, 01AZ031, 01AZ036, 01AZ038, 02AZ005, 02AZ007, 02AZ015, 02AZ018, 02AZ020, 02AZ023, 02AZ024, 02AZ025, 02AZ028, 02AZ034, 02AZ037, 02AZ038, 02AZ050, 02AZ054, 02AZ058, 02AZ060, 02AZ062, 02AZ066, 02AZ069, 02AZ077, 02AZ080, 02AZ083, 02AZ089, 02AZ093, 02AZ098, 98GEMZ005, 98GEMZ007, 98GEMZ008, 99GEMZ011, 99GEMZ011(2), 99GEMZ012, 99GEMZ013, 03GEMZ002, 03GEMZ003, 03GEMZ007, 03GEMZ008, 03GEMZ009, 03GEMZ011, 03GEMZ012, 03GEMZ014, 03GEMZ016, 03GEMZ017, 03GEMZ020, 03GEMZ021, 03GEMZ022, 03GEMZ024, 03GEMZ027, 03GEMZ030, 03GEMZ031, 03GEMZ032, 03GEMZ034, 03GEMZ035, 03GEMZ036, 03GEMZ038, 03GEMZ040, GU945074, GU945075, GU945077, GU945079, GU945080, GU945082, GU945090, CB_155, CB_156, CB_599,

GU945091, GU945092, GU945096, GU945101, GU945102, GU945106, GU945107, GU945108, GU945109, GU945113, GU945127, GU945132, GU945140, GU945141, GU945142, GU945143, GU945160, GU945161, GU945162, GU945163, GU945164, GU945165, GU945166, GU945169, GU945170, GU945173, GU945174, GU945175, GU945189, GU945190, GU945194, 44, UG273, 65, 164, 180, 401, 419, 429, 440, 441, 453, 496, 503, 542, 574, 584, 599, 604, 608, 622, 623, 646, 649, 652, 653, 660, 687, 702, 704, 718NT6W, 734, 742, 792, 807, 841, 13, 173, 185, 444, 879, 522, 532, 534, 545, 577, 592, 699, 710, 790, 847, 871, UG275, 96CM_MP512, 97CM_MP582, 97CM_MP640, 98CM_MP1019, 98CM_MP1094, 98CM_MP1173, 00GAB_LBV02_08, 00GAB_LBV18_08, 00GAB_01B, 00GAB_03, 00GAB_04B, 00GAB_14M, 00GAB_18N, 00SN_69HPD_4_9_00, 97CD_KCC2, 97CD_KTB13, 02CD_KTB035, 02CD_KS035, 02CD_LBS024, 02CD_KSS013, 02CD_KIST061, 02BU_U1808, 02BU_U2226, 02BU_U2929, 02BU_U3121, 02BU_U0706, 02BU_U1524, 02BU_U1617, 02BU_U1803, 97CM_MP812, 99CM_MP1370, 99CM_MP1433, 99CM_MP1411, DDI579, DDJ360, DDJ369, 02_1354KP, 03_9073NR, 01_300NG, 03_9146NM, 01_1378MJ, 03_9410NS, 03_9366WG, 03_9412NS, 03_9538NG, 03_9456KM, 999_7NM, 04_0567NR, TC025104, cm1657, cm1672, cm2bb, cm8bb, cm10bb, cm161bb, BKO_1037, SGO_08, SGO_27, SGO_47, SGO_71, SGO_86, ARV26, CHU12, HXB2-LAI-IIIB-BRU, 445_124360B, UG031, UG029, 92RW008, 92UG037, ETH_G_079, 92RW025A, 93RW037A, UG275, Q23_17, 94RW13, 97TZ02_AF361872, 97TZ03_AF361873, 113913U, 115106J, 123208J, 123793K, 123898B, 125636U, 126533T, 130526D, 132074M, 132318B, 132591A, 132685M, 134248P, 136677C, 137499Q, 138162O, 138430V, 138904N, 139177O, 140727Y, 141601A, 142181Q, 144449S, 145027T, 145484O, 146922V, 147521R, 148838K, 150014T, 151848F, KER2008, KER2009, KER2012, KNH1088, KNH1135, KNH1144, KNH1199, KNH1207, KNH1209, KNH1211, KSM4021, KSM4024, KSM4030, MSA4069, MSA4070, MSA4072, MSA4076, MSA4079, NKU3005, 99UGA07072, 99UGG03379, 98UG57134, 98UG57135, 98UG57136, 98UG57142, ML170_1986, 98SN_15HALD, 98SN-69HPD, 99SN_61HPD_15_10_99, 98SN_5hald_20_7_98, CK2_75, CK22_50, MKK248, SU478, SU238, 119758j_185, 124727i_164, 97CIRMF21, 22, 54, A173, A341, ML013_10, ML013_2, ML605_3, ML752, 667, 850, 842, 22, 45, 51, 185, 164, 180, 401, 429, 453, 485, 545, 584, 534, 532, 604, 623, 599, 596, 667, 652, 704, 687, 801, 742, 790, 807, 816, 841, 847, 850, 879, 871, 14, 4, 31, 46, 34, 32, 57, 85, 69, 58, 157, 97, 192, 225, 175, 162, 252, 204, 209, 231, 199, 228, 221, 253, 408, 421, 407, 460, 439, 446, 415, 451, 403, 508, 437, 458, 488, 514, 445, 427, 487, 520, 551, 576, 560, 589, 651, 562, 572, 582, 573, 635, 614, 619, 659, 661, 662, 697, 701, 686, 679, 684, 580, 672, 674, 690, 725, 737, 731, 507, 761, 754, 757, 744, 771, 726, 776, 798, 820, 821, 838, 845, 811, 819, 866, 839, 877, 861, 869, 881, 882, A_122_335, A_140_223, A_161_287, A_194_793, A_225_706, A_264_615, A_264_643, A_326_642, A_ABB08, A_ABB102, A_ABB165, A_ABB213, 97KE104, 97KE173, 98KE178, 99KE432, 96KE449, 00KE561, 99KE579, 02ZA_DR17, 93RW_024, 92UG_029, 111186L, 111740R, 111976H, 112979I, 114892W, 121635E, 121766I, 121874N, 122286N, 122530R, 122742D, 122744X, 123114S, 123768V, 126151S, 126721B, 126951L, 128124P, 128315G, 128375T, 129324I, 132590N, 133395E, 133592I, 133725L, 135274M, 136075T, 136579G, 136668K, 137067Y, 138829X, 139277H, 139330R, 141070V, 141906R, 142114X, 142425V, 144857P, 144963H, 145058J, 146342V, 147272W, 147418O, 147572M, 148095P, 148639N, 148882I, 149254I, 149594A, 150205H, KER2008_patent, TC029004, TC026104, TC025204, TC024604, TC023504, TC022504, TC022404, TC022204, TC021904, TC021704, TC021404, TC021204, TC020504, TC011303, TC010603, TC009903, TC009003, TC008903, TC007803, QC890, QB424, QB596, QB670, 04ZASK162B1, CM_01_133, CM_01_568, P0034_090205, P0054_020305, P0060_070305, P0079_160305, P0110_120405, P0185_130605, P0218_130705, P0289_140905, MBA1021_170805, MBA1025_180805, MBA1028_230805, MBA1037_060905, MBA1053_201005, MBAR005_120405, MBAR007_120405, MBAR016_020505, MBAR017_030505, 004_106665H, 014_108422Z, 022_109985S, 031_106897U, 032_109180U, 034_109734M, 044_110390I, 045_109688N, 046_110300R, 051_110961Y, 057_110308G, 097_117239C, 180_118080i, 199_114660F, 225_119778C, 401_120082s, 403_123871S, 407_124279Q, 408_125049Q, 421_125042G, 427_125290E, 429_125330t, 439_126157P, 440_126149o, 446_126268V, 458_126651R, 466_124362l, 476_126649x, 485_127960U, 487_126068K, 496_127964E, 503_127844l, 508_128740N, 514_128712J, 522_128205c, 534_129066r, 542_129863t, 551_129563G, 560_129773M, 562_128605K, 572_131039B, 574_130458m, 576_131187D, 577_131301I, 580_130747L, 582_130283Q, 584_131489e, 592_129775v, 608_131002r, 614_133371R, 622_133749h, 623_134023j, 635_134512Y, 638_134023j, 646_133038m, 649_134513d, 651_135200J, 652_133291l, 653_134024g, 661_135234B, 662_135975N, 667_136472S, 672_135604H, 674_137072A, 679_137708Q, 684_134386Y, 686_137213K, 690_138008N, 697_138073B, 699_136283t, 701_138581J, 702_136536A, 704_138427k, 725_138071W, 726_138858J, 731_138858J,

734_140160a, 737_140207H, 742_139543x, 744_140410S, 754_140838E, 757_139187G, 141262J, 771_140407R, 776_141841G, 792_142856n, 798_141839U, 801_143188S, 802_142555O, 805_143343E, 807_141911f, 811_145165P, 819_144897T, 820_144896A, 821_144894D, 838_145431G, 839_146276F, 841_146190g, 842_144505j, 850_145864u, 861_146980X, 866_147223K, 869_147440T, 871_146977m, 876_147573X, 877_147575W, 879_147225s, 881_147226R, 117948S, 129779W, 133750R, 139830I, ML752, ML1901, ML1945, ML1990, ML2014, ML1901PCR2, ML1990PCR, GH_007, GH_142, GH_182, GH_243, GH_244, TBK002, TBK004, TBK005, TBK011, TBK017, TBK018, TBK023, TBK024, TBK037, TBK039, TBK040, TBK045, TBK048, TBK056, TBK060, TBK064, TBK067, TBK090, TBK095, TBK099, TMS001, TMS002, TMS006, TMS007, TMS011, TMS016, TMS018, TMS025, TMS103, TMS108, TMS115, TMS121, TMS123, TMS124, TMS127, TMS131, TMS134, TMS137, TMS141, TMS142, TMS147, TMS206, TMS218, TMS301, TMS305, TMS306, TMS314, TMS319, TMS320, TMS336, TMS337, TMS341, TMS343, IAEAM003, IAEAM004, IAEAM007, IAEAM012, IAEAM013, IAEAM014, IAEM017, IAEAM019, IAEM022, IAEM028, IAEM031, IAEM033, IAEM038, IAEM039, IAEM041, IAEM043, IAEM045, IAEM049, IAEM050, IAEM054, IAEM056, IAEM060, IAEM065, IAEM067, IAEM078, BKO_902, BKO_1023, BKO_1047, SGO_49, 335346R, 341741Y, 349844H, 361358A, 362363H, 363290D, 326163_M, 365770P, 366975Y, 367927_J, 369096E, 377986_M, 380521_H, 387519_L, 390900V, 394901B, 400177E, 400196E, 409924L, 413282X, 413433L, 415189_B, 417659J, 430066A, 460471T, 476384H, 289005, SK162, 07_156895, 07_156927, 07_156928, 07_156929, 07_156944, 07_156948, 07_156956, 07_156961, 07_156967, 07_156976, 07_156979, 07_156997, 07_156998, 07_157001, 08_101504, 08_101507, 08_101510, 08_101518, 08_101521, 08_101544, 08_101577, 08_102864, 08_102868, 08_102902, TR178, TR185, TR257, 06KECst_006, 06KECst_007, 06KECst_021, 06KECst_019, 06KECst_028, 06KECst_009, 06KECst_005, 06KECst_020, 06KECst_016, 06KECst_022, 06KECst_013, 06KECst_025, 06KECst_001, 06KECst_017, TV239, TV314, Cameroon_20, Cameroon_55, ZZP625, ZZP628, ZZP649, ZZP671, ZZP877, ZZP963, ZZQ675, ZZQ946, ZZR40, ZZQ955, ZZQ956, ZZR160, ZZR475, 06TGHT045, 06TGHT055, 06TGHT077, 06TGHT085, 06TGHT126, 08GQ182, 08GQ508, 08GQ281, 114HALD, 337HALD, 61HPD, 935hald, 1700_hald, A1_07CD1246, A1_07CD0509, A1_07CD0444, A1_07CD3665, A_07CD3012, A_07CD3729, URF_A_CRF22_07CD3407, A_07CD0050, A_07CD2110, A_07CD4292, A_07CD1425, A_07CD1525, URF_A_A3_A_07CD2084, A_07CD3373, A1_07CD3784, A_07CD0335, A_07CD0040, A1_07CD2974, A_07CD2657, A_07CD3537, A_07CD3191, A_07CD4078, 13020D_P1, 13228D_P1, 13240D_P1, 13263D_P1, 13305D_P1, 13354D_P1, 13929D_P1, 14006D_P1, 14275D_P1, 14324D_P1, 14411D_P1, 14493D_P1, 14888D_P1, 15017D_P1, 15460D_P1, 15545D_P1, 15710D_P1, 15712D_P1, 15911D_P1, 16079D_P1, 16182D_P1, 16455D_P1, 16712D_P1, 17219D_P1, DT0016_05_P1, DT0971_05_P1, DT1515_04_P1, DT1745_05_P1, DT2172_05_P1, DT2373_05_P1, DT2794_05_P1, DT2829_05_P1, DT2986_05_P1, DT3165_05_P1, DT3282_05_P1, DT3441_05_P1, DT3482_05_P1, DT3513_05_P1, DT3516_05_P1, DT3736_05_P1, DT3793_05_P1, DT3994_05_P1, DT4014_05_P1, DT3340_05_P1, 301V03000895, 301V03000981BM, 301V03000985, 301V04000057BM, 301V04000184, 301V04000665BM, 301V04000684, 301V04000864, 301V04000900BM, 301V04000911BM, 301V04002129, 301V04002157BM, 301V04002594, 301V04002668BM, 301V04003339BM, 301V04003358, 301V04003475BM, 301V04003479, 301V04003733, 301V04003895, 301V04003959BM, 301V04004030BM, 301V04004298, 301V04004337BM, 301V04004694, 301V04004804BM, 301V04005655, 301V04005698BM, 301V04005942, 301V04005975BM, CYV75, CYR20, CYV085, CYR028, CYN851, CYN732, CDL80, SRCDD0242, HG_0857, HG_0943, HG_0130, HU_0017, SK_0004, CK1_1, CK1_20, CK1_46, CK1_45, CK1_40, CK1_25, CK1_47, VP_LT_111, VP_SW_2, VP_SW_1, VP_SW_11, ORAV279, ORAV280, ORAV271, 07ZA17579, CM04_073, CM04_094, CM04_096, CM04_064, SN756HALD, SN826HALD, SN2720HALD, 320024_A, 320208_V, 322176_A, 337194_A, 337314_Y, 337373_S, 338325_D, 341569_X, 341815_K, 348121_H, 350052_R, 384132_R, 436546_P, 436547_D, 439208_K, 439552_J, 440672_L, 317296_E, 318244_S, 318883_Z, 319763_R, 320477_T, 321288_A, 321541_D, 326455_J, 328958_W, 332768_D, 337081_S, 338135_K, 338796_K, 341279_H, 362487_R, 370903_R, 371850_K, 385463_H, 385494_T, 393958_V, 404615_W, 413614_D, 427986_C, 430253_Z, 436063_G, 393250_S, 393405_H, 416275_M, 425113_S, 431730_B, 519810_R, 521684_A, 529392_J, 482043_M, 503727_G, 504080_C, 509635_D, 516673_E, CP076300084, CP076300141, CP076300688, 1_0212_KiBS_V7, 1_066_KiBS_V5, 1_0517_KiBS_V5, 1_0360_KiBS_V3, 1_0472_KiBS_V5, 1_0457_KiBS_V3, 1_0496_6_KiBS_V3, 0_0158_KiBS_V16, 0_0472_KiBS_V16, 1_0437_KiBS_V5, 1_0357_KiBS_V5, 1_0289_KiBS_V5, TZ_10_002026_A1, TZ_08_014497_A1, TZ_09_032657_A1, TZ_09_006010_A1, TZ_09_032639_A1, TZ_10_003313_A1, TZ_09_001336_A1, TZ_08_017180_A1,

TZ_08_022112_A1, TZ_08_014496_A1, TZ_08_023282_A1, TZ_08_017177_A1, TZ_08_017190_A1, TZ_10_002029_A1, TZ_08_023280_A1, TZ_08_017183_A1, TZ_10_003320_A1, TZ_10_003309_A1, TZ_09_032646_A1, TZ_09_006021_A1, TZ_10_003314_A1, TZ_08_017174_A1, TZ_08_023285_A1, TZ_09_006015_A1, TZ_08_017187_A1, TZ_08_017178_A1, TZ_09_032655_A1, TZ_09_001326_A1, TZ_08_029745_A1, C435S_J0, C455S_J0, D344T_J0, D375T_J0, NTCPS729, AICPS434, AICPS399, NTCPS726, NTCPS724, NTCPS720, NTCPS719, NTCPS731, NTCPS717, NTCPS716, AICPS335, AICPS107, AICPS435, AICPS438, AICPS495, AICPS432, NTCPS733, AICPS157, AICPS336, AICPS303, AICPS283, AICPS225, AICPS216, AICPS180, NTCPS709, NTCPS707, NTCPS702, NTCPS701, AICPS165, AICPS158, AICPS106, AICPS005, AICPS001, AICPS006, AICPS009, AICPS035, UGNTC737, 07NG_SN022, 07NG_SN305, 07NG_SN223, 07NG_SN077, 07NG_SN277, 07NG_SN132, 07NG_SN157, SK6031_M18, SK6039_M24, SK9037_PI, SK6074_PI, SK3015_PI, SK8040_PI, SK5010_PI, SK9027_PI, SK5013_PI, 0765HR, 2917HR, 3132HR, 3272HR, 3445HR, 4668HR, 6356HR, 6699HR, 7475HR, 7518HR, 7672HR, 3083HR, 4834HR, 3153HR, 3635HR, 3683HR, 6891HR, U455_U455A, 97CDKFE4, 02CD_KBS187, 02CD_KSS004, A2_225_692, QA779, QA284, 01CM_1445MV, P32_16_1a10, P32_16_1b3, P32_16_1b4, P32_16_1b5, P32_16_1b7, P32_16_1c2, P32_16_1c5, P32_16_1c7, P32_16_1e5, P32_16_1f1, P32_16_1f3, P32_16_1f2, P32_16_2a6, P32_16_2b3, P32_16_2c7, P32_16_2c8, P32_16_2d2, P32_16_2d6, P32_16_2e1, P32_16_2e9, P32_16_4a2, P32_16_4a4, P32_16_4b6, P32_16_4b7, P32_16_4b11, P32_16_4c12, P32_16_4d4, P32_16_4d6, P32_16_4d7, P32_16_4e6, P32_16_4e8, P32_16_4f4, P32_24_2e7, P32_24_1a3, P32_24_1a5, P32_24_1a8, P32_24_1b9, P32_24_1c7, P32_24_1c8, P32_24_1e8, P32_24_2a10, P32_24_2b7, P32_24_2b12, P32_24_2c6, P32_24_2e2, P32_24_2e4, P32_24_2f6, P32_24_2f9, P32_24_3a2, P32_24_3a8, P32_24_3d4, P32_24_3d11, P32_24_3e9, P32_24_3f8, P32_24_4a3, P32_24_4a4, P32_24_4a12, P32_24_4b5, P32_24_4b9, P32_24_4b10, P32_24_4c4, P32_24_4c12, P32_24_4f8, P32_44_1a8, P32_44_2a3, P32_44_2a8, P32_44_2c8, P32_44_2d3, P32_44_3a1, P32_44_3a4, P32_44_3a12, P32_44_3c5, P32_44_3d3, P32_44_3f8, P32_44_3f10, 71377PL10, A57, A58, A41, A44, A46, A51, A55, A56, A63, A69, A71, A72, A52, A54, A42, A43, A62, A5, A3, A61, A65, A21, A26, A31, A19, A25, A28, A29, A70, A73, A33, A22, A24, A32, A39, A40, A48, 94CY017_41, CY008, CY009, CY012, CY015, CY057, CY058, CY111, CY112, CY153, CY169, CY170, CY171, CY173, CY178, CY021, CY022, CY023, CY051, CY064, CY106, CY121, CY140, CY164, CY182, CY185, CY196, CY207, CY208, CY209, CY213, CY218, CY230, CY231, CY235, CY236, CY243, CY255, 67429PL12, 85344PL1, 71264PL11, 87516PL1, 89629PL1, 89794PL1, 71372PL14, 75925PL8, 82472PL1, 75925PL9, SL17852PL1, 94275PL1, 95074PL1, 95310PL1, 94412PL1, 95314PL1, 94412PL2, 94389PL1, 75925PL10, 97023PL1, 94412PL3, 88882PL1, 88662PL1, 88372PL1, 89160PL1, 89161PL1, 88518PL1, 88723PL1, 98821PL1, 75925PL11, 99255PL1, 99517PL1, 99925PL1, 99772PL1, 100614PL1, 72193PL19, 102032PL1, 94412PL5, 101297PL1, 75925PL12, 102766PL1, 102620PL1, 94412PL6, 89160PL3, 104198PL1, 104100PL1, 103562PL1, 103995PL1, 103880PL1, 104103PL1, 103887PL1, 104589PL1, 94412PL8, 76153PL5, 105739PL1, 71264PL12, 98821PL2, 106804PL1, 94412PL10, 75925PL14, 108393PL1, 109856PL1, 76153PL6, 111637PL1, 111640PL1, 112522PL1, 112688PL1, 113058PL1, 114213PL1, 113727PL1, 114099PL1, 114942PL1, 115520PL1, 115989PL1, 76153PL7, 116508PL1, 116214PL1, 116928PL1, 116880PL1, 117167PL1, 117345PL1, 111863PL1, 117052PL1, 107836PL4, 119984PL1, 117058PL1, 120094PL1, 107836PL6, 121001PL1, 71164PL14, 122395PL1, 121653PL1, 122644PL1, SI4037_02

Dataset B

EU345825, EU345824, EU345823, EU345822, EU345821, EU345820, EU345819, EU345818, EU345817, EU345815, EU345803, EU345802, EU345801, EU345800, EU345799, EU345773, EU345771, EU345770, EU345767, EU345766, EU345765, EU345763, EU345762, EU345761, EU345760, EU345722, EU345721, EU345719, EU345718, EU345716, EU345715, EU345714, EU345713, EU345711, EU345709, EU345905, EU345904, EU345903, EU345902, EU345901, EU345900, EU345899, EU345898, EU345897, EU345895, EF545110, EF545109, AY500393, FJ864679, FJ822134, FJ822133, FJ822132, FJ822131, FJ822130, FJ822129, FJ822128, FJ822127, FJ822126, FJ009813, FJ009812, FJ009811, FJ009809, FJ009808, FJ009807, FJ009806, FJ009805, FJ009804, FJ009803, GQ867635, GQ867631, GQ867630, GQ867628, GQ867627, GQ867625, GQ867624, HQ161930, HQ161929, HQ161928, HQ161927, HQ161926, HQ161925, HQ161924, HQ161923, HQ161921, HQ130000, HQ129999, HQ129998, HQ129997, HQ129996, HQ129995, HQ129994, HQ129993, HQ129991, HQ129990, HQ129989, HQ449403, HQ449402, HQ449398, HQ449397, HQ449395, HQ449393, HQ896679, HM345560, HM345558, HM345554, HM345536, HM345579, HM345575, HM345569, HM345565, HM345553, HM345551, HM345539, HM345537, JN673405, JQ292898, JQ292896, JQ292894, JQ292899, JQ292895, JQ292893, JX141227, JX141225, JX141223, JX141221, JX141219, JX141217, JX141215, JX141213, JX141211, JX141209, JX141226, JX141224, JX141222, JX141220, JX290238, JX290236, JX290234, JX290232, JX290230, JX290228, JX290226, JX290224, JX290222, JX290218, JX500694, KC156532, KC156533, KC254612, KC254643, KC254641, KC254639, KC254637, KC254635, KC254633, KC254631, KC665939, KC665938, KC665937, KC665936, KC665934, KC665933, KC665932, KC665931, KC665930, KC665929, KC665928, KF257884, KF257882, KF257881, KF257880, KF257879, KF257878, KF257877, KF257876, KF257875, KF257874, KF257873, KF257870, KF374888, KF374887, KF374886, KF374885, KF374884, KF374883, KF374882, KF374881, KF374880, KF374879, KF374878, KF374877, KF374876, KF374875, KF374872, KF374871, KF374870, KF374869, KF374868, KF374867, KF374866, KF374865, KF374864, KF374863, KF374862, KF374861, KF374860, KF374859, KF374858, KF374857, KF374856, KF374855, KF374854, KF374853, KF374852, KF374851, KF971945, KF971938, KF971937, KF971936, KF971933, KF971929, KF971926, KF971920, KF971919, KJ499635, KJ499634, KJ499633, KJ499632, KJ499631, KJ499630, KJ499628, KJ499626, KJ499625, KJ499622, KJ499621, KJ499620, KJ499619, KJ499618, KJ499617, KJ499616, KJ499615, KJ499614, KJ499612, KJ499611, KJ499610, KJ499609, KJ499592, KJ499591, KJ499590, EU345703, EU345738, FJ009725, FJ009731, FJ009736, FJ009750, FJ009751, FJ009754, FJ009768, FJ009774, FJ822080, KF716492, KF716491, JX500695, JX500696, KJ499563, KJ499562, KJ499561, KJ499560, KJ499559, KJ722123, KJ722122, KJ722121, KJ722120, KJ722119, KJ722118, KJ722117, KJ722116, KJ722115, KJ722114, KJ722113, KJ722109, KJ722075, KJ722074, KJ722073, KJ722072, KJ722071, KJ722070, KJ197223, KJ197220, KJ197218, EU345840, EU345839, EU345838, EU345837, EU345836, EU345835, EU345834, EU345833, EU345832, EU345831, DQ823367, DQ823366, DQ823365, DQ823361, DQ823360, DQ823359, DQ823358, DQ823357, DQ823356, DQ055315, DQ055313, DQ055310, DQ055308, DQ055307, DQ055306, DQ055304, DQ055303, DQ055302, DQ055301, DQ055299, DQ055298, DQ055297, DQ055295, DQ055293, DQ055291, DQ055290, DQ055289, DQ055288, DQ055285, DQ055283, DQ055282, DQ055281, DQ055279, DQ055278, DQ055277, DQ055276, DQ055274, DQ055273, DQ055272, DQ055271, DQ055270, DQ055269, DQ055268, DQ055267, DQ055266, DQ055265, DQ055264, DQ055263, DQ055262, DQ055261, DQ055260, DQ055259, DQ055257, DQ055256, DQ055255, DQ055254, DQ055253, DQ055252, DQ055251, DQ055248, DQ055247, DQ055246, DQ055245, DQ055244, DQ055243, DQ055242, DQ055241, DQ055240, DQ055239, DQ055238, DQ055235, DQ055234, DQ055233, DQ055232, DQ055231, DQ055230, DQ055229, DQ055225, DQ055224, DQ055223, DQ055221, DQ055219, DQ055210, DQ055206, DQ055204, DQ055199, DQ055196, DQ055195, DQ055194, DQ055193, DQ055192, DQ055191, DQ055190, DQ055189, DQ055188, DQ055187, DQ055186, DQ055184, DQ055183, DQ055182, DQ055181, DQ055180, DQ055178, DQ055177, DQ055176, DQ055175, DQ055174, AF413987, HQ161910, HQ161909, HQ161908, HQ161907, HQ161906, HQ115075, HQ115071, HQ115070, HQ115068, HQ115067

Dataset C

DQ823367, DQ823366, DQ823365, DQ823361, DQ823360, DQ823359, DQ823358, DQ823357, DQ823356, AF413987, AF025596, AF025595, AF025594, AF025593, AF025592, AF025591, AF025590, AF025589, AF025588, AF025587, AF025586, AF025585, AF025584, AF025583, AF025582, AF025581, AF025580, AF082486, U93676, U93665, AF100939, AF100934, AF098953, EF545108, AY500393, FJ864679, GU481683, GU481660, GU481635, GU481589, GU481565, GU481536, GU481518, GU481471, GU481420, GU481391, GU481370, GU481328, GU481276, GU481253, GU481202, GU481155, HQ385842, HQ385840, HQ385839, HQ385838, HQ385837, HQ385830, HQ385820, HQ385819, HQ385818, HQ385817, HQ338117, HQ338113, HQ338112, HQ338111, HQ616096, HQ616092, HQ616084, HQ616085, HQ616083, JF327807, JF430896, JQ292900, JQ292898, JQ292896, JQ292892, JQ292897, JQ292895, HQ834965, HQ896492, HQ896489, JF952017, JF952015, JF952013, JF952018, JF952016, JF952014, JF952012, JX500695, JX500696, JX5006

Appendix 2

(Chapter 3)

Migration events in the full combined dataset, 2014-2015; HPD – Highest Posterior Density

From	To	N	95% HPD	%	$\sum$ exp [95% HPD]	% [95% HPD]	$\sum$ imp [95% HPD]	% [95% HPD]
Center	Center	-	-	-	4.83 [0, 11]	1.59 [0, 3]	78.76 [74, 84]	25.96 [24, 28]
	Crimea	0.57	[0, 2]	0.19				
	East	0.38	[0, 2]	0.13				
	Kyiv	0.87	[0, 3]	0.29				
	Odessa	1.01	[0, 4]	0.33				
	South	0.93	[0, 4]	0.31				
	West	1.05	[0, 4]	0.35				
Crimea	Center	0.09	[0, 1]	0.03	0.52 [0, 2]	0.17 [0, 0.6]	10.39 [10, 12]	3.43 [3, 4]
	Crimea	-	-	-				
	East	0.05	[0, 0]	0.02				
	Kyiv	0.08	[0, 1]	0.03				
	Odessa	0.09	[0, 1]	0.03				
	South	0.10	[0, 1]	0.03				
	West	0.11	[0, 1]	0.04				
East	Center	73.57	[65, 80]	24.27	279.73 [264, 292]	92.27 [87, 96]	1.83 [0, 5]	0.60 [0, 2]
	Crimea	7.55	[2, 10]	2.49				
	East	-	-	-				
	Kyiv	65.33	[57, 72]	21.55				
	Odessa	47.88	[41, 53]	15.79				
	South	68.55	[59, 75]	22.61				
	West	16.85	[10, 21]	5.56				
Kyiv	Center	2.14	[0, 6]	0.71	7.00 [1, 13]	2.31 [0, 4]	67.65 [61, 74]	22.31 [20, 24]
	Crimea	0.60	[0, 2]	0.20				
	East	0.52	[0, 2]	0.17				
	Kyiv	-	-	-				
	Odessa	0.59	[0, 3]	0.19				
	South	0.73	[0, 3]	0.24				
	West	2.47	[0, 6]	0.81				
Odessa	Center	0.77	[0, 3]	0.25		1.36 [0, 3]	50.88 [46, 55]	16.78 [15, 18]
	Crimea	0.37	[0, 2]	0.12				
	East	0.33	[0, 2]	0.11				

	Kyiv	0.49	[0, 2]	0.16	4.13 [0, 9]			
	Odessa	-	-	-				
	South	1.69	[0, 4]	0.56				
	West	0.44	[0, 2]	0.15				
South	Center	2.00	[0, 6]	0.66	5.36 [1, 12]	1.76 [0, 4]	72.24 [65, 78]	23.83 [21, 26]
	Crimea	0.87	[0, 3]	0.29				
	East	0.43	[0, 2]	0.14				
	Kyiv	0.64	[0, 3]	0.21				
	Odessa	0.75	[0, 3]	0.25				
	South	-	-	-				
	West	0.68	[0, 3]	0.22				
West	Center	0.21	[0, 1]	0.07	1.74 [0, 5]	0.57 [0, 2]	21.60 [19, 24]	7.12 [6, 8]
	Crimea	0.40	[0, 2]	0.13				
	East	0.12	[0, 1]	0.04				
	Kyiv	0.20	[0, 1]	0.07				
	Odessa	0.58	[0, 2]	0.19				
	South	0.23	[0, 1]	0.08				
	West	-	-	-				
Total number of migration events					303.18 [289, 315]			

Appendix 3

(Chapter 3)

Distribution of migration events in 10 subsets of the Ukrainian dataset (after balancing the number of sequencing per location); HPD – Highest Posterior Density; the main exporter is marked in bold.

Subsample		1			2		
From	To	N	95% HPD	%	N	95% HPD	%
Center	All exp.	3.98	[0, 9]	1.6	4.11	[0, 9]	1.7
	All imp.	56.38	[52, 60]	23.2	56.88	[52, 61]	23.4
	Crimea	0.49	[0, 2]	0.2	0.54	[0, 2]	0.2
	East	0.32	[0, 2]	0.1	0.33	[0, 2]	0.1
	Kyiv	0.85	[0, 3]	0.4	0.73	[0, 3]	0.3
	Odessa	0.72	[0, 3]	0.3	0.83	[0, 3]	0.3
	South	0.66	[0, 3]	0.3	0.85	[0, 3]	0.3
	West	0.97	[0, 3]	0.4	0.83	[0, 3]	0.3
Crimea	All exp.	0.70	[0, 2]	0.3	0.70	[0, 2]	0.3
	All imp.	10.49	[10, 12]	4.3	10.49	[10, 12]	4.3
	Center	0.11	[0, 1]	0.0	0.11	[0, 1]	0.0
	East	0.05	[0, 1]	0.0	0.06	[0, 1]	0.0
	Kyiv	0.13	[0, 1]	0.1	0.12	[0, 1]	0.0
	Odessa	0.11	[0, 1]	0.0	0.10	[0, 1]	0.0
	South	0.14	[0, 1]	0.1	0.14	[0, 1]	0.1
	West	0.14	[0, 1]	0.1	0.15	[0, 1]	0.1
East	All exp.	**219.71**	**[204, 232]**	**90.6**	**220.53**	**[207, 234]**	**90.6**
	All imp.	1.68	[0, 5]	0.7	1.73	[0, 5]	0.7
	Center	50.90	[44, 58]	21.0	52.12	[43, 58]	21.4
	Crimea	7.46	[3, 11]	3.1	7.60	[3, 11]	3.1
	Kyiv	50.02	[43, 57]	20.6	48.82	[41, 56]	20.1
	Odessa	47.26	[40, 53]	19.5	47.23	[39, 53]	19.4
	South	46.94	[40, 54]	19.3	47.87	[41, 55]	19.7
	West	17.01	[10, 22]	7.0	16.84	[11, 22]	6.9
Kyiv	All exp.	6.27	[0, 12]	2.6	5.90	[1, 12]	2.4
	All imp.	52.65	[46, 58]	21.7	51.43	[45, 57]	21.1
	Center	1.95	[0, 6]	0.8	1.36	[0, 5]	0.6
	Crimea	0.50	[0, 2]	0.2	0.52	[0, 2]	0.2
	East	0.43	[0, 2]	0.2	0.44	[0, 2]	0.2

	Odessa	0.56	[0, 2]	0.2	0.55	[0, 2]	0.2
	South	0.77	[0, 3]	0.3	0.63	[0, 3]	0.3
	West	2.10	[0, 5]	0.9	2.47	[0, 6]	1.0
Odessa	All exp.	4.50	[0, 10]	1.9	4.78	[0, 11]	2.0
	All imp.	50.01	[44, 55]	20.6	50.10	[44, 55]	20.6
	Center	1.14	[0, 4]	0.5	1.15	[0, 4]	0.5
	Crimea	0.67	[0, 3]	0.3	0.57	[0, 2]	0.2
	East	0.36	[0, 2]	0.1	0.36	[0, 2]	0.1
	Kyiv	0.76	[0, 3]	0.3	0.87	[0, 3]	0.4
	South	0.87	[0, 3]	0.4	1.19	[0, 4]	0.5
	West	0.67	[0, 3]	0.3	0.63	[0, 3]	0.3
South	All exp.	5.47	[0, 12]	2.3	5.29	[0, 12]	2.2
	All imp.	49.73	[43, 55]	20.5	50.98	[45, 57]	21.0
	Center	1.91	[0, 5]	0.8	1.86	[0, 5]	0.8
	Crimea	1.00	[0, 3]	0.4	0.83	[0, 3]	0.3
	East	0.44	[0, 2]	0.2	0.43	[0, 2]	0.2
	Kyiv	0.59	[0, 3]	0.2	0.61	[0, 3]	0.2
	Odessa	0.65	[0, 3]	0.3	0.72	[0, 3]	0.3
	West	0.85	[0, 4]	0.3	0.79	[0, 3]	0.3
West	All exp.	2.02	[0, 5]	0.8	2.05	[0, 5]	0.8
	All imp.	21.78	[18, 24]	9.0	21.73	[19, 24]	8.9
	Center	0.31	[0, 2]	0.1	0.27	[0, 1]	0.1
	Crimea	0.40	[0, 2]	0.2	0.43	[0, 2]	0.2
	East	0.14	[0, 1]	0.1	0.15	[0, 1]	0.1
	Kyiv	0.28	[0, 1]	0.1	0.28	[0, 1]	0.1
	Odessa	0.62	[0, 2]	0.3	0.63	[0, 2]	0.3
	South	0.27	[0, 1]	0.1	0.29	[0, 1]	0.1
Total N		242.63	[231, 255]	100.0	243.31	[230, 255]	100.0

Subsample		3			4		
From	To	N	95% HPD	%	N	95% HPD	%
Center	All exp.	3.90	[0, 9]	1.6	4.36	[0, 10]	1.8
	All imp.	55.63	[51, 60]	23.0	57.18	[52, 61]	23.6
	Crimea	0.54	[0, 2]	0.2	0.53	[0, 2]	0.2
	East	0.30	[0, 2]	0.1	0.82	[0, 3]	0.3
	Kyiv	0.63	[0, 3]	0.3	0.38	[0, 2]	0.2

	Odessa	0.78	[0, 3]	0.3	0.91	[0, 3]	0.4
	South	0.77	[0, 3]	0.3	0.76	[0, 3]	0.3
	West	0.89	[0, 3]	0.4	0.92	[0, 3]	0.4
Crimea	All exp.	0.64	[0, 2]	0.3	0.64	[0, 2]	0.3
	All imp.	10.48	[10, 12]	4.3	10.46	[10, 12]	4.3
	Center	0.11	[0, 1]	0.0	0.12	[0, 1]	0.0
	East	0.06	[0, 1]	0.0	0.10	[0, 1]	0.0
	Kyiv	0.12	[0, 1]	0.1	0.06	[0, 1]	0.0
	Odessa	0.10	[0, 1]	0.0	0.10	[0, 1]	0.0
	South	0.13	[0, 1]	0.1	0.12	[0, 1]	0.0
	West	0.13	[0, 1]	0.1	0.15	[0, 1]	0.1
East	All exp.	**220.18**	**[204, 233]**	**91.1**	5.68	[0, 12]	2.3
	All imp.	1.61	[0, 4]	0.7	49.11	[42, 55]	20.3
	Center	50.85	[42, 57]	21.0	1.07	[0, 4]	0.4
	Crimea	7.60	[3, 11]	3.1	0.64	[0, 2]	0.3
	Kyiv	49.49	[43, 55]	20.5	0.42	[0, 2]	0.2
	Odessa	46.64	[39, 53]	19.3	0.58	[0, 2]	0.2
	South	47.23	[39, 54]	19.5	2.36	[0, 6]	1.0
	West	18.34	[13, 23]	7.6	0.61	[0, 2]	0.3
Kyiv	All exp.	3.75	[0, 9]	1.6	**219.05**	**[203, 233]**	**90.4**
	All imp.	51.74	[47, 57]	21.4	1.86	[0, 5]	0.8
	Center	0.70	[0, 3]	0.3	52.97	[45, 59]	21.9
	Crimea	0.51	[0, 2]	0.2	7.33	[2, 10]	3.0
	East	0.36	[0, 2]	0.2	46.55	[38, 53]	19.2
	Odessa	0.53	[0, 2]	0.2	47.30	[40, 54]	19.5
	South	0.71	[0, 3]	0.3	46.09	[38, 54]	19.0
	West	0.96	[0, 3]	0.4	18.84	[13, 23]	7.8
Odessa	All exp.	5.12	[0, 11]	2.1	5.08	[0, 11]	2.1
	All imp.	49.53	[43, 55]	20.5	50.31	[45, 56]	20.8
	Center	1.00	[0, 4]	0.4	1.03	[0, 4]	0.4
	Crimea	0.47	[0, 2]	0.2	0.50	[0, 2]	0.2
	East	0.35	[0, 2]	0.1	0.47	[0, 2]	0.2
	Kyiv	0.72	[0, 3]	0.3	0.39	[0, 2]	0.2
	South	1.92	[0, 5]	0.8	2.17	[0, 6]	0.9
	West	0.66	[0, 3]	0.3	0.53	[0, 2]	0.2
South	All exp.	5.91	[0, 13]	2.4	5.38	[0, 12]	2.2
	All imp.	50.98	[45, 56]	21.1	51.75	[45, 57]	21.3
	Center	2.66	[0, 7]	1.1	1.70	[0, 5]	0.7
	Crimea	0.94	[0, 3]	0.4	1.01	[0, 3]	0.4
	East	0.42	[0, 2]	0.2	0.95	[0, 3]	0.4
	Kyiv	0.53	[0, 2]	0.2	0.41	[0, 2]	0.2

	Odessa	0.67	[0, 3]	0.3	0.65	[0, 3]	0.3
	West	0.67	[0, 3]	0.3	0.61	[0, 3]	0.3
West	All exp.	2.23	[0, 5]	0.9	2.19	[0, 5]	0.9
	All imp.	21.66	[19, 24]	9.0	21.69	[18, 24]	8.9
	Center	0.30	[0, 2]	0.1	0.29	[0, 1]	0.1
	Crimea	0.45	[0, 2]	0.2	0.46	[0, 2]	0.2
	East	0.15	[0, 1]	0.1	0.21	[0, 1]	0.1
	Kyiv	0.27	[0, 1]	0.1	0.19	[0, 1]	0.1
	Odessa	0.80	[0, 2]	0.3	0.80	[0, 2]	0.3
	South	0.25	[0, 1]	0.1	0.25	[0, 1]	0.1
Total N		241.76	[228, 253]	100.0	242.39	[228, 255]	100.0

Subsample		5			6		
From	To	N	95% HPD	%	N	95% HPD	%
Center	All exp.	3.84	[0, 9]	1.6	4.12	[0, 9]	1.7
	All imp.	55.54	[51, 60]	23.0	55.13	[51, 60]	22.8
	Crimea	0.55	[0, 2]	0.2	0.48	[0, 2]	0.2
	East	0.35	[0, 2]	0.1	0.73	[0, 3]	0.3
	Kyiv	0.75	[0, 3]	0.3	0.70	[0, 3]	0.3
	Odessa	0.82	[0, 3]	0.3	0.84	[0, 3]	0.3
	South	0.71	[0, 3]	0.3	0.36	[0, 2]	0.2
	West	0.58	[0, 3]	0.2	1.04	[0, 4]	0.4
Crimea	All exp.	0.69	[0, 3]	0.3	0.67	[0, 2]	0.3
	All imp.	10.48	[10, 12]	4.3	10.50	[10, 12]	4.3
	Center	0.12	[0, 1]	0.0	0.12	[0, 1]	0.0
	East	0.06	[0, 1]	0.0	0.10	[0, 1]	0.0
	Kyiv	0.11	[0, 1]	0.0	0.12	[0, 1]	0.0
	Odessa	0.11	[0, 1]	0.0	0.10	[0, 1]	0.0
	South	0.12	[0, 1]	0.1	0.07	[0, 1]	0.0
	West	0.16	[0, 1]	0.1	0.15	[0, 1]	0.1
East	All exp.	**218.14**	**[204, 230]**	**90.5**	5.97	[1, 12]	2.5
	All imp.	1.78	[0, 5]	0.7	49.82	[42, 56]	20.6
	Center	51.78	[44, 57]	21.5	1.87	[0, 5]	0.8
	Crimea	7.45	[3, 11]	3.1	0.52	[0, 2]	0.2
	Kyiv	46.61	[39, 53]	19.3	1.12	[0, 4]	0.5

	Odessa	48.32	[41, 54]	20.0	0.73	[0, 3]	0.3
	South	47.27	[39, 54]	19.6	1.18	[0, 4]	0.5
	West	16.78	[11, 22]	7.0	0.49	[0, 2]	0.2
Kyiv	All exp.	5.71	[0, 11]	2.4	6.82	[0, 13]	2.8
	All imp.	48.89	[43, 55]	20.3	50.82	[44, 57]	21.0
	Center	0.91	[0, 3]	0.4	2.11	[0, 5]	0.9
	Crimea	0.44	[0, 2]	0.2	0.62	[0, 2]	0.3
	East	0.42	[0, 2]	0.2	0.55	[0, 2]	0.2
	Odessa	0.55	[0, 2]	0.2	0.72	[0, 3]	0.3
	South	0.82	[0, 3]	0.3	0.54	[0, 2]	0.2
	West	2.56	[0, 6]	1.1	2.26	[0, 5]	0.9
Odessa	All exp.	5.43	[0, 11]	2.2	4.01	[0, 9]	1.7
	All imp.	51.21	[45, 56]	21.2	50.80	[46, 56]	21.0
	Center	0.93	[0, 3]	0.4	0.78	[0, 3]	0.3
	Crimea	0.64	[0, 3]	0.3	0.46	[0, 2]	0.2
	East	0.39	[0, 2]	0.2	0.71	[0, 3]	0.3
	Kyiv	0.53	[0, 2]	0.2	0.86	[0, 3]	0.4
	South	2.22	[0, 6]	0.9	0.45	[0, 2]	0.2
	West	0.76	[0, 3]	0.3	0.72	[0, 3]	0.3
South	All exp.	5.42	[0, 12]	2.2	**217.96**	**[203, 232]**	**90.2**
	All imp.	51.41	[45, 57]	21.3	2.75	[0, 6]	1.1
	Center	1.64	[0, 5]	0.7	49.99	[42, 56]	20.7
	Crimea	1.00	[0, 3]	0.4	8.01	[4, 11]	3.3
	East	0.43	[0, 2]	0.2	47.54	[40, 55]	19.7
	Kyiv	0.62	[0, 3]	0.3	47.75	[39, 54]	19.8
	Odessa	0.81	[0, 3]	0.3	47.68	[41, 54]	19.7
	West	0.94	[0, 4]	0.4	17.05	[11, 22]	7.1
West	All exp.	1.90	[0, 5]	0.8	2.03	[0, 5]	0.8
	All imp.	21.81	[19, 24]	9.0	21.73	[19, 24]	9.0
	Center	0.22	[0, 1]	0.1	0.28	[0, 1]	0.1
	Crimea	0.41	[0, 2]	0.2	0.41	[0, 2]	0.2
	East	0.14	[0, 1]	0.1	0.18	[0, 1]	0.1
	Kyiv	0.26	[0, 1]	0.1	0.26	[0, 1]	0.1
	Odessa	0.58	[0, 2]	0.2	0.71	[0, 2]	0.3
	South	0.29	[0, 1]	0.1	0.18	[0, 1]	0.1
Total N		241.13	[229, 253]	100.0	241.54	[228, 254]	100.0
Subsample		7			8		

From	To	N	95% HPD	%	N	95% HPD	%
Center	All exp.	3.57	[0, 8]	1.5	**222.25**	**[208, 236]**	**91.2**
	All imp.	56.20	[52, 60]	23.1	2.33	[0, 6]	1.0
	Crimea	0.55	[0, 2]	0.2	7.93	[4, 11]	3.3
	East	0.28	[0, 1]	0.1	49.58	[41, 57]	20.4
	Kyiv	0.57	[0, 3]	0.2	50.50	[43, 57]	20.7
	Odessa	0.66	[0, 3]	0.3	47.60	[41, 53]	19.5
	South	0.76	[0, 3]	0.3	48.25	[40, 55]	19.8
	West	0.74	[0, 3]	0.3	18.18	[13, 23]	7.5
Crimea	All exp.	0.68	[0, 2]	0.3	0.68	[0, 2]	0.3
	All imp.	10.50	[10, 12]	4.3	10.51	[10, 12]	4.3
	Center	0.13	[0, 1]	0.1	0.07	[0, 1]	0.0
	East	0.06	[0, 1]	0.0	0.11	[0, 1]	0.0
	Kyiv	0.12	[0, 1]	0.1	0.14	[0, 1]	0.1
	Odessa	0.10	[0, 1]	0.0	0.11	[0, 1]	0.0
	South	0.12	[0, 1]	0.1	0.11	[0, 1]	0.0
	West	0.14	[0, 1]	0.1	0.15	[0, 1]	0.1
East	All exp.	**221.65**	**[209, 233]**	**91.3**	6.44	[0, 13]	2.6
	All imp.	1.67	[0, 4]	0.7	52.30	[45, 59]	21.5
	Center	51.26	[43, 57]	21.1	0.57	[0, 2]	0.2
	Crimea	7.92	[4, 11]	3.3	0.52	[0, 2]	0.2
	Kyiv	47.35	[41, 54]	19.5	1.64	[0, 5]	0.7
	Odessa	47.75	[40, 53]	19.7	0.90	[0, 3]	0.4
	South	49.98	[43, 56]	20.6	2.14	[0, 6]	0.9
	West	17.36	[12, 22]	7.1	0.66	[0, 3]	0.3
Kyiv	All exp.	6.45	[0, 12]	2.7	4.80	[0, 11]	2.0
	All imp.	49.63	[43, 55]	20.4	53.66	[48, 59]	22.0
	Center	2.10	[0, 5]	0.9	0.52	[0, 2]	0.2
	Crimea	0.57	[0, 2]	0.2	0.63	[0, 3]	0.3
	East	0.43	[0, 2]	0.2	0.86	[0, 3]	0.4
	Odessa	0.55	[0, 2]	0.2	0.58	[0, 3]	0.2
	South	0.52	[0, 2]	0.2	0.52	[0, 2]	0.2
	West	2.26	[0, 5]	0.9	1.67	[0, 5]	0.7
Odessa	All exp.	4.35	[0, 10]	1.8	3.92	[0, 9]	1.6
	All imp.	50.39	[44, 55]	20.7	50.41	[45, 55]	20.7
	Center	1.01	[0, 4]	0.4	0.41	[0, 2]	0.2
	Crimea	0.48	[0, 2]	0.2	0.46	[0, 2]	0.2
	East	0.37	[0, 2]	0.2	0.66	[0, 3]	0.3
	Kyiv	0.67	[0, 3]	0.3	0.62	[0, 3]	0.3

	South	1.14	[0, 4]	0.5	1.14	[0, 4]	0.5
	West	0.69	[0, 3]	0.3	0.60	[0, 3]	0.2
South	All exp.	4.29	[0, 10]	1.8	3.49	[0, 8]	1.4
	All imp.	52.75	[48, 58]	21.7	52.56	[48, 58]	21.6
	Center	1.37	[0, 5]	0.6	0.58	[0, 2]	0.2
	Crimea	0.65	[0, 3]	0.3	0.52	[0, 2]	0.2
	East	0.39	[0, 2]	0.2	0.84	[0, 3]	0.3
	Kyiv	0.64	[0, 3]	0.3	0.43	[0, 2]	0.2
	Odessa	0.72	[0, 3]	0.3	0.55	[0, 2]	0.2
	West	0.57	[0, 3]	0.2	0.60	[0, 3]	0.2
West	All exp.	1.92	[0, 5]	0.8	2.04	[0, 5]	0.8
	All imp.	21.77	[19, 24]	9.0	21.80	[19, 24]	8.9
	Center	0.27	[0, 1]	0.1	0.17	[0, 1]	0.1
	Crimea	0.39	[0, 2]	0.2	0.48	[0, 2]	0.2
	East	0.16	[0, 1]	0.1	0.21	[0, 1]	0.1
	Kyiv	0.24	[0, 1]	0.1	0.32	[0, 2]	0.1
	Odessa	0.63	[0, 2]	0.3	0.59	[0, 2]	0.2
	South	0.24	[0, 1]	0.1	0.29	[0, 1]	0.1
Total N		242.88	[230, 254]	100.0	243.62	[230, 256]	100.0
Subsample		9			10		
From	To	N	95% HPD	%	N	95% HPD	%
Center	All exp.	4.19	[0, 9]	1.7	4.11	[0, 9]	1.7
	All imp.	55.97	[51, 60]	23.1	54.85	[50, 59]	22.7
	Crimea	0.57	[0, 2]	0.2	0.49	[0, 2]	0.2
	East	0.32	[0, 2]	0.1	0.30	[0, 2]	0.1
	Kyiv	0.76	[0, 3]	0.3	0.75	[0, 3]	0.3
	Odessa	0.87	[0, 3]	0.4	0.88	[0, 3]	0.4
	South	0.89	[0, 3]	0.4	0.67	[0, 3]	0.3
	West	0.82	[0, 3]	0.3	1.07	[0, 4]	0.4
Crimea	All exp.	0.65	[0, 2]	0.3	0.64	[0, 2]	0.3
	All imp.	10.47	[10, 12]	4.3	10.50	[10, 12]	4.3
	Center	0.12	[0, 1]	0.0	0.10	[0, 1]	0.0
	East	0.07	[0, 1]	0.0	0.04	[0, 0]	0.0
	Kyiv	0.10	[0, 1]	0.0	0.10	[0, 1]	0.0
	Odessa	0.11	[0, 1]	0.0	0.12	[0, 1]	0.1
	South	0.11	[0, 1]	0.0	0.12	[0, 1]	0.1
	West	0.19	[0, 1]	0.1	0.13	[0, 1]	0.1
East	All exp.	**219.97**	**[205,**	**90.8**	**220.51**	**[205,**	**91.3**

			233]			**233]**	
	All imp.	1.74	[0, 5]	0.7	1.59	[0, 4]	0.7
	Center	51.08	[43, 58]	21.1	51.13	[44, 57]	21.2
	Crimea	7.76	[4, 11]	3.2	7.78	[4, 11]	3.2
	Kyiv	46.71	[38, 53]	19.3	48.21	[41, 55]	20.0
	Odessa	47.01	[40, 53]	19.4	47.72	[41, 54]	19.8
	South	50.11	[43, 56]	20.7	48.65	[41, 55]	20.1
	West	17.41	[11, 22]	7.2	17.08	[10, 22]	7.1
Kyiv	All exp.	5.78	[0, 11]	2.4	5.83	[1, 12]	2.4
	All imp.	49.38	[42, 55]	20.4	50.65	[45, 57]	21.0
	Center	1.88	[0, 5]	0.8	1.60	[0, 5]	0.7
	Crimea	0.44	[0, 2]	0.2	0.53	[0, 2]	0.2
	East	0.41	[0, 2]	0.2	0.38	[0, 2]	0.2
	Odessa	0.64	[0, 3]	0.3	0.49	[0, 2]	0.2
	South	0.53	[0, 2]	0.2	0.71	[0, 3]	0.3
	West	1.91	[0, 5]	0.8	2.17	[0, 5]	0.9
Odessa	All exp.	4.84	[0, 11]	2.0	4.61	[0, 10]	1.9
	All imp.	50.08	[44, 55]	20.7	50.46	[44, 55]	20.9
	Center	1.08	[0, 4]	0.4	1.06	[0, 4]	0.4
	Crimea	0.51	[0, 2]	0.2	0.64	[0, 2]	0.3
	East	0.37	[0, 2]	0.2	0.36	[0, 2]	0.1
	Kyiv	0.94	[0, 4]	0.4	0.63	[0, 3]	0.3
	South	1.35	[0, 4]	0.6	1.37	[0, 4]	0.6
	West	0.59	[0, 2]	0.2	0.59	[0, 2]	0.2
South	All exp.	4.61	[0, 11]	1.9	3.60	[0, 9]	1.5
	All imp.	53.13	[47, 58]	21.9	51.78	[46, 57]	21.4
	Center	1.56	[0, 5]	0.6	0.78	[0, 3]	0.3
	Crimea	0.73	[0, 3]	0.3	0.57	[0, 2]	0.2
	East	0.38	[0, 2]	0.2	0.33	[0, 2]	0.1
	Kyiv	0.64	[0, 3]	0.3	0.65	[0, 3]	0.3
	Odessa	0.67	[0, 3]	0.3	0.65	[0, 3]	0.3
	West	0.65	[0, 3]	0.3	0.65	[0, 3]	0.3
West	All exp.	2.20	[0, 5]	0.9	2.04	[0, 5]	0.8
	All imp.	21.58	[19, 24]	8.9	21.72	[19, 24]	9.0
	Center	0.27	[0, 1]	0.1	0.29	[0, 2]	0.1
	Crimea	0.46	[0, 2]	0.2	0.44	[0, 2]	0.2
	East	0.17	[0, 1]	0.1	0.16	[0, 1]	0.1
	Kyiv	0.29	[0, 1]	0.1	0.31	[0, 2]	0.1
	Odessa	0.80	[0, 2]	0.3	0.61	[0, 2]	0.3
	South	0.21	[0, 1]	0.1	0.25	[0, 1]	0.1
Total N		242.36	[228, 254]	100.0	241.45	[229, 254]	100.0

Region		Average % of migration events across all 10 sub-samples
Center	All exporting	10.6
	All importing	20.9
Crimea	All exporting	0.3
	All importing	4.3
East	All exporting	64.4
	All importing	6.7
Kyiv	All exporting	11.2
	All importing	19.0
Odessa	All exporting	1.9
	All importing	20.8
South	All exporting	10.8
	All importing	19.3
West	All exporting	0.9
	All importing	9.0

Appendix 4

(Chapter 4)

Accession numbers of sequences in the combined dataset

JN670059, DQ345495, KC238135, DQ328525, DQ328544, JX300449, DQ177224, EU248424, EF517426, JQ619609, AF112923, AF112925, FJ594168, FJ591487, JQ619569, HM533994, KR066204, KU820995, HE657455, KT167848, KT168171, EU693765, KC924507, KU050462, KT893490, GU345488, GU345191, FJ388930, AY694277, AY694373, KJ771746, KJ771370, KJ770931, KJ771085, JX299770, GQ400853, GQ400988, AJ419491, JX431654, KC340854, MF404092, HM460462, HM460461, KC340211, MF403567, MF403718, JX300128, JX301131, JX300410, GQ399180, K02013, FJ525771, L31963, FJ030710, JQ395622, KX661912, KX661944, KX662723, KX661920, KU142957, KU142995, JN101878, JN100840, D10112, FJ030791, JN101426, JN100813, KU143003, JN100946, JN100951, DQ205260, DQ205261, DQ205273, DQ205284, GQ400282, EU439774, LC163454, AB640129, AB867294, AB864148, AB865346, AB867545, AB866945, KC018507, AY166476, GQ399256, KY386756, GU945114, GU945117, GU945126, GU945144, DQ631416, KJ474627, EU910737, EU910746, KM283961, KF157442, KY857892, KY857902, KY857900, KY857905, KY857898, KU645850, KU645851, KJ722132, KJ499567, KX517389, KX446885, KX446886, EU345856, EU345857, EU345849, FJ822088, FJ822112, KX517402, FJ009810, KX446877, FJ822089, KC254614, KF971944, KF971924, GQ400343, GQ401007, KP013666, KP013673, AJ971104, AJ971099, AJ971093, KF745361, KF745574, AY173951, MF456662, MF456427, KY190134, GU560977, KF469812, KF469820, JX160372, KT167951, KT168154, KT168136, KT168157, KT168163, M38431, KT168119, KT168108, KJ704793, KJ704794, AY835777, AY835778, AY835749, AY835750, AY156801, AY156802, AY030876, AF514038, AY032504, FJ469749, EF136782, EU346342, EU346344, EF121398, AY030788, AY835755, KT167958, KJ722836, EU693719, KF470525, JX161104, FJ469734, FJ983392, KC312528, HM056593, JX160812, KC814288, KY037022, KX830506, AY845747, GQ202852

Appendix 5

(Chapter 5)

Accession numbers for the reference subtype G sequences

MF999247, MF999251, KY857895, KY275364, KT121457, KU685565, KX183809, MF434982, MF434974, MF434968, KP090090, KU128340, KX183780, KU685592, KU536034, KT276261, KU536014, KU536019, KU536011, KU535963, KT592479, KT592433, KT592430, KT592427, KT592426, KT592419, KT592416, KT592396, KT592392, KT592386, KY579520, KU536001, KU536000, KU535956, MF124864, KU535952, MF124844, KF745675, KF745360, KU535985, KU535942, KJ499636, KP231397, KP231395, KC755269, KC755265, KC755263, KC755262, KC755261, KC755260, KC755258, KC755257, KC755255, KC755253, KC755252, KC755250, KC755249, KC755247, KC755244, KC755241, KC755240, KC755239, KC755238, KU535938, KJ194662, KP231375, KY386751, KY386735, KY386722, KY386707, KY579776, KY579616, MF434855, MF434854, MF404967, MF404964, MF404957, MF404949, KF735836, KJ636019, KP231384, MF456983, KP231381, KX389648, KX389646, AB747543, KJ194688, KJ194660, KJ194654, KJ194650, KJ194645, KJ194643, KF751450, KT315999, KR047854, KR047853, KR047814, KT379606, KY386790, KY386724, KY386717, KY386695, KY579523, KX073711, KX073706, KX073703, KX073701, MF404959, MF404958, MF404956, MF404952, MF404948, KF241487, KF241488, KX467054, KF241484, KF241478, KX389645, KX389642, KX389641, KX465813, KP231391, KP231394, KJ936918, KC350318, KF241500, KC350301, KF241485, KF241480, KC350376, KF241505, JQ796143, JQ585483, AB747478, JX460764, JX460724, JX460714, JX460706, JX460654, JX460653, JX460648, JX460646, JX460634, JX460626, JX460623, JX460620, JX460599, JX460576, JX460573, JX460570, AB868710, AB864401, AB864400, KJ194657, KJ194655, KF751434, KJ771584, KJ771556, KP231378, KP688122, KT315985, KT315981, KT315970, KT315966, KY386798, KY386752, KY386747, KY386739, KY386703, KY386698, KY658701, MF404955, MF404954, KY953202, KY953201, KY953200, KX466989, KC350113, KC350111, KC350115, KC350329, KC350193, KF241481, KF241510, KF241502, KR229863, KF241483, KF241508, KF241501, KR229817, KC350110, HE657492, KC350181, JN673706, KF241504, KX389615, KF907218, KX389618, KF241509, KX389620, KX389619, KC350205, HQ655391, HQ655388, HQ845957, HQ845954, HQ845950, HQ845948, HQ845947, HQ845945, HQ845944, HQ845941, HQ845940, HQ845939, JN000022, JQ585425, JQ585415, JQ585434, JQ585392, JX140676, JX499018, KC018757, KC018758, KC340494, JX460541, JX460539, JX460398, KF751448, KF751446, KF751443, KF751442, KF751438, KF751433, KF751431, KF751430, KF751429, KJ771444, KP109502, KY386774, KY386771, KY386770, KY386769, KY386746, KY386675, KY386669, KY579502, KY579415, MF404966, KC350310, KC350338, KC350216, KJ936912, KC350320, JN673707, KX389625, KX389626, JN673705, KX389628, KX389631, KX389635, KX389636, KC350114, KC350109, KC350218, KC350214, KC350208, JN132378, JN132372, JN132370, JN132373, JN132371, JN132365, GU326176, GU326175, JN132377, GU326166, GU326165, JN673700, JN132367, HQ267657, HQ267655, HM002537, JN132368, GU362882, MF157743, KX465812, GU326159, KC971042, KF241460, HQ667706, HQ667698, HQ267634, HQ267631, JN248593, JN248591, JN248586, JN248584, HQ845937, HQ845936, HQ845935, HQ845934, HQ845932, HQ845930, JF803990, JF803987, JQ796107, JQ796146, JQ585317, JQ585346, JQ585340, JX140675, KC018675, KC340490, JX460363, JX460352, JX460310, JX460295, JX460233, HG529956, KF716477, KF192120, KF192119, KF192118, KF192117, KF192116, KC350112, KF751451, KF751449, KF751447, KF751445, KF751444, KF751441, KF751437, KF751436, KF751435, KF751432, KJ770943, HG518368, HG518365, HG518359, LT578246, LT578231, LT578226, LT578202, KX228801, KY386686, KY386676, MF404962, HQ267630, HQ267582, HQ267581, HQ267577, JN132350, JN132351, HQ267569, JN132355, HQ267617, HQ267546, HQ267539, KR233379, KF241463, FJ670530, HQ267607, KR233375, GU326110, GU326109, KF241470, KF241467, KF241475, KF241468, KF241474, FN557306, JX271487, JX301049, GU969570, GU969569, GU969547, HM135512, HM135511, HM135510, HQ864369, JN248582, HQ845928, HQ845927, HQ845926, HQ845921, FR715290, FR715281, FR715270, HQ541943, HQ541939, HQ541926, HQ541916, HQ541888, HQ995018, JN106043, JN704992, JN704987, JX983038, KC340465, KC340403, KC340399, KC340498, JX460174, JX460155, JX460150, JX460147, JX460135, JX460070, HG529938, HG529936, AB868307, AB863818, KP688139, KP718925, KU499411, KU499409,

KU499406, KU499403, KU499400, KU499395, KU499394, KU499393, KU499392, KU499390, KU499389, KU499385, KU499384, KU168300, KY386673, KY305747, KY579808, KX465925, JX301099, JX300797, JQ361672, JX300265, JX300208, KF241495, EU741670, JX300581, JX300044, EU255518, EU545193, FJ659678, GQ241001, GQ241000, GQ241119, GQ241117, GQ241116, GQ241110, GQ131619, GQ131618, GU191666, GU191639, GU191633, GU191597, GU191553, GU969517, FR666644, FR666643, FR666630, FR666627, FR666616, HQ843680, HQ843679, HQ843677, HQ843676, HQ843674, HQ843668, HQ843667, HQ843661, HQ843657, HQ843655, HQ843652, HQ843650, HQ843648, HQ843643, HQ843640, HQ843638, HQ843637, HQ843635, HQ843634, HQ843633, HQ843632, HQ843630, HQ843625, HQ843624, HQ843622, HQ843621, HQ843620, HQ843619, HQ843618, HQ843615, HQ843613, HQ843612, HQ843611, HQ843610, HQ843604, HQ843603, HQ843598, HQ843597, HQ843594, HQ843593, HQ843592, HQ843587, HQ843585, HQ843584, HQ843582, HQ843579, HQ843578, HQ843573, HQ843571, HQ843566, HQ843565, HQ843561, HQ843558, HQ843557, HQ843552, HQ843551, HQ843550, HQ843548, HQ843547, HQ843546, HQ843545, HQ843541, HQ843540, HQ843539, HQ843537, HQ843535, HQ843534, HQ843531, HQ843530, HQ843524, HQ843523, HQ843516, HQ843511, HQ864368, HQ864367, HQ864366, HQ864365, JF505618, JF929084, JF929058, HQ994227, HQ994159, HQ994149, HQ994139, HQ994125, HQ994123, HQ994115, HQ994095, HQ994085, HQ994079, HQ994075, HQ994057, HQ994049, HQ994037, HQ994031, HQ994019, HQ994017, HQ994011, HQ993987, HQ993979, HQ993709, JN639458, HQ994160, HQ994158, HQ994150, HQ994146, HQ994144, HQ994130, HQ994114, HQ994104, HQ994098, HQ994094, HQ994056, HQ994052, HQ994046, HQ994042, HQ994036, HQ994030, HQ994026, HQ994018, HQ994016, HQ994014, HQ994010, HQ994004, HQ994000, HQ993996, HQ993994, HQ993986, HQ993982, HQ993824, HQ993788, HQ993704, JN639459, JN639457, JN630903, JQ480201, JQ480199, JQ480195, JQ480191, JQ480206, JQ480202, JQ480200, JQ480198, HQ170613, KC340398, JX460039, JX460021, HG780854, KJ723366, KJ770596, KP178442, KP688105, KP718923, KR017776, KM248760, KU499410, KU499398, KY305772, KY579481, KY579444, MF404965, MF404963, MF404960, MF404953, MF404951, HM067749, JX300473, KC517039, KC516874, GU991882, GU991880, GU991875, GU991874, GU991869, GU991863, GQ167754, GQ167751, GQ167739, JX301141, EU330723, EU330714, FJ688263, JX300455, JX300616, JX301013, AM886216, JX300249, JX300970, GU991883, GU991881, GU991878, GU991870, GU991868, GU991865, GU991864, EU255477, EU255391, EU342797, GQ167766, GQ167763, GQ167757, GQ167753, GQ167737, GQ167732, GQ167731, FM955738, FM955734, FM955721, FM955716, FM955715, FM955712, FM955711, FM955702, FM955700, FM955693, GQ240996, GQ240992, GQ241105, GQ241097, GQ241056, GU191677, GU191665, HQ843536, HQ843522, HQ843517, HQ843509, JF929074, JN639446, KC340589, JX459988, JX459982, KJ906889, KP718915, KT022379, MF404950, HQ896682, HQ896681, HQ896680, KX465924, EU099972, EU099971, HQ896678, HQ896677, KX465811, GQ167760, GQ167750, GQ167742, FJ670520, EU099973, EU099970, EU786670, EU255345, EU255342, EF199585, EF199560, GQ240947, GQ241048, GQ400368, GQ400213, GQ400130, GQ400115, GQ399947, GQ399924, GQ399797, GQ399643, GQ399521, GQ399404, GQ399365, GQ399357, GQ399108, GQ398880, GQ398866, EU715229, FN599788, JF929089, JX140674, KJ770152, KJ770080, KX228812, EU741669, GU991879, GU991877, GU991876, GU991872, GU991871, GU991867, GU991866, GQ167765, GQ167762, GQ167759, GQ167756, GQ167744, GQ167738, GQ167735, GQ167733, HM460489, HQ896676, HF586667, HF586670, FM210721, FM210708, FM210706, FM210704, FM210697, FM210696, FJ389366, FJ389365, FJ389364, FJ389363, GQ167729, GQ400708, GQ400493, GQ400469, GQ400421, GQ400393, GQ400323, GQ400290, GQ400287, GQ400225, GQ400218, GQ400215, GQ400171, GQ400133, GQ400126, GQ400111, GQ400083, GQ400016, GQ399828, GQ399808, GQ399758, GQ399678, GQ399661, GQ399349, GQ399283, GQ399239, GQ399192, GQ399178, GQ399022, GQ399006, GQ398969, GQ398892, GQ398857, GQ398851, GQ398850, FJ481667, HM135509, HM135507, HM135506, HM135505, HM135502, HM135501, HM135500, HM135499, HM135497, HM135496, HM135495, HM135494, HM135493, HM135492, HM135489, KM248762, KY386688, DQ878227, DQ879086, DQ878218, DQ878533, AB287004, AB287003, DQ877755, AB231893, GU991873, GQ167764, GQ167758, GQ167748, GQ167746, GQ167741, GQ167736, GQ167730, GQ167727, DQ879056, DQ877867, DQ878322, EU615561, EU612844, EU612202, EU248481, EU248466, EU248445, EU248413, EU248374, EU248347, EU248320, EU248547, EF174563, DQ113308, DQ113206, DQ113189, DQ113133, DQ113132, DQ113103, DQ113089, DQ113067, DQ113061, DQ113056, DQ113052, DQ113043, DQ113034, AM181796, GQ400723, GQ400471, GQ400448, GQ400386, GQ400357, GQ400356, GQ400252, GQ400238, GQ400235, GQ400223, GQ400191, GQ400076, GQ400013, GQ399962, GQ399921, GQ399908, GQ399843, GQ399842, GQ399830, GQ399810, GQ399783, GQ399671, GQ399618, GQ399441, GQ399375,

GQ399358, GQ399331, GQ399327, GQ399279, GQ399262, GQ399225, GQ399196, GQ399121, GQ398994, GQ398983, GQ398968, GQ398914, GQ398906, GQ398852, GQ398828, GU207222, HM135488, HM135487, HM135485, HM135483, HM135482, HM135481, HM135480, HM135479, HM135478, HM135477, HM135476, HM135475, HM135474, HM135473, HM135472, HM135471, HM135470, AB640341, JQ698691, KU168278, KU168277, KU168302, GQ167761, GQ167755, GQ167752, GQ167749, GQ167747, GQ167745, GQ167743, GQ167740, GQ167734, GQ167728, GU199524, EU178067, EU178026, EU178021, EU178002, EU608993, DQ009061, AM260319, EF122523, AM041046, AM041043, AM041028, AM041012, AM041000, AM181795, AJ578178, AJ577914, KJ906677, AM285285, AJ583750, AY371121, DQ168579, DQ168576, DQ168575, DQ168573, DQ013282, DQ013280, DQ013279, DQ013278, DQ013275, DQ013272, DQ013270, DQ013269, AJ577968, AJ577913, AJ577879, AJ578174, FJ389367, GU207114, GU207087, FN599705, KJ948662, KY579463, AJ313412, AJ313401, AJ313393, AY366535, AF423760, DQ445117, AJ419501, AJ419474, AY377436, AY694317, AF347293, GQ241129, KM057330, EU345865, AY586549, AY586548, AY586547, AM279386, AM279378, AM279377, AF450098, AY181082, AY181081, AY181080, AY181078, AY181076, AY181075, AY181074, AY181073, AY181072, AY181071, AY181070, AY181069, AY181068, AJ577967, AY017459, AY017458, AY017457, KM057328, KX465243, AF295287, AM279392, AJ286982, AJ286965, AJ286964, AF212297, AM279373, AY492756, AJ270559, AJ286963, AY772535, AF084936, FJ030730, AF212296, DQ009070, AF061641, AF061640, AF447856, AF061642, AB485663, AB485662, U88826, AF212295