THE UNIVERSITY
of EDINBURGH

TRANSMISSION NETWORKS INFERRED FROM HIV SEQUENCE DATA

Manon Ragonnet-Cronin

This dissertation is submitted for the degree of Doctor of Philosophy at the Institute of Evolutionary Biology,

University of Edinburgh

October 2015

DECLARATION

This thesis is submitted to the University of Edinburgh in accordance with the requirements for the degree of Doctor of Philosophy in the faculty of Science. I declare that this thesis is my own composition and that the work described herein is my own, except where explicitly stated below. This work has not been submitted for any degree or professional qualification.

Signed:_____

Date:_____

Manon Ragonnet-Cronin

Lay Summary of Thesis

The lay summary is a brief summary intended to facilitate knowledge transfer and enhance accessibility, therefore the language used should be non-technical and suitable for a general audience. (See the Degree Regulations and Programmes of Study, General Postgraduate Degree Programme Regulations. These regulations are available via: http://www.drps.ed.ac.uk/.)

Name of student:	Manon Ragonnet-Cronin	UUN	S1136422
University email:	Manon.ragonnet@ed.ac.uk		
Degree sought:	PhD	No. of words in the main text of thesis:	39537
Title of thesis:	TRANSMISSION NETWORKS INFERRED FROM HIV SEQUENCE DATA		

Insert the lay summary text here - the space will expand as you type.

HIV can be transmitted through sexual contact (heterosexual or homosexual), through sharing needles and during childbirth. 100,000 people are infected in the UK and around 45% of these are men who have sex with men (MSM), 45% are heterosexuals and ~5-10% are people who inject drugs.

HIV displays huge genetic diversity: numerous subtypes exist and the virus is distinct in between even transmission-recipient pairs. If viruses sampled from different individuals are highly genetically similar, those individuals may have infected each other or be associated with the same transmission chain. In the UK, many HIV genetic sequences are available because these are generated as standard practice during medical treatment. The relationships between all UK viruses can be reconstructed in the form of large trees of life, or phylogenies, to assess the probability of transmission events at the national level. These trees contain thousands of sequences, but within them, “clusters”, or groups of related sequences, are of particular interest because they contain sequences associated with the same transmission chain.

In Chapter 3, I develop software tools that identify clusters within large phylogenies automatically and describe them in terms of risk group. In Chapter 4, the tools are used to analyse and compare the UK and Swiss HIV epidemics. Switzerland generates HIV sequences from Swiss patients in the same way as is done in the UK and the two countries have the two largest HIV sequence databases in the world. I found clustering patterns in the two countries to be broadly similar, although the UK epidemic comprised more men who self-identified as heterosexuals yet clustered with MSM, suggesting that they are more likely to have been infected through sex with men.

In Chapter 4, I analysed the transmission patterns of HIV subtypes that have entered the UK only recently. Subtype B predominated until the 1990s, but non-B subtypes now circulate and account for 60% of new diagnoses each year. In contrast in the rest of Europe, subtype B still represents >80% of infections. These non-B subtypes infect mostly heterosexuals. I found that these clusters were more likely to be smaller than subtype B clusters among MSM; and between 2007 and 2010, the non-B clusters were less likely to grow. However, a small proportion of non-B clusters comprised MSM and these clusters were more likely to grow. It thus appears that crossover of non-B subtypes into MSM has caused expansion of those subtypes within the UK.

In Chapter 5, I use simulations of HIV epidemics to evaluate how well identified clusters match true transmission chains.

Together with other evidence, this thesis demonstrates that the UK HIV epidemic continues to be driven by transmission among MSM. New subtypes historically associated with heterosexual transmission now circulate among MSM, most likely due to MSM also having sex with women.

ABSTRACT

HIV in the UK in the 1980s was concentrated within men who have sex with men (MSM) and people who inject drugs (PWID) but heterosexual sex is now the most frequently reported risk behaviour. As these risk groups are associated with different virus populations, this is reflected in the subtype diversification of the UK epidemic, which was historically dominated by subtype B.

I have made use of a national database of HIV sequences collected during routine clinical care, which also contains data on age, sex, route of exposure & ethnicity. The 2014 release of the UK HIV Drug Resistance Database contained data from over 60,000 patients.

In this thesis, I first describe the development of novel tools that rapidly and automatically identify HIV clusters within phylogenetic trees containing tens of thousands of sequences because they represent transmission chains within the larger infected population.

I use these tools to compare the HIV subtype B epidemics in the UK and Switzerland, which had both been described separately but using different approaches. Working with Swiss colleagues, I was able to analyse the epidemics in exactly the same way without having to share sensitive data. I found clustering in the UK to be much higher at relaxed thresholds than in Switzerland (34% vs 16%) indicating that the UK database is more likely to capture transmission chains. Down sampling revealed that this pattern is driven by the larger size of the UK epidemic. At tighter cluster thresholds, the epidemics were very similar.

I next use these tools to analyse the spread of emerging subtypes A1, C, D and G in the UK. I found both risk group and cluster size to be predictive of cluster growth, which I tested using simulations and a GLM. Growth of MSM and crossover clusters was significantly higher than expected for subtypes A1 and C, indicating that crossover from heterosexuals to MSM has contributed to their expansion within the UK.

Numbers were small for subtypes D and G but the proportion of new diagnoses linking to MSM and crossover clusters was similar to A1 and C, suggesting that the same pattern may be emerging for D and G.

I conclude by evaluating the accuracy of a method previously described by our group to generate transmission networks from HIV sequences. The interpretation of clustering patterns from phylogenetic trees is difficult because of the absence of a standardised statistical framework. In contrast, a body of work exists that relates disease transmission to networks. Using large simulated datasets, I developed algorithms which eliminate improbable links. I then reconstructed improved UK transmission networks for subtypes A1, B and C and compare network metrics (such as the degree distribution) between risk groups.

Together with other evidence, this thesis demonstrates that the UK HIV epidemic continues to be driven by transmission among MSM. The UK epidemic is no longer compartmentalised and the crossing over of subtypes across risk groups has been facilitated by MSM also having sex with women.

PUBLICATIONS ARISING FROM THIS THESIS

Ragonnet-Cronin M., Hodcroft E., Hue S., Fearnhill E., Delpech V., Leigh Brown A. J. and Lycett S. (2013). Automated analysis of phylogenetic clusters. *BMC.Bioinformatics.* **14**, 317.

This paper forms the basis of Chapter 3.

Ragonnet-Cronin M., Lycett S., Hodcroft E., Hue S., Fearnhill E., Brown A. E. , Delpech V., Dunn D. and Leigh Brown, A. J. Transmission of non-B HIV subtypes in the UK is increasingly driven by large non-heterosexual clusters. (*submitted, JID*).

This paper forms the basis of Chapter 5.

During my PhD, I have also contributed to the following publications:

Yebra G, **Ragonnet-Cronin M**, Ssemwanga D, Parry CM, Logue CH, Cane PA, Kaleebu P and Leigh-Brown AJ. Analysis of the History and Spread of HIV-1 in Uganda using Phylodynamics. *J Gen Virol* 2015.

Birungi J, Min JE, Muldoon KA, Kaleebu P, King R, Khanakwa S, Nyonyintono M, Chen Y, Mills EJ, Lyagoba F, **Ragonnet-Cronin M**, Wangisi J, Lourenco L and Moore DM. Lack of Effectiveness of Antiretroviral Therapy in Preventing HIV Infection in Serodiscordant Couples in Uganda: An Observational Study. *PLoS ONE* 2015; **10(7)**:e0132182.

Hue S, Brown AE, **Ragonnet-Cronin M**, Lycett SJ, Dunn DT, Fearnhill E, Dolling DI, Pozniak A, Pillay D, Delpech VC and Leigh Brown AJ. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 2014; **28(13)**:1967-1975.

AUTHOR CONTRIBUTIONS

Chapter 3

Dr. Samantha Lycett and Dr. Emma Hodcroft designed and wrote the software. I participated in the initial design of the tools and carried out all the analyses and testing.

Chapter 4

The work described in this Chapter was planned as a collaborative comparison of two national datasets to which access was restricted. Accordingly, processing of Swiss data was done by Mohaned Shilaih at University Hospital Zurich. The logistic regression was carried out by Mohaned.

Modification of the Cluster Picker code was done in collaboration with Dr. Samantha Lycett.

Chapter 6

Dr. Samantha Lycett provided the base code for the Discrete Spatial Phylo Simulator, and the original code to reconstruct networks (which was edited to return tMRCA)

Dr. Emma Hodcroft wrote the HIV simulator and provided model output for analysis.

ACKNOWLEDGEMENTS

I have immensely enjoyed my PhD experience. I have been extremely fortuitous to have a remarkably supportive supervisor, Andrew Leigh Brown. Andy has built my confidence in my abilities by pushing me when I needed to work harder and supporting me when I had doubts. I am grateful for his encouragement to present my work at conferences incessantly and for the kindness he has shown in introducing me to everyone he knew (which is a lot of people) at all those conferences. Despite the delays this caused to me finishing, he supported me fully in endeavours that interfered with my PhD, like going to Uganda and interning for the UK Government. I have felt that I could go to him for advice both professional and personal and I hope he's willing to continue to fulfil that role as I progress through my career.

My PhD sister, (Dr.) Emma Hodcroft, has been with me every step of the way. Every meeting, conference, work dinner, hotel room and train ride has been made fun by having her with me. We have travelled the world together and made the most of every trip. I also have to thank her profusely for running RaxML for me at least 724000 times and for building a fantastic HIV epidemic simulator.

Samantha Lycett is an expert problem-solver and programmer. I have felt reassured throughout my PhD that if I really couldn't figure something out, I could always go to Sam. She is a formidable researcher and mine of information who must be bothered sparingly but seems to know all the answers.

All the members of our lab group have taught me how to use new programs, shared scripts with me, helped me with my presentations, and many times listened to seemingly senseless streams of thought until I could make sense of them. Gonzalo Yebra, Melissa Ward, Lu Lu and Mojca Zelnika, it has been amazing working with you and sharing offices with you and all of you have contributed to my work over the years. I've also had the benefit of having a second research group who opened me to the completely new field of network analysis. My co-supervisor Rowland Kao has helped direct my work and discussions with his research group have been extremely intellectually stimulating.

I was incredibly lucky to share an office with Darren Obbard and Gytis Dudas for the first two years of my PhD. Both have given me endless advice and help with programming and statistics, some concrete (Darren) and some more wacky (Gytis). Darren thank you for continuing to offer me useful and straightforward R and statistics advice even after I left the office, including on weekends and over the phone. I intend to continue pestering you.

I have been lucky to have access to one of the best HIV databases in the world to do my PhD work. None of the work in this thesis would have been possible without the clinicians, technicians, public health analysts and bioinformaticians who contribute to and maintain the UK HIV database. In particular I appreciate the guidance and insights offered by David Dunn, Esther Fearnhill, Stephane Hue, Anna Tostevin, Valerie Delpech and Alison Brown, on issues as large as the UK HIV epidemic and as small as the database process of de-duplication.

Ashworth is a fabulous place to work because of its friendliness and quirky traditions and because for the inspiring minds that walk the corridors. I would particularly like to extend my gratitude to Matthew Hall, Andrew Rambaut, Trevor Bedford and other members of Virus Club. I have made wonderful friends in Ashworth who have made my PhD such an enjoyable experience, including Hannah Froy, Reuben Nowell, Sam Lewis, Sarah Mattey, Elisa Anastasi, Jess Flood, Kevin Donnelly, Lucy Carter and Laura Ross.

I am grateful to Anna van Weringh, my personal copy editor, who has now read both my MSc and PhD theses with tremendously well feigned enthusiasm. I thank my parents for their love and encouragement do whatever I want. I am exceptionally fortunate to have received unflagging support from my husband and rock, Dave.

I'd like to thank my examiners, Andrew Rambaut and Simon Frost for their time and their valuable input. Finally, I would like to acknowledge funding from the Biology and Biotechnology Science Research Council and the Bill & Melinda Gates Foundation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Abbreviations and Acronyms

AIDS	Acquired Immunodeficiency Syndrome
ART	Anti-Retroviral Treatment
AUC	Area Under the Curve
AZT	Azidothymidine
CH	Confoederatio Helvetica (Switzerland)
CI	Confidence Interval
CM	Cluster Matcher
CP	Cluster Picker
CRF	Circulating Recombinant Forms
CTL	Cytotoxic T-Lymphocytes
DDQC	Degree Distribution Quantification and Comparison
df	Degrees of Freedom
DRM	Drug Resistant Mutations
DSPS	Discrete Spatial PhyloSimulator
FN	False Negative
FP	False Positive
GLM	Generalised Linear Model
HAART	Highly Active Anti-Retroviral Therapy
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
KS	Kolmogorov-Smirnov
LANL	Los Alamos National Laboratories
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MSM	Men who have Sex with Men
MTCT	Mother to Child Transmission
NATSAL	National Survey of Sexual Attitudes and Lifestyles
NB	Negative Binomial
NNRTIs	Non-Nucleotide/Nucleoside Reverse Transcriptase Inhibitors
NRTIs	Nucleotide/Nucleoside Reverse Transcriptase Inhibitors

PIs	Protease Inhibitors
PR	Protease
PreP	Pre-exposure Prophylaxis
PWID	People Who Inject Drugs
ROC	Receiver Operating Characteristic
RT	Reverse Transcriptase
SDRM	Surveillance of Drug Resistance Mutations
SIR	Susceptible-Infected-Recovered/ Removed
SIV	Simian Immunodeficiency Virus
STI	Sexually Transmitted Infection
tMRCA	time to Most Recent Common Ancestor
TN	True Negative
TP	True Positive
UK	United Kingdom
UK HIV RDB	UK HIV Drug Resistance Database
USA	United States of America
WHO	World Health Organisation

List of Appendices

1 TRANSMISSION AND EVOLUTION OF HIV

1.1 The HIV Epidemic

1.1.1 Global distribution of HIV

The Human Immunodeficiency Virus (HIV) is the causative agent of Acquired Immunodeficiency Syndrome (AIDS). AIDS was first recognized in 1981, when previously healthy young men were diagnosed with unusual infections that suggested their immune systems were malfunctioning [1]. The AIDS-causing virus was isolated first in France in 1983 [2] then later that year in the United States of America (USA) [3], establishing a definite link between the virus and AIDS. The appellation "HIV" was internationally adopted in 1986 [4].

The immune system of people living with HIV progressively fails until they are no longer able to fight off other infections and die from AIDS. The number of people living with HIV continues to rise. In 2013, 35 million people were infected, 2.1 million people became newly infected and 1.5 million died from AIDS [5]. The World Health Organization (WHO) considers HIV a global pandemic but infections are unequally

distributed across the globe. Sub-Saharan Africa is the most severely affected region, with 67% of infections, or 24.7 million people living with HIV. HIV is a leading cause of death in a number of countries, leaving millions of children orphaned. AIDS patients are unable to work and require significant medical care. As such HIV is a major social and economic burden. In addition, HIV is a highly stigmatised condition due to its associations with homosexuality, drug use and race. Thirty years into the epidemic, HIV has no cure or vaccine and continues to be a major global heath priority.

1.1.2 HIV in the UK

In the UK in 2013, an estimated 107,000 people were living with HIV, a quarter of whom were unaware of their infection [6]. At the beginning of the HIV epidemic in the 1980s, HIV was concentrated within men who have sex with men (MSM) and people who inject drugs (PWID). Harm reduction and needle exchange programs have decreased HIV incidence in PWID substantially since then. Since 1999, heterosexual sex has become increasingly frequently reported as the risk behaviour for HIV acquisition. Currently, MSM and heterosexuals account for around 45% each of prevalent cases and PWIDs for 5% [6].

1.1.3 Origin of HIV

HIV arose by zoonotic transfer of simian immunodeficiency virus (SIV) from non-human primates. A number of cross-species transmission events have led to HIV's current genetic diversity [7, 8]. HIV-1, discovered first, is closely related to SIV in chimpanzees (SIVcpz) [9]. HIV-2 was isolated from West African patients with AIDS in 1986. HIV-2 is related to HIV-1 but has distinct antigenic components and evolved from SIV in sooty mangabeys (SIVsm) [9] (Figure 1.1). HIV has been classified according to genetic similarities: HIV-1 contains four groups, M, N, O and P, while HIV-2 separates into eight, A to H [10]. HIV-1 group M is the pandemic-causing strain, and it comprises a further 11 major subtypes (A1, A2, B, C, D, F1, F2, G, H, J, K) as well as an increasing number of circulating and unique recombinant forms (CRF, URF) [11]. The HIV-1 groups M, N, O and P are each individually closer to SIV

isolates than they are to each other and so appear to have originated from distinct cross-species events [8] (Figure 1.1).

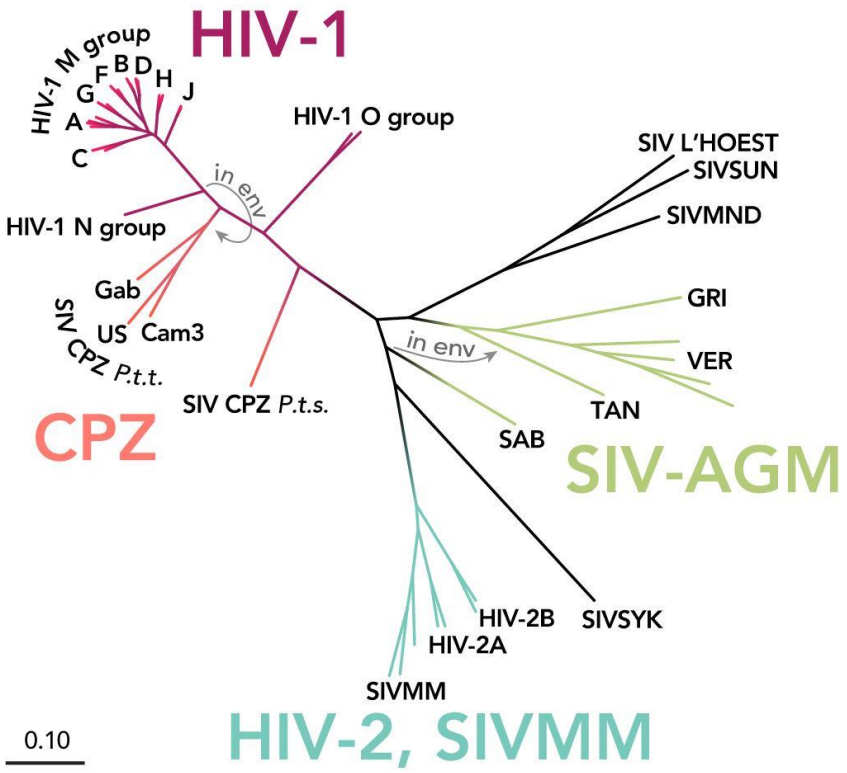


Figure 1.1: Primate lentivirus phylogenetic relationships based on the *pol* region. Scale in nucleotide substitutions per site. By Thomas Splettstoesser and based on Kuiken *et al.* [12], via Wikimedia Commons.

Thirty nine species of non-human primates possess a strain of SIV [13]; however the virus does not normally cause serious illness in its natural host (for exceptions, see [14, 15]). This resistance to disease has been shown to arise in part due to a down-regulation of immune activation in naturally SIV-infected cercopithecoid monkeys [16] and points to a long-term adaptive co-evolution, in contrast to the recent emergence and high pathogenicity of HIV in humans and rhesus macaques. Modern methods of analysing ancient samples support the hypothesis that HIV-1 group M was introduced into human populations around 1900 [17] in Kinshasa in the Democratic Republic of Congo [18, 19], the country displaying the highest genetic diversity of HIV. It is likely that transmission occurred through the practice of primate bushmeat hunting, through

exposure to infected blood due to animal bites or during butchering [8]. The oral polio vaccine theory suggested that HIV was the result of vaccination trials conducted in the Belgian Congo in the late 1950s; but this theory has been discredited. No HIV virus nor chimpanzee DNA were found in remnant batches of the vaccine [20], and molecular clock analyses (see section 2.2.2.2) place the common ancestor to circulating HIV strains before the 1950s [21, 22].

1.2 Molecular virology of HIV

1.2.1 HIV life cycle and disease progression

Two aspects of HIV biology make it a particularly challenging disease. Firstly, HIV directly attacks the host's immune system. HIV infects CD4+ T cells responsible for regulating immune responses, progressively compromising the host's ability to fight off infection. Secondly, infected hosts do not show any symptoms for 8 to 12 years without treatment but are still able to transmit the virus during this period of clinical latency.

HIV is classified as a member of the lentiviral genus in the *Retroviridae* family. Lentiviruses contain single-stranded positive-sense RNA genomes. The HIV particle is spherical and 120nm in diameter and each particle contains two copies of its full genome. The genome is ~10,000bp and consists of three main genes, *gag*, *pol*, and *env* and six accessory genes: *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu*, encoding a total of 19 proteins.

The major target cells of HIV are CD4+ T cells vital to the human immune system, but macrophages and dendritic cells can also be infected. Entry into the cell is dependent on binding to the CD4+ receptor and a co-receptor (one of the chemokine receptors CCR5 or CXCR4). After the viral capsid has entered the cell, the viral enzyme reverse transcriptase (RT) converts viral RNA in to double-stranded DNA (dsDNA). The viral protein integrase then integrates the dsDNA into the host cellular DNA, often within active transcription units [23]. Viral genes become highly expressed and infected cells produce and release numerous viral particles, triggering cytotoxic defences. Alternatively, the integrated virus can enter viral latency and lie dormant within the host DNA, expressed at low levels or not at all. Latent virus can

persist indefinitely in memory T-cells and this latent reservoir is a major barrier to curing HIV [24].

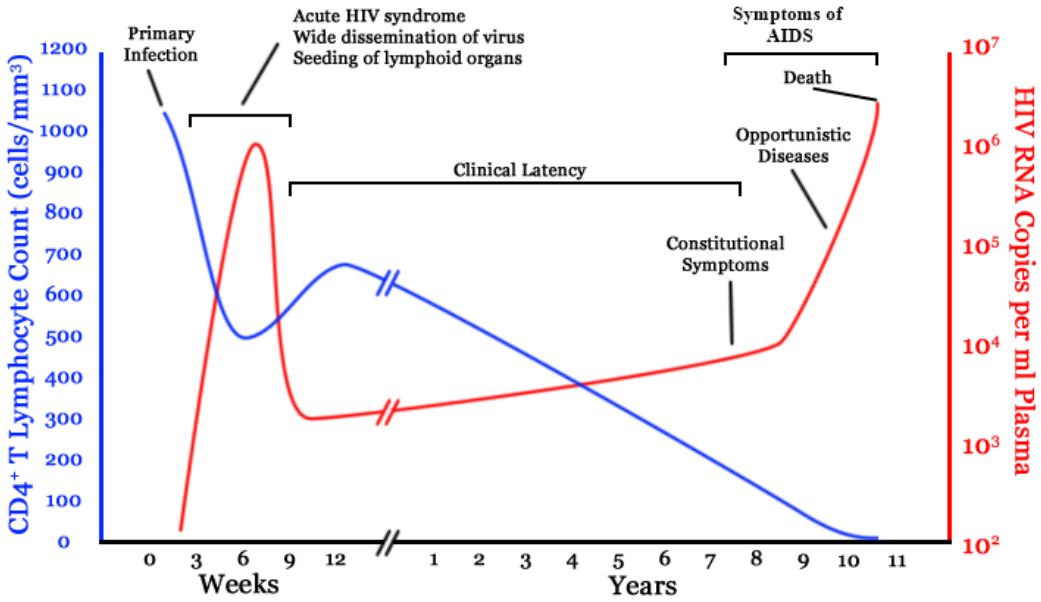


Figure 1.2: HIV course of infection. CD4+ counts and HIV copies in an untreated patient. By Jurema Oliveira, and based on Figure 1 in Pantaleo *et al.* [25]. Via Wikimedia Commons, GNU Free Documentation License.

HIV disease progression to AIDS is divided into three stages (Figure 1.2). During the first 2-4 weeks (acute infection) the virus replicates rapidly, occasionally causing generic symptoms of immune activation such as a fever, headache and sore throat. Extremely high viral loads (10^7 viral RNA copies/mL blood) are associated with high infectiousness during this stage of infection [26]. CD4+ counts falls and are never recovered. Initiation of host adaptive immune responses causes patients to enter the latent period of HIV infection. During the next 8-12 years the patient does not show any overt symptoms. At the start of latency the infection is controlled by the host immune system, but immunity fails over time in the absence of treatment. Viral loads increase gradually, correlating with the decline of CD4+ T cells in the peripheral blood and a progressive failure of the immune system. When CD4+ cell counts drop below a critical level, around 200 cells/mm^3, cell-mediated immune responses cannot

function. The AIDS stage has been reached and patients are no longer able to fight off opportunistic infections. Patients usually die within a year of developing AIDS [27]. Antiretroviral therapy (ART) prevents the virus from multiplying and destroying the host's immune system so that patients never progress to AIDS (see section 1.2.3).

1.2.2 Viral replication and molecular evolution

Many mutations happen during reverse transcription. The enzyme that synthesises complementary DNA from viral RNA, RT, is extremely error-prone and has no proofreading ability. One error is made every 1,000-10,000 bases causing 5 to 10 nucleotide substitutions per new HIV genome synthesized [28]. In addition HIV has an extremely high replication rate: in the absence of treatment 10^{10} virions are produced daily by infected individuals [29]. On top of this, two genomic copies are packaged into HIV virions and these can produce recombinants. During reverse transcription, both copies are used to produce double stranded DNA through a process of template switching [30]. Recombination occurs at a minimum rate of 2.8 crossovers per genome per cycle [31]. Occasionally, virions become coinfected with divergent strains leading to intra and inter-subtype (section 1.4.2) recombinant viruses [32, 33].

HIV thus displays high genetic variability both within a host and at the level of the population. High within-host variation allows selection of HIV variants able to evade human immune responses and resist HIV drugs. The extreme genetic diversity of circulating isolates contributes to the difficulties in producing a vaccine.

HIV undergoes a major bottleneck at transmission; in 60-80% of infections, a single founder virus is transmitted [18, 34-36]. Intra-patient genetic diversity then increases over the course of infection [37] as a consequence of both genetic drift and immune selection. Selection acts on the genetically diverse quasispecies as whole, allowing the viral population to rapidly adapt to its host. Viral escape mutants that reduce recognition and destruction by the immune system are strongly selected [38-40]. Neutralising antibodies drive the diversification of HIV envelope in early infection [41]. The most important genetic determinant of disease progression identified so far is the Human Leukocyte Antigen (HLA) [42, 43]. Some HLA alleles that confer a

notable protective effect against HIV disease, such as B57, are called 'super-controller' alleles [44]. Additionally, non-immune related selection is driven by ART which results in drug-resistant viruses (see section 1.2.4).

1.2.3 HIV treatment

Azidothymidine (AZT) was the first HIV drug approved by the US government in 1987. Since then, around 30 more drugs have been approved. Many stages of the viral life cycle cannot be targeted because they utilise host machinery and drugs targeting host machinery are harmful to the host. Existing drugs each target different aspects of the viral life cycle and are sorted into five classes based on their target. Nucleotide Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleotide Reverse Transcriptase Inhibitors (NNRTIs) both block the viral RT (see section 1.2.2). Protease Inhibitors (PIs) prevent the protease enzyme from cleaving newly synthesised proteins to make them functional. Integrase inhibitors block double-stranded viral DNA from integrating into the host genome. Fusion inhibitors prevent the virus from docking and entering into new cells. All drugs are used in combinations of 3 drugs from 2 or more classes and can fully suppress replication of wild type virus below detectable levels. Prescribing drugs in combination (Highly Active Anti-Retroviral Treatment or HAART) has changed HIV to a chronic disease: viral loads are lowered, optimally to undetectable levels, and CD4+ cell count increases. It has been shown that with optimal therapy the life expectancy of an infected individual lies within the normal range [45]. At present, these drugs prevent disease but do not cure HIV as they are not able to remove integrated latent provirus (section 1.2.1).

1.2.4 Drug Resistance

Incompletely suppressed virus can lead to mutations arising in targeted enzymes that allow the virus to develop resistance to treatment. The opportunity to develop resistance is reduced as HIV replication is more efficiently blocked and as HAART significantly decreases replication, drug resistant mutations (DRM) are far less likely to emerge and be selected for. However, viral rebound, for example as a consequence of incomplete adherence to drug regimens, can favour the emergence of DRM.

Widespread use of ART has led to the development of DRM in circulating strains. Drug resistant strains can be transmitted (transmitted drug resistance or TDR) and can considerably reduce the efficacy of first line therapies. Resistance exists to all available classes of drugs but mutations are most common in the *pol* region of HIV because drugs that have been available the longest target protease (PR) and RT. Because of increasing prevalence of TDR [46, 47], genotyping (sequencing of the *pol* region) is recommended for all newly diagnosed individuals and is routinely performed in a number of countries [48-50]. The identification of specific DRM allows for tailoring of drug regimens. In the UK the virus is genotyped to detect drug resistance mutations before the initiation of therapy and in the case of immune failure, so that drug regimens can be tailored (see section 2.1.2).

1.3 HIV transmission and prevention

HIV is transmitted mainly through sexual intercourse. HIV can also be transmitted through contact with infected blood and from mother to child (MTCT). In sub-Saharan Africa, heterosexual sex is the dominant route of transmission while in Europe it is sex between men. In Asia, the use of contaminated needles has emerged as a driver of the epidemic. In the USA, both sex between men and needle sharing are major risk factors. Accurate estimates of per act transmission are essential for designing HIV interventions and for mathematically modelling the HIV epidemic (see section 1.5).

1.3.1 Sexual transmission of HIV

During intercourse, the receptive partner is at higher risk for HIV than the insertive partner. The risk per (unprotected) sexual act in developed countries has been calculated in meta-analyses as highest for receptive anal intercourse: 1.38% (95 CI 1.02-1.86%), 0.11% for insertive anal intercourse, 0.08% for receptive vaginal intercourse and 0.04% for insertive vaginal intercourse [51].

Original estimates for transmission were derived from longitudinal analyses of heterosexual sero-discordant couples in Uganda [52] and extrapolated to MSM and high resources settings. Since then a number of studies have been conducted in heterosexuals and MSM in the developed world (reviewed in [51]) and in low resource

settings (reviewed in [53]), in the pre and post-HAART eras. In low resource settings, calculated estimates 2-8 times higher [53].

MSM have a higher risk of contracting HIV than heterosexuals because of increased risk during anal intercourse, the higher prevalence of HIV among MSM, sexual role versatility and possibly increased risk-taking behaviour in this group [54].

Condoms prevent sexual HIV transmission [55] and have been available since the beginning of the epidemic but lack of compliance have limited their ability to curb the epidemic. The risk of sexual transmission of HIV is increased in the presence of other sexually transmitted infections (STI) [56]. Genital ulcer disease, for example, increases HIV transmission risk fivefold [53]. Thus, STI treatment is an effective prevention strategy [56]. Additionally, circumcision reduces the risk of heterosexual HIV acquisition in men [26] and voluntary medical male circumcision programmes have been initiated in 14 eastern and southern African countries.

HIV transmission risk is highly dependent on viral load [57]. In turn, viral load is dependent on a number of factors, including infection stage (see section 1.2.1) [58]. High viral loads during the first few weeks of infection may lead to a large proportion of transmissions originating from recently infected patients [26]. ART (see section 1.2.3) suppresses viral loads and thus decreases HIV transmission [59, 60]. Treatment and viral load monitoring are thus key prevention strategies, coining the expression “treatment as prevention”. Moreover, clinical trials have demonstrated that HIV acquisition can be decreased by the prophylactic use of antiretrovirals [27]. The use of pre-exposure prophylaxis (PrEP) is currently supported by guidelines in the USA but not in the UK. HIV testing is considered an important component of HIV prevention, particularly in the context of such high numbers of undiagnosed cases. It is possible that increased transmission from patients with recent infections is in part due to transmitters not being aware of their HIV positive status. HIV testing enables patients to make informed decisions about safer sex and HIV treatment, thus reducing their likelihood of transmitting the virus.

1.3.2 Transmission of HIV through needle sharing

Exchange of blood through sharing needles among PWID is a major route of HIV transmission. Globally, 10% of new infections occur through PWID, although if sub-Saharan Africa is excluded, this surges to 30% [61]. PWID, their partners and children account for one third of AIDS cases in the USA [62]. Infectivity per act for needle sharing is 0.63% (95% CI 0.41%-0.92%) [51], and with PWID engaging in hundreds of injections every year, lifetime risk is very high. The estimate was calculated for subtype B among PWID cohort in Thailand and is considered to the most relevant to the current US epidemic.

The transmission of HIV through injection drug use can be completely prevented by the provision of clean needles. The size of the HIV epidemic among PWID in different countries is a direct consequence of that country's policy towards harm reduction and legislation surrounding drug use. Needle exchange programmes have drastically reduced the incidence of HIV in countries where they exist, including the UK.

1.3.3 Iatrogenic transmission of HIV

Iatrogenic transmission is the transmission of infectious agents in a medical setting. The estimated risk of HIV transmission is highest for blood transfusions; in a retrospective analysis, 90% of HIV+ blood recipients became infected [51]. Early in the epidemic, 14,000 people became HIV positive through infected blood in Europe and the USA [63], primarily haemophiliacs [64]. Testing of donor blood has practically eliminated this route of infection in countries where it is implemented. In Africa 6% of infections are estimated to be the result of blood transfusions [65].

The re-use of injection equipment in hospitals is estimated by the WHO to contribute 1.6% of new infections in Africa [65] but this number is disputed, with one research group claiming that up to half of infections are linked to medical exposure [66]. Risk per exposure is estimated at 0.23% (95% CI 0-0.46%) [51].

1.3.4 Mother to child transmission (MTCT) of HIV

MTCT can occur at three stages: during pregnancy, during delivery or through breastfeeding. In the absence of treatment, the total risk across these stages ranges from 15% to 45% and has been estimated at 22.6% in developed country settings [67]. In 1996, a placebo-controlled trial with AZT reduced MTCT to 7.6 % [67]. Access to antenatal combination ART reduced MTCT to 0.46% in the UK in 2010-2011 [68]. MTCT approached 0 when the mother's viral loads were controlled. This year, Cuba became to first country to have eliminated MTCT of HIV. Globally, only 57% of pregnant women living with HIV received ART in 2011 and MTCT accounts for 10% of new infections annually (http://apps.who.int/gho/data/node.main.627).

1.4 Phylogenetic analysis of HIV

1.4.1 Rationale

The accumulation of HIV mutations occurs on a time scale similar to that of the ecological processes to which it is subject, for example spatial dynamics, changes in virus population size (due to epidemic spread and transmission bottlenecks) and host immune pressure. HIV phylogenies reconstructed from viruses sampled over time can be used to infer the evolutionary and transmission history of the pathogen [69-71]. HIV phylogenetics was initially used to confirm transmission pairs in contact investigations, the first of which established the transmission of HIV from a dentist in Florida to a number of his patients in 1992 [72]. More recently, viral linkage has been used in clinical trials that evaluate whether treating the infected partner decreases their risk of transmitting HIV to their HIV-negative partner [73]. In the event of the partner of a treated patient becoming infected, phylogenetic analysis is used to test whether the infection may have come from outside the partnership. Erroneously attributing the transmission to the treated partner would underestimate the efficacy of the intervention.

In addition to phylogenies informing us on the relationships between samples, methods have been developed that incorporate sample dates to estimate the timing of events in the tree (the molecular clock, see Methods) [74]. The purpose of the molecular clock

is to correlate genetic distance with calendar time. Demographic reconstructions can also be enhanced by the inclusion of geographic parameters into analyses [75]. Therefore sequentially sampled viral isolates can reveal past population dynamics, such as dates of introduction of the virus into different populations, population size (incidence) through time and space, dispersal and epidemic growth rate and doubling time [76]. Together, phylogenetic and phylogeographic analyses have been used to elucidate the origin of HIV, reconstruct its spread around the world and to resolve ongoing transmission patterns within countries and smaller areas.

1.4.2 Subtype diversification and global spread

Subtypes are defined based on the high genetic distances found between them: 25-35%, as compared to 5-20% found within a subtype [77]. Diversification into subtypes is the result of a number of random founder effects of HIV-1 group M outwards from Kinshasa [19, 22, 78]. The present HIV subtype distribution is geographically heterogeneous. Globally, subtype C is most prevalent and accounts for nearly 50% of infections [11]. In Southern Africa and in India, >90% of infections are subtype C [11]. Phylogenetic and phylogeographic analyses place the origin of subtype C around 1960 in the Democratic Republic of Congo [79] and then indicate spread eastwards and southwards in Africa. Subtype C was introduced into India around 1971 [80] and, as in sub-Saharan Africa, is transmitted mainly heterosexually.

B is the most dispersed subtype and was the first discovered. B predominates in Europe and North America but is responsible for only 12% of infections globally [11]. Bayesian phylogenetic analysis testing strongly supports that subtype B was exported from Africa to Haiti around 1966 and then to USA around 1969 [81]. Before the end of the 1960s, subtype B was also circulating in South America [82], but was not introduced into Eastern Europe until the 1980s [83]. Comparative demographic analyses of subtypes B and C suggest the B epidemic initially exploded in the West among PWID and MSM (growing twice as fast as it had in Africa) but then stabilised in the 1990s, whereas the subtype C epidemic in sub-Saharan Africa still appears to be growing exponentially [84]. In Eastern Europe, where a high proportion of HIV

transmission is through PWID [61], the subtype B epidemic also seems to be growing exponentially [83].

Subtype A accounts for another 12% of infections [11] and is most common in East Africa (where it co-circulates with subtype D) and in Eastern Europe where over 70% of infections are subtype A [83]. The subtype A epidemic in Eastern Europe began in Ukraine in the mid-1990s among PWID and rapidly spread to other Eastern European countries and Russia [85]. The absence of a selection during blood to blood infection (due to the absence of a mucosal immune barrier), combined with the rapidity of the spread through the region, led to subtype A displaying particularly low genetic diversity. Subtype AE, the first recombinant discovered in Thailand in the 1980s [86], dominates the southeast Asian epidemic [11].

The dates of viral introductions based on sequence analysis into different parts of the worlds are consistent with our overall understanding of HIV spread across the world. HIV epidemic growth patterns arise as a consequence of the underlying characteristics of the transmission networks. The most notable difference is that in sub-Saharan Africa, a large proportion of HIV transmission occurs through heterosexual sex and from mother to child, while in the rest of the world MSM and PWID are most affected. Within smaller populations and geographical areas other epidemiological factors are likely to affect the rate of spread, including times of introduction into different risk populations, sexual and drug behaviours, as well as laws, health policies and education relating to HIV. Only rarely have subtype-specific patterns that could play a role in the dynamics of these distinct epidemics (such as the increased transmissibility [87] and disease progression [88] of subtype D) been observed.

HIV was introduced into the UK MSM population around 1980 [89], probably from the USA [90], and separately into UK PWID around 1983 [91, 92] Until the early 1990s the UK epidemic was dominated by the transmission of subtype B among MSM and PWID. In the 1990s, cases among heterosexuals became more common [93] and these were often linked to immigration [90]. The prevalence of non-B subtypes has increased over time (Figure 1.3). 48% of HIV diagnoses between 2002 and 2010 were among heterosexuals [93]. The predominant viral subtype was B (39.9%), followed by

C (34.3%), A (5%), recombinants (12%), D (2.4%), G (2.7%) and AE (2%). The genetic diversity of the UK epidemic is a reflection of immigration in to the UK: subtype C comes from sub-Saharan Africa and India and subtype A comes from East Africa and Eastern Europe. However, non-B subtypes are increasingly being acquired within the UK [94, 95] (Chapter 5). For some time the UK epidemic was compartmentalised by subtype and risk group, with B predominantly circulating among MSM and non-B virus among heterosexuals. Yet recently a small but significant proportion of non-B infections are being acquired through sex between men [93] (Chapter 5).

1.4.3 Findings from local epidemics

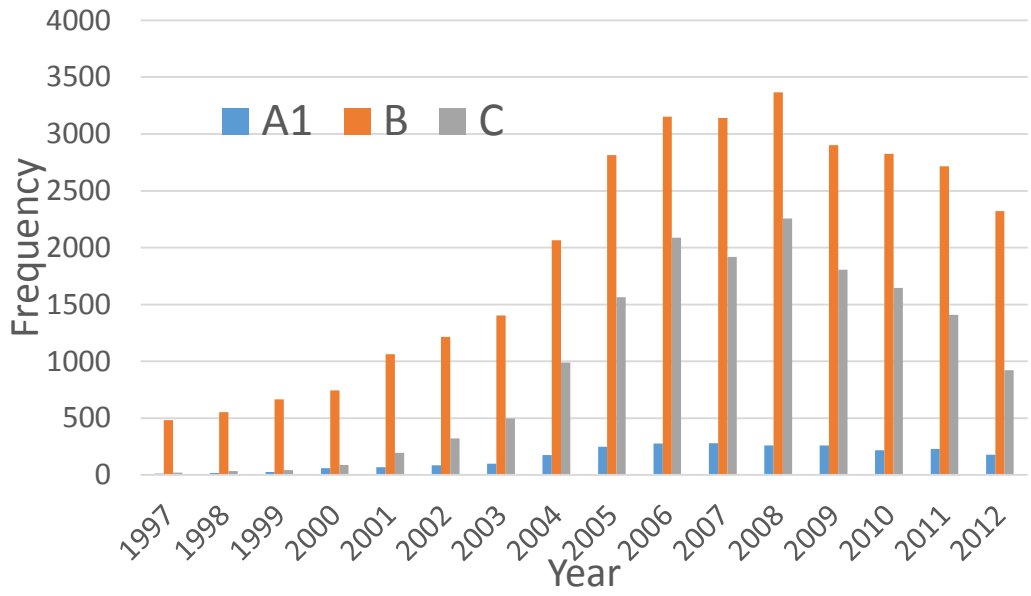


Figure 1.3: HIV diagnoses by subtype and by year in the UK HIV Drug Resistance Database. Generated from the database.

Local and national epidemics can be monitored through phylogenetic analysis to help guide intervention strategies and make predictions about the future of the epidemic [96]. In addition, molecular epidemiology has been applauded for providing an alternative source of information on epidemics in parallel to that revealed through collected epidemiological data, , and can do so when epidemiological data are lacking.

For example, phylogenetic analyses can reveal whether imported strains are spreading locally [97] (in which case local and foreign sequences are intermingled in the tree). Linkages between national epidemics have been established within Europe [98] and globally [99]. Within Switzerland, phylogenetic analysis established that only 25% of non-B infections were found to arise within Swiss-specific subtrees, indicating a growing role for immigration in the Swiss HIV epidemic [97]. In France, a phylogenetic analysis of sequences sampled from recently acquired infections concluded that at least 20% of non-B subtype infections had been acquired in France [100]. In Chapter 5, I investigate the spread of non-B subtypes within the UK using phylogenetic analysis.

Another question that has been addressed using phylogenetics is the contribution of recent infections towards onward transmission. Because recently infected patients have high viral loads and often do not know their status, it has been suggested that they are responsible for a disproportionate number of onward transmissions. Accordingly, their viruses have shown a stronger tendency to cluster in phylogenies than those from chronic infections [101, 102], meaning that they are found close together in trees, separated by small genetic distances. (Cluster definitions and methods are explained in detail in section 3.2.) However, clustering of recent infections does not necessarily mean that onward transmission occurred during early infection [103]. In fact a large proportion of clustered recent infections will be recipients sampled soon after they themselves became infected [104]. Clustering of recent infections will also be more likely because sequences will not yet have diverged [104].

The extent of overlap between risk groups and the role of bridging populations in driving the epidemic remains an open question. In Switzerland, phylogenetic analysis has demonstrated a lack of overlap between MSM and heterosexual/PWID HIV epidemics. The heterosexual subtype B epidemic appears to have been continually reseeded by PWID and not to be self-sustaining [105]. In the UK, between 1 and 11% of self-identified heterosexual men appear to have become infected with subtype B through sex with men [106], providing a potential bridge between the MSM and

heterosexual subtype B epidemics. In Chapter 5, we examine the extent of non-B subtype bridging in the other direction, form heterosexuals to MSM.

The higher the sample fraction, the more detailed analyses can be. In the UK, tens of thousands of sequences are available through the UK HIV RDB: sample coverage approaches 60%. Phylogenetic analysis has revealed six large transmission chains among MSM, reflecting independent introductions into the UK. The epidemic displayed exponential growth in the 1980s, then stabilised in the 1990s but continues to be highly clustered [107]. Within large MSM clusters, the internal architecture has been further resolved indicating that the epidemic has unfolded in bursts with 25% of transmissions occurring within 6 months of infection [108]. Heterosexuals in the UK displayed far less clustering and much slower epidemic dynamics: only 2% of transmissions occurred within 6 months of infection [109].

The Swiss HIV Cohort Study similarly contains HIV sequences from >60% of the HIV-infected population. The UK and Switzerland have led the way in the field of molecular epidemiology and phylodynamics. The two countries' epidemics are compared in Chapter 4.

1.4.4 From phylogeny to transmission network

Recently, there has been huge interest in using HIV phylogenies to gain information about the HIV transmission network and sexual contact network. Networks have been developed extensively in mathematical models to study the spread of infectious diseases and benefit from a rigorous statistical framework (see section 1.5). Meanwhile, sequences provide what mathematical models are often lacking: data.

1.5 Network modeling of epidemics

1.5.1 Classical modeling approaches

Mathematical models can be used to understand and predict the spread of diseases using sets of equations relating the variables in the system. Models are intended to be a simplification of the real world, but capture the essential features of the system, for example the number of infected individuals over time. Additionally, models offer a

setting in which to test the effect of interventions, for example they can be used to estimate the reduction in the total number of infected individuals expected by using a vaccine with a particular efficacy.

Classical epidemiological models divide the population into classes (or compartments), such as "susceptible" and "infected" and represent the flow of individuals between compartments using equations. One of the simplest models, the SIR (susceptible/ infected/ recovered) model, can be represented as:

Where S represents the proportion of susceptible individuals at time t, I the proportion of infected individuals and R the proportion of recovered (or removed) individuals.

Individuals move from compartment S to compartment I according to the transmission rate of the disease and the contact rate between infecteds and susceptibles (amalgamated into β) and the relative numbers of susceptibles and infecteds. Individuals then recover at a rate γ. Thus:

$$\frac{dS}{dt} = -\beta SI \quad (1)$$

$$\frac{dI}{dt} = \beta SI - \gamma I \quad (2)$$

$$\frac{dR}{dt} = \gamma I \quad (3)$$

An example of the movement of a population through the compartments for set rates of β and γ is shown in Figure 1.4.

From looking at equation (2), it is clear that if people are infected slowly and recover rapidly ($\beta SI < \gamma I$), there will not be anyone in the infected compartment and the disease will never take off. The relative rate of exit from the infected compartment as

compared to that of entry into the compartment (the relative recovery rate or γ/β) must be small enough for the disease to spread. The inverse of the relative recovery rate is the reproductive ratio R_0.

$$\boldsymbol{R_0 = \frac{\beta}{\gamma}} \tag{4}$$

R_0 is defined as the number of secondary cases arising from an infection in a completely susceptible population. An infection will only take off in a population if R_0>1.

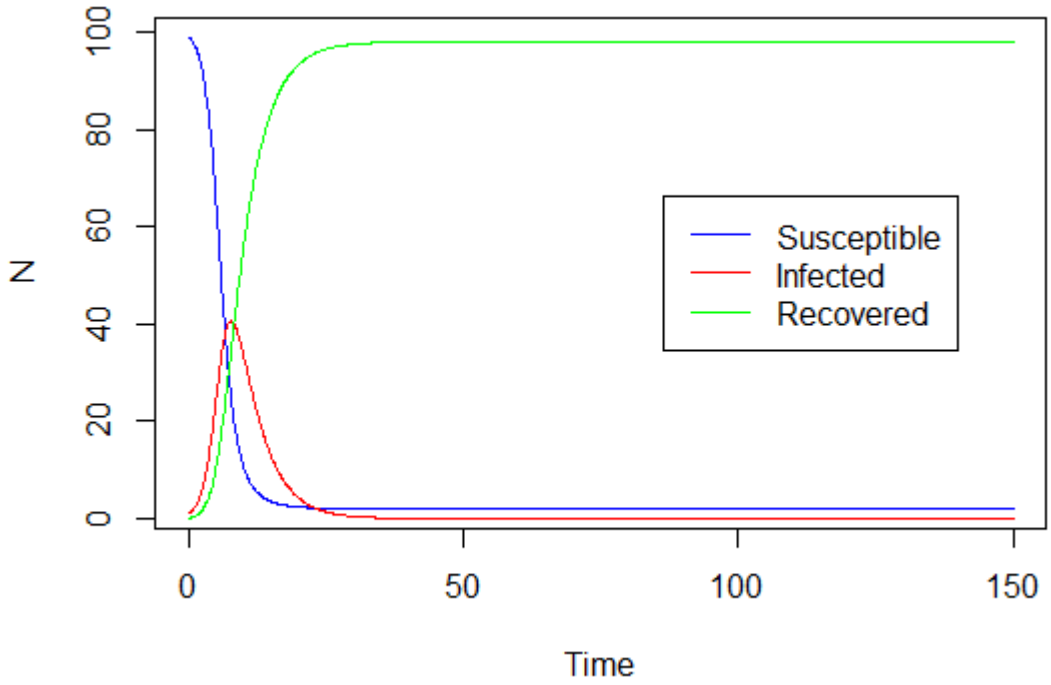


Figure 1.4: Infection dynamics of an SIR model. This figure was plotted assuming β=1 and γ=0.25 for a population of 100 with a single individual initially infected.

Compartmental models often assume random mixing, which means that all susceptible individuals are equally likely to come in contact with infected individuals. As random mixing is a poor assumption in the case of sexually transmitted infections such as HIV [110], some models incorporate high and low risk groups which interact more or less. Compartmental models in fact be extended to include numerous compartments, for

example disease stages (acute, chromic, AIDS) or age group, so this allows for some differences in transmissibility and behaviour between groups. However, the higher the number of compartments the more parameters are required. The error in each introduces uncertainty into the model.

The first mathematical models of HIV were used to predict the number of future HIV and AIDS cases. Blower *et al.* conducted an analysis of the HIV epidemic in New York City [111]. In order to make their model realistic, they included three risk groups (MSM, PWID and heterosexuals) and two sexes, with interaction terms within and between each risk group. However, the model contained over 30 parameters, and due to the uncertainty in the parameters the predicted number of infections over time had very wide confidence intervals. Williams and Anderson also tried to predict HIV cases for England and Wales [112]. As well as risk group, they included categories for levels of sexual and drug-related activity. Again, so many parameter combinations fit past trends that they could not make reliable predictions, although they emphasised the importance of bridging populations to future trends.

Compartmental models are used extensively, and usefully, to test and predict the impact of different intervention strategies. However, they suffer from three major limitations. Firstly, as mentioned above, they assume random mixing despite the pool of partners realistically being much smaller than the population as a whole. Secondly, compartmental models cannot take into account differences at the individual level, despite heterogeneity in number of partners being central to the spread of STIs such as HIV (see section 1.5.5). While additional compartments can be added, the large number of parameters required introduce uncertainty. Thirdly, the simplest compartmental models are deterministic so that the output of the model is fully determined based on the initial conditions and the parameter values. A deterministic model does not take into account randomness and will produce the same output if it is run with the same starting parameters.

These limitations can be addressed by models which make explicit the contact structure in a population by using a network and model the spread of disease on that network.

1.5.2 Contact network models

Contact network models capture the diversity in host behaviours and interactions by modeling each individual in the population as a discrete object with its own parameters. As well as variation in the number of partners, other features of the network, for example assortativity (the tendency of individuals sharing characteristics to partner) or heterogeneity in transmission rates can easily be incorporated. Network models usually still have an underlying SIR model, but progress through disease stages for each individual can be determined by that individual's characteristics. Whilst transmission and contact rates in compartmental models were amalgamated into one parameter β, they can now be decoupled. In some networks (heterosexual networks where partnerships are assumed to form randomly given the heterogeneity in contact number), it remains possible to directly calculate R_0 [113]:

$$R_0 = \frac{\bar{\tau}}{\tau_c} \quad (5)$$

Where $\bar{\tau}$ is the average likelihood of transmission and τ_c is the inverse of the contact rate or level of connectivity of the network. The higher the contact rate, the lower the transmission probability required in order for an infection to take off. Hence we refer to τ_c as the epidemic threshold (see section 1.5.5).

Incorporating stochasticity is more straightforward in network models as these contain inherent randomness. As such, models with the same starting parameters can have different outcomes. Finally, network models do not require large population sizes. Recently developed dynamic network models even allow for the structure of the contact network to change over time.

One remaining problem however, is how to specify the underlying contact network of a population.

1.5.3 Network theory

In a network, each individual is represented by a node, and if they are linked, they share an "edge" that connects them. For some diseases, the definition of a contact is ambiguous; however for STIs, an edge indicates partnership, with sexual contacts

occurring as events along that edge. Edges can be directed (for example, showing the transmission of an infection from one individual to the next) or undirected (showing that the two individuals interact with each other). In a directed network, edges might be represented by arrows while, in an undirected network, nodes would be linked by a symmetrical line. Importantly, in the context of epidemiology, an infection can only be transmitted between an infected and a susceptible node that are linked on the network.

Network theory has been developed independently in two fields: social sciences and physics, yielding a wealth of metrics which can be used to describe network structure and be useful for modeling disease transmission. Some metrics are measured at the network level. The number of edges connected to each node is referred to as its degree. If the degree of every node in a network is estimated, the degree distribution of the network as a whole can be plotted, showing the frequency of nodes with each number of connections. For example in Figure 1.5, node 2 has degree 3 and all other nodes have degree 1. The degree distribution is an important characteristic of a network because it provides information of the process that has led to the formation of the network. The distance between nodes is the number of edges between them on the shortest path, and the diameter of the network is the longest of all its pairwise distances. The diameter of the network in Figure 1.5 is 3. The clustering coefficient measures the tendency of nodes in a graph to cluster together, and can be estimated for a single node (local clustering) or for the network as a whole (global clustering).

On one hand, data has been collected to construct real networks. However, determining the mixing behaviours between every individual in a population is time consuming and expensive and impossible for large populations. In the case of an STI like HIV, information on sexual contacts may not be readily volunteered (particularly in view of the criminalisation of HIV transmission), and there may be issues of recall. In parallel, algorithms have been developed to simulate networks. The simplest network that can be generated is random. In a random network such as those generated by the Erdõs-Rényi model (Figure 1.6A), the probability of any node being connected to another is constant and independent [114]. In a network of n nodes, with a probability p for each

edge, the expected degree of each node will be $(n-1)p$ and the degree distribution of the network will be binomial. The diameter in a random network is proportional to the number of nodes, and clustering is low.

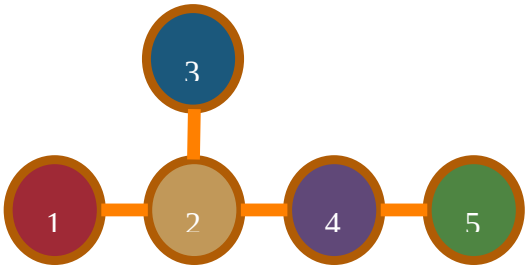


Figure 1.5: Example of an undirected sexual contact network.

1.5.4 Small world networks

Empirical observations suggested early on that many real networks, such as social networks, were not random. In a random network your friends would be no more likely to know each other than two individuals taken from the population at random. The small-world phenomenon was first illustrated by Stanley Milgram's experiment, in which he gave people packages which had to end up with a target final recipient [115]. The average number of steps between the initial recipient and the final recipient was much smaller than expected: around 6, hence the phrase "six degrees of separation". In social networks, the distance between two nodes is shorter than expected in a random network, degree is highly variable and clustering is high [116]. Watts and Strogatz developed the first small world network model by rewiring a lattice at random [116]. Networks generated through the Watts-Strogatz model have shorter path lengths and exhibit higher clustering than random networks (Figure 1.6B.). They demonstrated that their model was a good fit to the neural network of *C.elegans*, the power grid of the western USA, and collaborations between film actors, the latter being a proxy for social networks. Similar patterns have been observed in many other types of self-organising systems such as protein interaction networks and the World Wide Web.

In small world networks, instead of the distance between nodes growing proportionally to the number of nodes in the network, it grows proportionally to the logarithm of the number of nodes. Empirical analyses have demonstrated that the nodes in real networks frequently follow a power law distribution [117]. The probability $P(k)$ of a node having k connections can be expressed as:

$$\boldsymbol{P}(k) \sim k^{-\alpha} \qquad (6)$$

The scaling exponent α is a measure of the variance of the network and is the slope of the line when the distribution is plotted on a log log scale. If the degree distribution of a network follows a power law, the network is called scale-free.

1.5.5 Network Modeling of Epidemics

Disease spreads more easily on small world networks than on random networks. The behavioural heterogeneity underlying the power law distribution in sexual contact networks is thought to be central to the spread of STIs and may indeed be the reason STIs persist. In parallel, however, heterogeneity in sexual activity decreases the magnitude of the HIV epidemic so that some people are very unlikely to ever become infected [118].

Small-world networks offer obvious possibilities in terms of interventions. They are resilient to random failure but vulnerable if hubs are destroyed [119, 120]. Random interventions may have no impact as they are more likely to hit nodes that are not highly connected. In contrast, removal of hubs will disconnect parts of network and disassemble its structure. The power law exponent indicates to what extent targeted interventions are necessary.

As the variance of the degree distribution increases, the inverse of the contact rate from τ_c from Equation (5) decreases [121] and R_0 increases. If the exponent of the power law distribution lies in the range $2 < \alpha < 3$, the variance of the network degree distribution is theoretically infinite [113]. The level of connectivity τ_c from equation (5) tends towards 0 and there is theoretically no epidemic threshold [110, 122]. Even a pathogen with very low transmissibility will persist. In this case, *only* if highly connected nodes are targeted will it be possible to halt the epidemic [120].

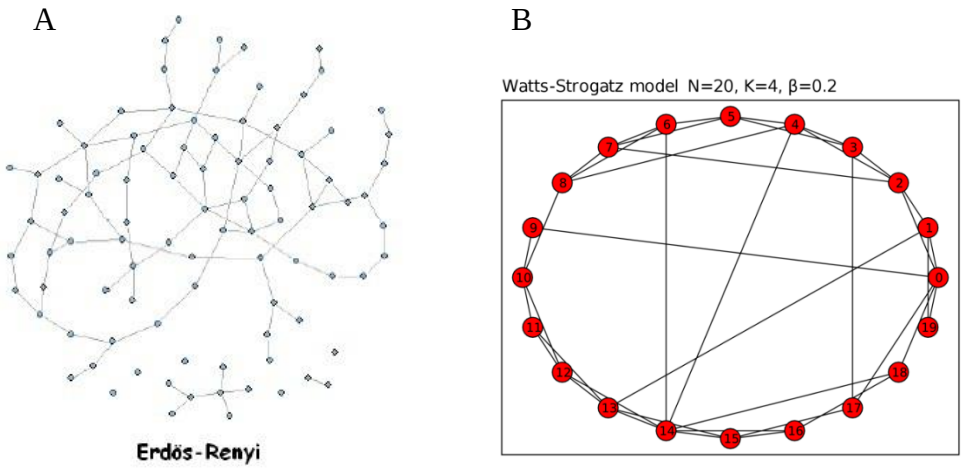


Figure 1.6: Networks generated under Erdõs-Rényi (A) and Watts-Strogatz (B) models (downloaded from Wikipedia under the creative commons licence).

While all studies of sexual contact networks agree that networks are highly heterogeneous in terms of number of contacts [118, 123-129], it is disputed whether variance is infinite. Liljeros *et al.* searched for the best fit distribution to sexual partnerships based on a Swedish survey of sexual behaviour [128]. They fitted a linear curve to the tail of the double-logged plot of the distribution. The exponent of the power law can be extracted from the slope of the line, and they found it to be between 2 and 3 for men and women. They concluded that the observed pattern is symptomatic of a preferential attachment scenario in which the most highly connected nodes tend to accumulate contacts over time ("the rich get richer"). Schneeberger *et al.* fitted power laws using maximum likelihood to the tail of the cumulative degree plots of British, USA and Zimbabwean sexual surveys [129]. In all cases they found the power law to be a good fit and the exponents to be between 2 and 3 in most cases. They note the particularly wide variance in the degree distribution of British MSM.

However, these papers have been criticised for their statistical methods [123, 124, 126]. The methods focus on the tail of the distribution to produce an estimate, but there is little information in the tail (because there are few data points) so fitting can be unreliable. In addition, uncertainty increases with increasing degree (18 is more likely to be rounded than a smaller number) which biases estimates of the exponent. Both

Handcock and Jones [124, 126] and Hamilton *et al.* [123] insist that fitting degree distributions without an underlying model for the formation of partnerships is insufficient.

For example, if all individuals in a population acquire partners at a fixed, homogeneous rate, the degree distribution is expected to follow a Poisson distribution. If the homogeneity assumption is relaxed, but people still acquire partners at a constant rate, this will overdisperse the data and the degree distribution will follow a Poisson lognormal or a Negative Binomial distribution. Importantly, both these distributions always have finite variance, although they do allow for long tails. The Yule [130] and Waring [131] distributions both result from preferential attachment models. They incorporate a probability proportional to k that a new link is made to a person of degree k, as well as a constant probability that a new link is made to a person of degree 0 (previously sexually inactive). In addition, the Waring incorporates a rate for non-preferential attachments, so that a partnership may form at random between two people, regardless of degree [124, 131]. The Yule and the Waring are power-laws that have infinite variance for $2 < \alpha < 3$.

Handcock and Jones fitted the Swedish data and sexual networks from Uganda and the USA to the distributions listed above through maximum likelihood [124-126]. They found that the Negative Binomial offered the best fit to most networks [124]. When they fitted a Yule distribution to the data (which still provides a better fit than the power law fitted by Liljeros *et al.*) the exponents were always >3, indicating the existence of an epidemic threshold.

Beyond fitting models, Hamilton *et al.* note that the degree distribution exponents of five USA sexual contact networks do not predict the epidemic behaviour that is observed [123]. They found only very small differences in fits between the models tests, underlining our inability to differentiate between them. They suggest that a single stochastic process is insufficient to explain the degree distributions or that the degree distribution is inadequate in representing the behaviour of the network.

1.5.6 Calibrating models

These studies underline the need for a better understanding of both the contact network and the disease transmission network. Creating a network representing a population requires knowledge of all the individuals in the population and their interactions. This is time-consuming and expensive, if not impossible, for anything but a small population [132]. At least, in the case of sexual networks, defining what constitutes a contact is straightforward, whereas this is much more difficult in the case of networks for modelling the transmission of airborne diseases [133].

Numbers of sexual partnerships are collected through surveys, for example the National Survey of Sexual Attitudes and Lifestyles in the UK (NATSAL) [134], and these can be used to estimate degree distributions (see section 1.5.5). One limitation of studies of sexual behaviours is that participants may be unwilling to share information. Collecting social network information rather than sexual contact network may yield better responses, for example through rapid questionnaires [135], or venue-based data collection [136]. Although ego-centrically collected, these data can be used to inform network models [137].

Any kind of random data collection, however, risks missing high degree nodes which are so central to capturing the variation in sexual contacts across the population. One solution to increase their likelihood of being sampled is to sample people with STIs [129]. Contact tracing, the act of obtaining a list of sexual contacts from patients diagnosed with an STI, has long been carried out both for testing and treatment. If any of those traced are infected, their contacts can also be identified. Contact tracing yields data on infected and uninfected partners which can be used to create contact transmission networks embedded within contact networks [138]. However, in the case of HIV, because of the delay between infection and diagnosis, people frequently incorrectly identify who infected them [139]. Instead, sequence data collected may provide additional insights into transmission network structure [140] and contact network structure [141] (Chapter 1).

2 MATERIALS AND METHODS

2.1 Data

2.1.1 Availability of HIV sequences

ART targets the protease and reverse transcriptase of HIV, selecting for drug resistant mutations (DRM) within the *pol* region that encodes them. HIV treatment is compromised by the presence of DRM (see section 1.2.4).

This has led to the accumulation of HIV *pol* sequences in databases. Early phylogenetic studies used *gag* and *env* sequences [72, 142, 143] while HIV *pol* was criticised for being too genetically conserved to accurately reconstruct HIV transmission. Hue *et al.* confirmed in 2004 that *pol* contains sufficient genetic diversity for the phylogenetic reconstruction of transmission events [144]. This region of the HIV genome is now extensively used in phylogenetic studies to address questions regarding HIV transmission dynamics. The *pol* region can also be used to classify the subtype of the infection.

2.1.2 The UK HIV Drug Resistance Database (UK HIV RDB)

Following recommendations from the British HIV Association [49], clinics sequence the HIV *pol* region of patients before they begin therapy and in the event of virologic failure (viral loads >1000 copies/ml). Since 2001, *pol* sequences generated for clinical purposes in the UK have been deposited in a centralised database: the UK HIV Drug Resistance Database (http://www.hivrdb.org). The purpose of the UK HIV RDB is to monitor the spread of drug resistance and serve as a research resource.

The UK epidemic is one of the most densely sampled HIV epidemics and the UK HIV RDB is the largest such database in the world. In 2011, the UK HIV DRB was estimated to contain sequences for approximately two-thirds of the subtype B MSM patients who were treated for HIV in the UK [140], and this proportion has increased since then. Sequences are available for around 50% of the infected population and >80% of patients diagnosed since 2005 (David Dunn, personal communication).

Data are provided by clinics to the UK HIV RDB in a pseudo-anonymised form with a phonetic version of the patient's name (soundex code), date of birth, date of diagnosis, sex and genetic sequence. Epidemiological data contributed by Public Health England included year of birth, age, gender, self-identified ethnicity (White, Black-African, Black-Caribbean, Black-other, Indian/ Pakistani/ Bangladeshi, other Asian/ Oriental, other-mixed, unknown), country of birth and self-reported most likely route of infection (PWID, heterosexual sex (HET), MSM, mother to child, blood product, or unknown). The epidemiological and subtype breakdown of the latest release is shown in Table 2.1.

Although the collection and management of data by the UK HIV RDB are rigorous, there is always a risk that data collected are incorrect. Epidemiological data collected are self-reported. It is also possible that a single patient moving between clinics could be recorded as two separate patients. In order to avoid duplication of records, if the soundex code and date of birth of multiple patients match, genetic distances between sequences are calculated. Where the distance falls below 3.5%, database entries are examined further. If soundex codes and dates of birth match, dates of diagnosis are within a month of each other and genetic distance is <3.5%, the sequences are assumed

to be from the same patient and are linked with the same patient identifier in the database.

Table 2.1 Epidemiological characteristics of 63163 unique patients in 2014 the UK HIV RDB.

		A1	B	C	Other	Total
Risk group	**MSM**	170	22296	662	2364	25492
	HET	1721	3421	11930	7912	24984
	IDU	92	879	128	303	1402
	Other/unknown	529	5053	3144	2559	11285
Sex	**Male**	883	25842	4995	5832	37192
	Female	1195	2051	8215	5183	16644
	NA	434	4115	2654	2123	9327
Ethnicity	**White**	495	22868	1729	3539	28631
	Black-African	1354	561	10221	5831	17967
	Black-Caribbean	49	1222	273	282	1826
	Indian/Pakistani/ Bangladeshi	44	350	235	130	759
	Other/unknown	570	6648	3406	3356	13980
Total		2512	31649	15864	13138	63163

Notes: The three major subtypes in the UK, and those that will be analysed in this thesis, are A1, B and C. Risk groups are men who have sex with men (MSM), heterosexuals (HET) and people who inject drugs (PWID).

Data are then fully anonymised and delinked so that results from analysing the database cannot be used to identify source partners, mainly because of the criminalisation of HIV transmission [145]. In order to further decrease the possibility of including multiple sequences from a single patient, identical sequences are discarded prior to analysis.

Deductive disclosure is the inference of an individual's identity based on a set of anonymised characteristics that together make that individual the only one to fit the description. For example there may only be one female of black Caribbean ethnicity in Dunfermline, born in May 1972, diagnosed with HIV in 2010. Data in the UK HIV

RDB are amalgamated to prevent deductive disclosure. The UK is divided into only 14 identified regions and data are aggregated from all centres within each region.

Originally, sequencing was performed before the initiation of ART, and many sequences in the database originated from patients in chronic infection who had been infected for a long time. Sequencing is increasingly performed at diagnosis for better estimation of TDR rates. As we will see in Chapter 1, the timing of sampling and sequencing will influence the likelihood of a sequence clustering (see section 3.2). Samples sequenced closer in time to each other are more likely to cluster, as are and samples sequenced closer to infection.

Three different versions of the UK HIV RDB were used in this thesis, as described in Table 2.2. For patients with more than one sequence in the database the earliest sequence was used, usually obtained before the initiation of ART.

Table 2.2: Versions of the UK HIV RDB used in this thesis

Year of release	Sequences up to	Number of unique patients	Number of sequences	Used in chapters
2009	2007	34469	26356	3
2012	2010	43002	55556	5
2014	2012	63163	106402	4, 6

2.1.3 The LANL database

HIV is the most widely sequenced organism. In addition to Genbank, HIV sequences made publicly available are curated by the Los Alamos National Laboratory in the LANL database. The LANL database (accessed 02/06/2015) contains 198428 *pol* sequences (minimum length 500bp) from 154 countries. The use of background sequences improves the reliability of inferences made through phylogenetic analysis. Because of the huge number of sequences in the UK HIV RDB and in LANL, we did not use all LANL sequences available. Instead, we downloaded all *pol* sequences (HXB2 nucleotide positions 2253-3459) available for subtypes A1, B, C, D and G (analysed in this thesis), of minimum length 500 bp. and excluding UK sequences, to create custom background sequence libraries. We then use the blast algorithm (Basic

Local Alignment Search Tool), as implemented in Viroblast [146] and Geneious [147], to return the ten closest LANL matches to each UK HIV RDB sequence. In many cases, the ten sequences returned for multiple UK sequences will be identical, thus the overall number of LANL matches included in the trees is far smaller than the overall number of LANL sequences available. While it would be possible to select the ten next sequences in cases where the first ten matches have already been included, this is not done because sequences more distantly related are unlikely to fall within clusters, and it is preferable not to increase the size of datasets for tree reconstruction any further.

2.1.4 The Swiss HIV Cohort Study (SHCS)

Switzerland is the only other country with an equivalent database and such high coverage of their epidemic (http://www.shcs.ch/). The SHCS was established in 1988 as a prospective observational study. Socio-demographic and behavioural data (year of birth, sex, last negative HIV test, most likely transmission route, confections) of new patients are collected. Patient data are updated with CD4+ counts, treatment and viral failure events. Since 2002, genotypic drug resistance test have been linked to the database, with many stored samples sequences retrospectively. Currently the database contains over >12000 sequences [148].

In Chapter 4, I performed a comparison of the Swiss and UK epidemics. I did not directly access the Swiss data, but rather developed a pipeline which we both submitted our data to.

2.2 Methods

2.2.1 Sequence manipulations

2.2.1.1 Sequencing and alignment

Sequencing for the UK HIV RDB is performed by 19 laboratories across the UK (see Appendix 1) using different sequencing algorithms. The majority of sequences (>90%) cover the entire protease gene and up to 900 bases of reverse transcriptase. Sequences generated using the Trugene® assay [149] are missing the first 9 bases of PR, the last 3 bases of PR and the first 120 bases of RT.

Sequences submitted to the UK HIV RDB are aligned using the Stanford HIVdb program [150], with manual checks for poor quality, and are released as alignments.

2.2.1.2 Subtype assignment

Aligned sequences were submitted to Subtype Classification Using Evolutionary Algorithms (SCUEAL) locally for subtyping [151]. SCUEAL uses a model-based phylogenetic method to assign viral subtype and leaves smaller numbers of sequences unclassified than other subtyping programs [151]. The UK HIV RDB subtypes sequences in parallel using REGA [152]. For pure subtypes (including all those analysed in the present thesis), subtype assignments across methods vary very little. SCUEAL-assigned subtypes were used in this thesis.

2.2.1.3 Sequence formats and editing

For each UK HIV RDB release analysed, a fasta file of aligned sequences and a csv file (containing epidemiological data) were provided to me by the UK HIV RDB. Alignments were checked manually in BioEdit [153] and Geneious. The fasta file and csv were processed in R [154], for example to edit sequence names, or select sets of sequences based on subtype or date of diagnosis. R was used to discard sequences which were missing either the entire PR or the entire RT.

Because drug selective pressure can cause convergent evolution in unrelated viruses, drug resistant sites are removed before phylogenetic analysis. The list of DRM has been standardised by the WHO [155] to facilitate comparisons between regional, national and international statistics and adjust for non-drug related polymorphic DRM [156, 157]. The 2011 updated drug resistance list [158] was used throughout this thesis, and sites were removed in R (Appendix 3).

2.2.2 Phylogenetic analysis

Phylogenetic trees are reconstructed based on the differences found between sequences. However, the number of possible trees increases super-exponentially with the number of sequences, rendering an exhaustive search impossible for large number of sequences. The most statistically robust methods, and those employed in this thesis, are maximum likelihood (ML) and Bayesian inference.

2.2.2.1 Maximum likelihood

The ML method is based on a model of evolution. The ML method is character-based, using all the information contained in the sequences, and is statistically robust and well-founded [159]. The approach calculates the probability of the observed data (the sequence alignment), given the proposed model (tree topology, branch lengths and the parameters of the evolutionary model). For each site in the alignment, the probability of the observed nucleotides is calculated as the sum of probabilities of every possible reconstruction of ancestral states, given the substitution model. As sites in the sequence alignment are considered to evolve independently, the probability of the observed sequences is the product of the probabilities for each site. Under ML, the best tree is the one with maximizes the likelihood, or the probability of observing the data. Overall, this method produces accurate results [160] but can be computationally intensive. Uncertainty is not incorporated into the final tree, but support values can be assigned to nodes in ML trees through bootstrapping (see section 2.2.5.1) and the fit of alternative trees can be compared based on their likelihoods [161].

Two phylogenetic tree reconstruction programs are used in this thesis: RaxML [162] and FastTree [163]. RaxML conducts a more thorough heuristic search for the best tree, but FastTree is faster. FastTree is used in Chapter 3 as part of the pipeline that was applied to the UK and Swiss datasets. They perform similarly in terms of HIV cluster reconstruction (see Appendix 2).

Both programs produce phylogenetic trees in Newick format. Newick tree format is a way of representing phylogenetic structure, branch lengths and bootstrap support values using parentheses and commas. Newick format is the minimal definition for a tree and branch lengths and bootstrap support values are optional. Other formats, Nexus being the most common, can store blocks of additional information associated with branches, nodes and tips, for example clock rates (see below) or colours for visualisation.

2.2.2.2 The molecular clock

2.2.2.2.1 The strict molecular clock

The molecular clock is a technique used in molecular evolution to date events on phylogenetic trees. Emil Zuckerkandl observed in 1962 that the number of differences in haemoglobin amino acid sequences sampled across species was proportional to the duration of time since the species diverged [164]. The suggestion that mutations accumulate in sequences linearly with time constitutes the molecular clock hypothesis:

$$\boldsymbol{r}ate = \frac{genetic\ distance}{2\ \times time} \qquad (2)$$

In the case of species divergence, the clock is calibrated using fossil records to infer rate of evolution. In the case of viruses, which evolve much faster, sequentially sampled isolates have been used to calculate evolutionary rate [74]. Application of the molecular clock uses time-stamped sequences to date events in the evolutionary history of the virus which can be very useful to test different hypotheses, for example those concerning the origin of HIV [21]. While the strict molecular clock is consistently rejected for HIV sequences [83, 84, 165-169], dating is possible nonetheless using relaxed clock models [170].

2.2.2.2.2 The relaxed molecular clock

According to the strict molecular clock, the rate of molecular evolution is constant and branch lengths are directly proportional to time. The relaxed clock models rates of evolution on one hand and the speciation process on the other so that the molecular clock can vary across the phylogenetic tree [171]. Earlier relaxed clock models were based on autocorrelated models of rate change, so that the rate of evolution of each branch was assumed to be inherited from the rate of the parental branch. Uncorrelated clocks depend only upon the mean clock rate associated with the whole tree and vary according to an underlying distribution, for example lognormal or exponential [172].

2.2.2.2.3 Least squares dating

Large phylogenetic trees (containing thousands of sequences) can be time-resolved using least-squares dating (LSD). LSD simultaneously estimates the substitution rate

and dates of ancestral nodes in phylogenies with dated tips using least squares data fitting approaches [173]. LSD is significantly faster than BEAST (see below), running in linear time for rooted trees, and (nearly) quadratic time when the root must be estimated. LSD requires a topology and dates of tips as input. The topology of the tree is not changed (other than re-rooting if requested by the user).

2.2.2.3 Bayesian evolutionary analysis by sampling trees (BEAST)

The Bayesian approach searches for the distribution of trees that is most probable given the sequence alignment under a particular model of evolution. The Bayesian approach makes use of likelihood and incorporates prior knowledge which enables the user to convey any expectations or uncertainty about some of the model parameters before looking at the data. Most importantly, Bayesian analysis generates multiple trees, consistent with our uncertainty about the final tree topology. The most popular program for Bayesian phylodynamic analysis is BEAST [76].

BEAST incorporates relaxed molecular clock models which model rates of evolution and time in parallel based on prior distributions [171]. Use of a clock allows for events on the phylogenetic tree to be dated. The posterior distribution of parameters is sampled through Bayesian Markov chain Monte Carlo (MCMC) chains.

2.2.2.4 Cluster identification

The rationale behind different cluster identification methods is explained in section 3.2. In Chapter 3, I discuss the Cluster Picker, developed by Samantha Lycett, which takes as input a phylogeny and a set of sequences and picks clusters based on user-selected genetic distance and bootstrap cut-offs. We compare the Cluster Picker to PhyloPart, which identifies clusters in large trees if the median of their genetic distances is below a user-input *t*-percentile threshold of the whole-tree distance distribution [174].

The Cluster Picker is used throughout the thesis using genetic distance cut-offs of 1.5% and 4.5% an bootstrap cut-offs from 70% to 99%. The Cluster Picker code was modified for Chapter 4 to recognise ambiguous nucleotide bases as matches as

determined by the International Union of Pure and Applied Chemistry (IUPAC) notation.

2.2.3 Epidemic simulation

2.2.3.1 The Discrete Spatial PhyloSimulator (DSPS)

The DSPS is a JAVA-encoded individual based model that runs in continuous time developed by Samantha Lycett [175]. The purpose of the DSPS is to simulate realistic epidemics along complex and variable contact networks. The program takes as input the parameters of the model (including network structure) and outputs a list of timed transmission events (transmitter/ recipient).

Within the DSPS, separate demes contain hosts that share the same parameters. Demes are connected to each other to form contact networks. The DSPS allows three different disease progression models: SI, SIR and SEIR (where are exposed but not infectious for a duration of time).

Throughout the simulation, a series of events (infection, recovery, migration) is created by assessing the current state of the population (for example the number of susceptible and infected individuals) and generating a weighted distribution of the likelihood of each type of event within each deme. The next event is chosen based on its probability and a random number to introduce stochasticity into the simulation.

The DSPS outputs a population log containing the state (S, E, I or R) of each individual in the population at each time point, an event log showing infections and death events, a transmission tree and a pruned transmission tree containing only sampled individuals.

2.2.3.2 HIV-specific DSPS (DSPS-HIV)

To make the DSPS appropriate for the simulation of HIV epidemics, a number of modifications have been made to the base code by Emma Hodcroft [176].

Because of the long term duration of HIV epidemics, demographic parameters (birth and death) have been added. Options exist for the population size to increase through time (growth) or to stay stable. The network structure of the full network (including

all nodes that will die and be born) is set out before the start of the simulation in this case, rather than the network structure being updated, but links are inactivated unless hosts are alive. In the simulations used in this thesis, demes were set to represent households, each containing one male and one female.

The DSPS-HIV has been developed to attach gender and orientation to each individual which determine the possible connections within the network. In this thesis, the networks were fully heterosexual with heterosexual couples existing together within households and a single deme containing female sex workers.

Transmission probability is high during acute stage and drops thereafter.

2.2.3.3 Sampling and sequence simulation

Realistic sampling schemes were executed in R by Emma Hodcroft. For any specified time period, hosts that are both alive and infected were sampled randomly with the desired sample proportion.

Matthew Hall's "Virus Tree Simulator" was then used by Emma Hodcroft to simulate viral phylogenies based on the infection and sample times of individuals, taking into account within host viral evolution. Sequences were simulated along the resulting phylogenies using piBUSS [177] according to an HKY substitution model with different parameters allowed for the $1^{st}/2^{nd}$ and 3^{rd} codon positions. Ancestral sequences for use in piBUSS were generated in BEAST from a collection of full-length subtype C samples by Gonzalo Yebra from southern Africa (South Africa, Botswana, Malawi and Zambia) obtained from LANL.

2.2.4 Network analysis

2.2.4.1 Network formats

All analysis of networks was done in R using two packages: network [178] and igraph [179]. The two packages contain many of the same and some different functions. Both accept as input networks in edge list format, whereby each line in the file represents an edge and displays the name of the nodes linked by that edge. Each line can in addition contain a numerical variable describing that edge (edge weight) which can be

used to represent the strength of the tie for example. Alongside this, the user can input a data frame containing information on each of the nodes (for example age, sex and risk group).

2.2.4.2 Network reconstruction

2.2.4.2.1 Clusters as networks

Networks in Chapter 4 were reconstructed from phylogenetically identified clusters in order to generate degree distributions instead of cluster size distributions. In this case, all individuals linked in a cluster would be linked to all other individuals in that cluster in the final network. An R script was written to convert each list of clustered sequences (output from the Cluster Picker) into an edge list (Appendix 3).

2.2.4.2.2 Time-based reconstruction

The method developed by the Leigh Brown group to convert phylogenies into networks does so based on the time to most recent common ancestor (tMRCA) for each pair of sequences [140]. Nodes are linked together in the network if their tMRCA is smaller than or equal to a user-selected cut-off. The networks returned from this method can be considered multifurcating trees with no time in which nodes are labelled by host (as well as epidemiological information and date of diagnosis). The eventual aim of the method is to produce transmission networks that resemble transmission chains, in which nodes represent hosts and links represent timed transmission events. The advantages of this method are discussed in section 3.2. R code was obtained from Samantha Lycett in order to perform this task, and was modified to return the tMRCA and pairwise genetic distance for each pair of nodes in addition to whether they were linked or not. This information was used to inform thinning algorithms implemented in Chapter 6 to eliminate unlikely links within reconstructed networks.

2.2.5 Statistical analyses

2.2.5.1 Bootstrapping

Bootstrapping is a statistical resampling technique used to generate a distribution from a single sample of data (because no additional data is available) [180]. The original

data are resampled with replacement, each time re-estimating the variable of interest to test its accuracy.

Within a sample (for example the sample of diagnosed HIV patients) it is impossible to know the sample statistics for the full population. By resampling sampled data, it is possible to make inferences about the true population statistics. This technique is used throughout the thesis. Bootstrapping allows us to determine whether our observed data diverges significantly from an effect that could be observed by chance.

In the case of phylogenetic analysis, bootstrapping is commonly used to assess the support of bifurcations within phylogenetic tree [181]. The tree is initially constructed with the true data. Sites in the sequence alignment and then resampled with replacement to produce bootstrap replicates of the alignment in which some sites will be present more than once and others will not be present at all. The phylogeny is constructed for each bootstrap replicate. Support to nodes in the original tree are given based on the proportion of bootstrap replicates which identified the same bifurcation as in the original tree [181].

2.2.5.2 Fisher's test/ chi-squared test

Fisher's exact test and the Chi squares test are both statistical significance tests used in the analysis of contingency tables. In both cases, these tests presume a categorical distribution of a variable and compare the observed distribution to the expected. Within large samples, a chi-squared test can be used instead of Fisher's test because the sampling distribution approximates the theoretical chi-squared distribution.

2.2.5.3 Comparing distributions

The Kolmogorov-Smirnov (KS) test is a very general non-parametric method for comparing distributions [182]. A distance is calculated between the cumulative distributions of the two compared samples under the null hypothesis that the two samples are drawn from the same distribution (the KS statistic). It is sensitive to differences in both location and shape between the two distributions, and has the huge advantage of being broadly applicable to any sets of data without knowing much about their distribution beforehand. However, in the case of comparing network degree

distributions, as we do in Chapter 6, the KS test is too sensitive to differences in size between networks. In addition, it does not provide any measure of the distance between two networks.

The recently developed Degree Distribution Quantification and Comparison (DDQC) [183] algorithm to compares network degree distributions and corrects for differences in population size. The range of node degrees of each distribution is divided into eight regions based on its minimum, maximum, mean and standard deviation. The probability of the degree of any node being contained within each interval is calculated. A vector of eight feature values is then extracted from each degree distribution and the distance between two networks is the sum of the absolute differences for each of the eight features extracted. JAVA code was obtained from the authors of this publication.

2.2.5.4 Fitting distributions

A more general method to compare distributions is to fit them to a theoretical distribution and compare the parameters of that model distribution. In the case of real-world network degree distributions, power laws are frequently a good fit and the exponent of the power law can be compared. As explained in the introduction (section 1.5.5), the exponent of those power laws is meaningful in terms of intervention strategies.

In addition, the tendency of a network to fit to some degree distributions is indicative of the process that led to the formation of that network. In the case of sexual contact networks, the process of partnership formation will have an impact on the degree distribution [123] (see section 1.5.5).

The R package degreenet [126] provides function to fit and simulate from the Waring, Pareto and Negative Binomial distributions.

2.2.5.5 Receiver Operating Characteristic (ROC) analysis

Receiver Operating Characteristic (ROC) analysis is a method for testing a binary classifier. The discrimination threshold of the classifier (a continuous variable) is varied, and each time, the agreement between the classifier and the true state is

measured. The number of cases correctly classified as positive (true positives or TP) and negative (true negative or TN) and those incorrectly classified as positive (false positive or FP) and negative (false negative or FN) are counted.

Based on these numbers, the sensitivity, specificity and precision of the classifier at that threshold can be calculated:

$$\boldsymbol{S}ensitivity = TP/(TP + FN) \quad (7)$$

$$\boldsymbol{S}pecificity = TN/(TN + FP) \quad (8)$$

$$\boldsymbol{P}recision = TP/(TP + FP) \quad (9)$$

The sensitivity is a measure of how well the classifier is at capturing positives, specificity measures how well negatives are excluded and precision tell us what proportion of the positives elected by the classifier were in fact positive. Depending on the aim of the classifier, sensitivity, specificity or precision can be optimised. For example, in the case of an HIV test, a high sensitivity test which captures all positives at the expense of specificity might be preferred in a first instance. Some positives will in fact be negative but this would become apparent in a second round of testing with a high specificity test.

		Classifier	
		Positive	Negative
True state	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 2.1: Agreement between classifier and true state during ROC analysis

ROC curves plot the sensitivity of the classifier against the inverse of its specificity for all thresholds, in order to estimate the overall performance of the classifier as the area under the curve (AUC). A reliable classifier has high sensitivity and high specificity and so the ROC curve increases very rapidly to yield an AUC close to 1.

In Chapter 7, I use ROC analysis to select a threshold for reconstructed tMRCA of two sequences that best classifies whether those sequences are connected through a direct transmission event.

3 AUTOMATED ANALYSIS OF PHYLOGENETIC CLUSTERS

3.1 Abstract

As sequence data sets used for the investigation of pathogen transmission patterns increase in size, automated tools and standardized methods for cluster analysis have become necessary. The Cluster Picker identifies monophyletic clades meeting user-input criteria for bootstrap support and maximum genetic distance within large phylogenetic trees. A second tool, the Cluster Matcher, automates the process of linking genetic data to epidemiological or clinical data, and matches clusters between runs of the Cluster Picker. I explore the effect of different bootstrap and genetic distance thresholds on clusters identified in a data set of publicly available HIV sequences, and compare these results to those of a previously published tool for cluster identification. To demonstrate their utility, I then use the Cluster Picker and Cluster Matcher together to investigate how clusters in the data set changed over time. I find that clusters containing sequences from more than one UK location at the first time point (multiple origin) were significantly more likely to grow than those representing only a single location. The Cluster Picker and Cluster Matcher can rapidly process

phylogenetic trees containing tens of thousands of sequences. Together these tools will facilitate comparisons of pathogen transmission dynamics between studies and countries.

3.2 Introduction: clustering methods

If all individuals within a transmission chain were sampled, it should in theory be possible to reconstruct the true sequence of transmission events from their genetic sequences [160]. However genetic diversity within the donor combined with a bottleneck at transmission [184], a delay between infection of the recipient and sampling and molecular evolution within patients mean that even with full coverage of a transmission chain, the evolution history may not be fully compatible with the transmission history [185]. In addition, directionality would not be resolved unless additional information (such as time since infection) were available. Thus phylogenetics does not reconstruct the true transmission chain but nevertheless enables inferences to be made about the epidemic that are useful to public health [186]. The aim of phylogenetic clustering analyses is to identify groupings within a phylogenetic tree that are epidemiologically meaningful.

Sequences within the tree that are more related to each other than to the rest of the tree are more likely to be epidemiologically linked. Closely linked infections reflect short durations of time between transmissions and thus clusters represent the leading edge of the epidemic. They are relevant for identifying transmission correlates or designing and evaluating prevention strategies. High levels of clustering in an incompletely sampled population indicates an explosive epidemic pattern and rapid transmission.

Cluster definition varies widely in the literature, based on genetic distance, time to most recent common ancestor and support for the bifurcation (the separation of the cluster from the rest of the tree indicated by bootstrap or posterior probability; Figure 3.1). The specific cut-offs will also vary from gene region to gene region because of the different rates of substitution across the HIV genome, but as this thesis uses only *pol*, I will focus on *pol* here.

At its simplest, it has been suggested that patients should be considered linked if their sequences cluster as closely as two sequences isolated from the same patient [144]. In practice, this has meant a cluster supported by a bootstrap equal or greater to 99% and a genetic distance lower or equal to 0.015 nucleotide substitutions per site (or 1.5%). This definition is used by the San Diego Primary Infection Cohort [187], where patients are diagnosed early in infection, theoretically before mutations have accumulated within patient. All samples have been collected within the same city. Sequences phylogenetically linked at such tight thresholds are highly likely to represent direct transmission events.

However, a large proportion of the UK HIV RDB is made up of patients who were either diagnosed late in infection or who were sequenced a long time after diagnosis (see section 2.1.2). In both cases, we expect direct transmission partners not to group so closely in the tree, as their sequences will have diverged since the transmission event. In addition, because sampling is incomplete, we may prefer to capture patients who whilst not directly linked themselves are nonetheless part of the same transmission chain (separated by 1 or 2 transmissions). If many sequences were contained within such a cluster, this cluster would be important for understanding transmission dynamics, even if some transmission links were missing.

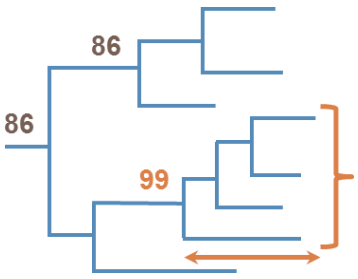


Figure 3.1: Diagram of a cluster of sequences in a phylogenetic tree defined by high confidence in the grouping (bootstrap) and low within cluster genetic distance. Sequences within a cluster are most likely epidemiologically related.

Two sequences taken from the population are on average 10% different within a subtype [188], but only 5% within a cohort of recently infected patients in San Diego [187]. Plotting the distribution of pairwise genetic distances among all UK sampled HIV sequences reveals two major peaks and one minor peak [108]. The two major peaks represent inter and intra-subtype genetic distances. The smallest peak represents comparisons between significantly more closely related sequences and is situated around 4.8%. Based on these findings, analyses of the UK HIV RDB have used genetic distance cut-offs of 1.5% to define infections that are likely to be direct links (strict definition) and 4.5% to capture transmission chains (relaxed definition). Those two cut-offs are used throughout this thesis. Bootstrap cut-offs are less important, as they are heavily influenced by the background of sequences in the tree. In this Chapter, I test a number of bootstrap cut-offs as sensitivity analyses to evaluate the robustness of the conclusions.

Evidently, clusters complying to a strict cluster definition are more likely to reflect true linkages (high specificity) but strict definitions risk missing true linkages (low sensitivity); while more relaxed definitions will incorrectly group sequences into clusters (low specificity) but are more likely to capture all true linkages (high sensitivity). In Chapter 6, I evaluate the sensitivity and specificity of cluster definitions on simulated data for the first time.

A combination of genetic distance and bootstrap is standard, but other cluster definitions have been developed and used by other groups. The Swiss HIV Cohort Study classifies as clusters monophyletic groups of at least ten sequences, of which ≥80% are Swiss [105]. This definition is appropriate for the data because the number of Swiss sequences is so small compared to the number of background sequences used (despite coverage being so high in Switzerland). They are interested in Swiss-based transmission and by enforcing a minimum size, clusters with ≥80% Swiss sequences are very unlikely to have occurred by chance. Prosperi *et al.* developed an algorithm which bases the cluster cut-off on the distance contained within the tree as a whole [174]. The user selects a t-percentile threshold and sequences are grouped into clusters if the median of their genetic distances falls below the t-percentile threshold of the full

distribution. One huge advantage of this algorithm is that it has been implemented as a user-friendly program (PhyloPart) that processes large trees containing 100,000s of sequences. Indeed another problem with cluster identification methods until this point was that no automated software could deal with the number of sequences available. In this Chapter, I discuss a program developed with colleagues to process 100,000s of sequences based on the more widely used genetic distance and bootstrap cut-offs.

It is noteworthy that all the definitions delineated above require clusters to be monophyletic. There are instances where this prerequisite will lead to epidemiologically linked infections not clustering. For example if sampling coverage of a large population were 100%, all sequences should realistically be linked to at least one other but the mean genetic distance would probably exceed 4.5% (and so the full tree would not be classed as a cluster). As an alternative, Wertheim *et al.*'s single linkage approach links sequences to each based solely on their pairwise genetic distance, without the need for a tree at all [99]. Another linkage method developed by the Leigh Brown group takes advantage of both the distances between sequences and the information in the tree, by linking sequences together based on their tMRCA in time resolved phylogenies [140]. As well as circumventing enforced monophyly, both the Wertheim and Leigh Brown methods allow for the connections between sequences to be visualised in a network. The Leigh Brown method is developed further in Chapter 6.With sequence data sets used for the reconstruction of phylogenies now containing tens of thousands of sequences, identifying clusters manually is infeasible. Using in-house pipe lines for detecting clusters is possible, but in order to compare results between studies, freely available software tools would be advantageous. Based on the support and genetic distance criteria commonly used, the Cluster Picker (CP) tool identifies clusters in phylogenetic trees. The Cluster Matcher (CM) describes identified clusters epidemiologically as well matches clusters between phylogenetic trees. I use the tools to examine subtype B cluster dynamics in the UK and compare CP performance to that of other available software.

3.3 Methods

The Cluster Picker (CP) and the Cluster Matcher (CM). The CP and CM were developed in Java 1.6 and are platform-independent. Both programs were released publicly as freely available software with accompanying tutorials, manuals and test files. Source code is available on Google code (http://code.google.com/p/cluster-picker-and-cluster-matcher/) under GNU GPLv3.

3.3.1 The Cluster Picker (S.J. Lycett)

3.3.1.1 Objective

The CP is a JAVA program that identifies clusters of sequences in a phylogenetic tree based on support for the node (bootstrap or posterior probability) and the maximum pairwise genetic distance within the cluster.

3.3.1.2 Input

The CP takes as input a set of aligned sequences in fasta format and a Newick tree built from those same sequences, with support values on the nodes. The user inputs the desired node support threshold and maximum genetic distance for clusters, as well as an initial support threshold for splitting the tree prior to analysis.

3.3.1.3 Algorithm

The CP utilizes a depth-first algorithm to explore the tree: starting at the root and working its way along each branch before backtracking when a leaf is reached. In order to minimize the number of pairwise distances computed (thus reducing running time), the tree is initially split. The user inputs an initial node support threshold, and starting from the root, the tree is divided into subtrees supported at this threshold. Further analyses will take place only within these subtrees; therefore, the initial support threshold must necessarily be smaller than or equal to the cluster support threshold. Starting from the root of the subtree, the CP proceeds to the first node exceeding the bootstrap support threshold. All sequences within the group are identified and their pairwise genetic distances are calculated. If the largest of these is smaller than or equal to the user-input maximum genetic distance threshold, the group of sequences is

identified as a cluster. If the maximum pairwise distance is larger than the threshold, the cluster is rejected and the algorithm proceeds to the next supported node and repeats the same analysis. When a leaf is reached, the CP backtracks to the last node whose children have not been fully analysed. When the algorithm has analysed the entire tree, a list of clusters matching the user-input criteria is generated. Note that because the algorithm proceeds from the root towards the tips, nested clusters are not identified and do not appear in the final list. For pseudocode, see Appendix 6.

3.3.1.4 Output

The CP outputs a log file listing for each cluster: cluster number, cluster size, maximum genetic distance within the cluster, support value and tip names. Also output are a fasta file in which sequence names are preceded by their cluster number and two trees, one in newick format and one in FigTree format (http://tree.bio.ed.ac.uk/software/figtree/). In both trees sequence names are preceded by cluster name, and in the FigTree file, sequence names are coloured by cluster.

3.3.2 The Cluster Matcher (E. Hodcroft)

3.3.2.1 Objective

The CM is a JAVA program which links sequences in clusters to epidemiological data. An automated tool for performing this task is necessary because of the size of the datasets analysed.

3.3.2.2 Input

The CM takes as input the newick files output by the CP and corresponding annotation files. The annotation file allows the user to select clusters based on those annotations. For example, if the annotation file contains risk group data, the CM could output only clusters containing at least 50% of sequences from MSM. The user can also choose to output clusters based on whether they contain a specified minimum number of sequences.

3.3.2.3 Algorithm

Traversing from root-to-tip, the CM first identifies all clusters present in each dataset, linking every sequence in a cluster to any epidemiological information provided. For each cluster, information is retrieved including its size, number of matching sequences, and the distribution of epidemiological traits attached to its sequences. This allows the clusters to be easily filtered when the user specifies cluster selection criteria, and is used to generate summary information for each cluster.

3.3.2.4 Output

The CM outputs a FigTree file for each cluster, as well as a log file detailing settings and summarizing results. The FigTree file contains two trees: the full tree with the cluster of interest highlighted and a zoom into each of those clusters, allowing for the visualization of single clusters within large phylogenies. In addition, the CM generates a spreadsheet containing the composition for each cluster identified by the CP; for example, the number of individual from each risk group, of each sex.

3.4 Analysis

3.4.1 Data

Publicly available HIV *pol* sequences from the UK HIV RDB (see Methods 2.1.2) were used to evaluate the CP and CM (Genbank IDs: EU236439 – EU236538 [3], GQ462027 - GQ462532 [18], JN100661 – JN101948 [23]). Sequences were subtyped in SCUEAL (see Methods 2.2.1.2), and drug resistance sites were stripped (Methods 2.2.1.3). In parallel, all unique subtype B *pol* sequences were downloaded from LANL (section 2.1.3) to perform a speed comparison between the CP and PhyloPart.

3.4.2 Effect of cluster thresholds on cluster distribution

Using the CP, I evaluated the effect of different cluster thresholds for genetic distance and cluster support on cluster identification among the UK subtype B sequences downloaded. One hundred replicate alignments were generated and a maximum likelihood tree with bootstraps was reconstructed in FastTree v2.1.4 [163] with a subtype C reference sequence (GenBank accession number: AY772699). I varied

genetic distance threshold between 1.5% and 7.5% and bootstrap between 70% and 99%, each time outputting the clusters. From the list of clusters I generated cluster distributions (the number of clusters of each size).

3.4.3 Automated analysis of cluster dynamics

Using both the CP and the CM I reconstructed cluster dynamics over time, analysing 409 non-B UK sequences as well as the 1381 subtype B sequences. Sequences were linked to sampling date and location information in the UK HIV RDB. An initial phylogenetic tree was constructed containing 1212 sequences of all subtypes collected up to 2005. Clusters supported by a bootstrap ≥90% and maximum genetic distance ≤4.5% were identified. The CM was used to sort clusters in 2005 based on whether they contained sequences from a single sample location ("single" origin) or more ("multiple" origin). A second tree was built from the entire dataset of 1790 sequences and clusters matched between the early and late trees. Cluster growth was then calculated for each cluster as the number of new sequences per initial sequence [26]. Patterns of change of single origin versus multiple origin clusters were compared (see Appendix 3) [30].

3.4.4 Comparison with PhyloPart

The CP was compared to PhyloPart, a recently released software tool for the identification of clusters [19]. PhyloPart generates the pairwise distance distribution for a tree and identifies a group of sequences as a cluster if the median of their genetic distances is below a user-input t-percentile threshold of the whole-tree distance distribution. The rooted subtype B tree containing 1381 sequences was analysed in PhyloPart, varying the t-percentile threshold for cluster identification from 1% to 30%. I compared cluster distributions to those produced by the CP.

In order to evaluate the performance of the two tools, I ran them both on 18 data sets sized 1000 to 18,000. 18,436 subtype B sequences were downloaded from the LANL database and a random sample of 18,000 sequences was selected to construct a maximum likelihood bootstrapped tree in FastTree. Then, sets of 1000 tips were dropped sequentially from the tree to generate trees with variable number of tips (see

Appendix 3). As PhyloPart does not print time to completion, it was launched from within a python script with an additional function to calculate running time.

3.5 Results

3.5.1 Clusters are robust to changes in genetic distance thresholds

Of 1831 downloaded sequences, 1381 unique subtype B sequences were used to examine the effect of cluster definition on cluster distribution using the CP. Although the phylogenetic tree contained a reference subtype C sequence, this outgroup was removed prior to analysis with the CP using the APE package v.3.0-8 in R [154, 189]. Initially, I fixed the bootstrap threshold in the CP at 90% and varied within-cluster maximum genetic distance between 1.5% and 7.5%. Between 4.5% and 7.5%, I found that for the most part, the same clusters were identified (Figure 3.2A). Within this range, the number of clusters stabilized around 128 (ranging from 126 to 131), with 2/3 containing only two sequences. At a genetic distance of 1.5%, only 63 clusters were identified. The proportion of sequences in clusters and average cluster size both increased as the genetic distance threshold was increased (Figure 3.3A). At a maximum genetic distance of 4.5%, 25% of sequences clustered, identical to the proportion found after a time-resolved analysis of the same sequences [108]. Beyond 4.5%, the ratio of these two measures became constant, indicating that as the genetic distance cut-off was relaxed sequences were being added equally to all clusters.

The effect of varying the cluster bootstrap threshold was different; fixing the genetic distance at 4.5%, the proportion of sequences in clusters decreased gradually as bootstrap thresholds were increased from 70% to 99% (Figure 3.2B and Figure 3.3B).

3.5.2 Automated analysis of cluster dynamics

The cluster dynamics of 1381 subtype B and 409 non-B sequences between 2005 and 2007 were analysed. These included 63 A subtypes, 219 C and 127 other non-B. In the phylogenetic tree constructed from sequences collected up to 2005, 148 clusters containing 431 sequences (35.6%) were identified. One hundred and eight of these clusters were pairs, with the largest containing seventeen sequences. A second tree was

built from the entire dataset of 1790 sequences and clusters were matched between the early and late trees so that cluster changes could be described. In support of the initial cluster definition, the genetic distance of the new clusters increased above 4.5% only in two clusters despite the addition of 578 sequences, while bootstrap dropped below 0.90 only for six clusters.

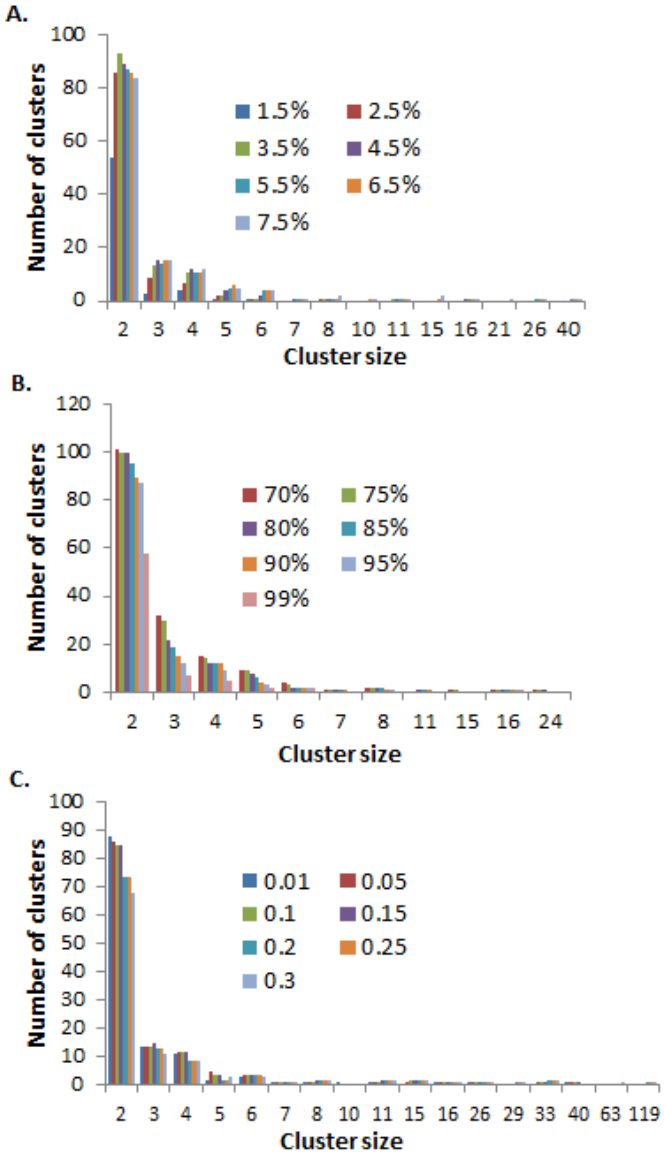


Figure 3.2: Cluster distributions. 1381 subtype B UK sequences from NCBI were processed (A) through the Cluster Picker, with bootstrap support threshold fixed at 90% and maximum genetic distance threshold varied between 1.5% and 7.5%, (B) through the Cluster Picker with maximum genetic distance threshold fixed at

4.5% and bootstrap support threshold varied between 70% and 99%, and (C) through PhyloPart, with the t-percentile threshold varied between 1% and 30%.

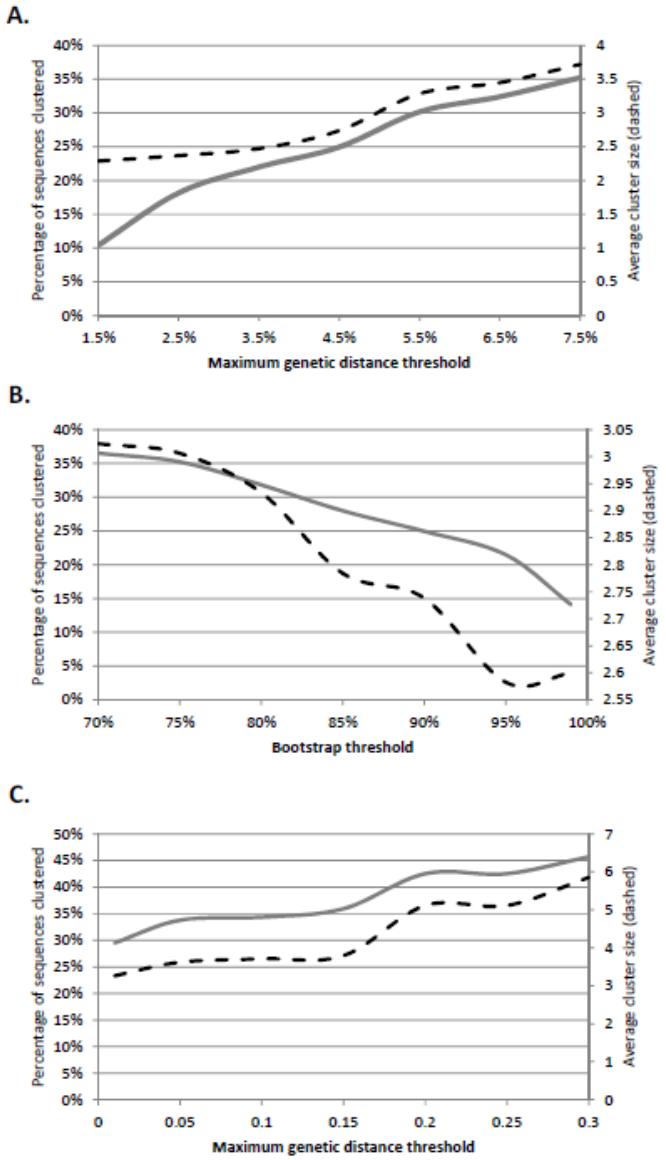


Figure 3.3: Clustering patterns. 1381 subtype B UK sequences from NCBI were processed (A) through the Cluster Picker, with bootstrap support threshold fixed at 90% and maximum genetic distance threshold varied between 1.5% and 7.5%, (B) through the Cluster Picker with maximum genetic distance threshold fixed at 4.5% and bootstrap support threshold varied between 70% and 99%, and (C) through PhyloPart, with the t-percentile threshold varied between 1% and

30%.distribution for varying bootstrap thresholds. For each threshold, we plotted the percentage of total sequences in clusters (grey line) and average cluster size (dashed line).

Sample location information was used to sort clusters in 2005 based on whether they contained sequences from a single sample location ("single" origin) or more ("multiple" origin). The UK HIV RDB categorizes geographical origin into 17 areas, all of which were represented in this dataset. A large proportion of sequences originate from the London area (one centre). Of 148 clusters, 63 were thus classified as multiple origin and 85 as single origin. Mean cluster growth differed significantly between single and multiple origin clusters (0.155 vs. 0. 302, respectively, Kruskal-Wallis test: p=0.0016).

3.5.3 Comparison with PhyloPart

The subtype B tree was analysed in PhyloPart, varying the *t*-percentile threshold for cluster identification from 1% to 30%. Upon examination of the output, it appeared that this range reflected median genetic distances within clusters from 4.5% to 9% in the data. Once again, cluster distribution was not very much affected by the cut-off (Figure 3.2C), but the proportion of sequences in clusters and average cluster size increased as cluster definition was relaxed (Figure 3.3C). As a *t*-percentile threshold of 0.01 and 0.05 corresponded to genetic distance cut-offs of 4.5%, and 6.5%, respectively, the CP and PhyloPart output were compared in more depth at each of these two matched thresholds. Each time, the number of clusters and the cluster distributions were near identical (KS test, p=0.9998 and p=1 for 4.5% and 6.5% respectively). However, as expected, individual cluster sizes were significantly reduced when maximum within cluster genetic distance was used instead of median (Figure 3.5; one-sample sign test, $p=6.1*10^{-5}$ and p=0.03 for genetic distances of 4.5% and 6.5%, respectively). At 4.5%, the CP and PhyloPart agreed on the cluster membership status of 94.5% of sequences and at 6.5% of 97.2% of sequences.

We ran both programs on datasets up to 18,000 sequences to compare performance. Both programs were able to process trees with up to 17,000 sequences in less than an

hour on a desktop (Table 3.1), although PhyloPart did not terminate on the largest dataset (n=18,000 sequences). The CP completed on average three times faster than PhyloPart.

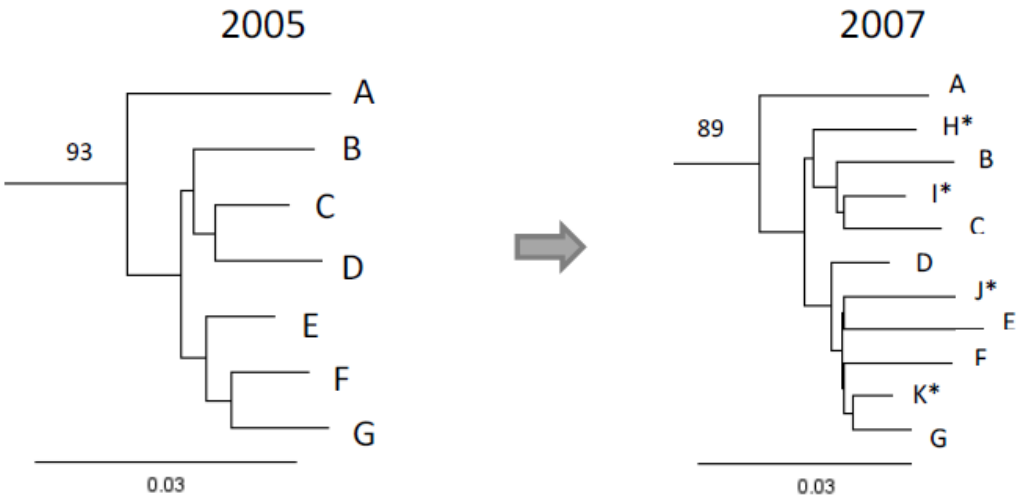


Figure 3.4: Dynamics of a single cluster 2005-2007. In this example, the cluster identified in 2007 no longer matches the initial cluster definition as bootstrap support has dropped from 93% to 89%. Sequences A to G are those already in the cluster in 2005, starred sequences (H to K) have been added to the cluster in the intervening years.

3.6 Discussion

The tools that presented here can be used to investigate the dynamics of pathogen transmission. The CP is able to rapidly identify clusters in an automated way in large datasets, based on criteria demonstrated previously to accurately delineate epidemiologically relevant clusters [144]. Because in many cases cluster studies seek to combine genetic with epidemiological or clinical data (such as risk group or stage of infection), we have also made available the CM, which links clusters between runs and to epidemiological data. In contrast to other tools available for the analysis of trait-annotated phylogenies [190, 191], the CM does not require any assumptions to be made about the heritability of the traits examined. As an example, I used the tools

together to investigate the dynamics of single vs. multiple origin HIV clusters in the UK.

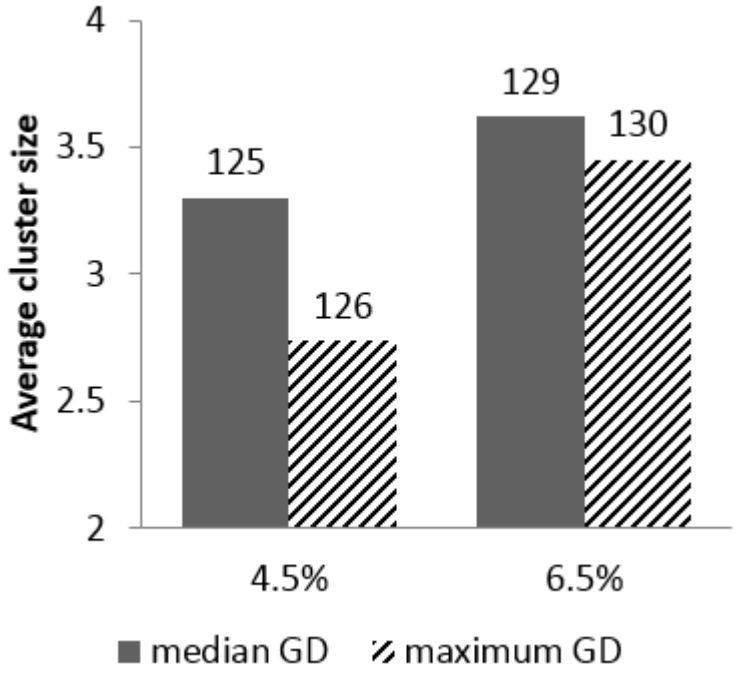


Figure 3.5: Average cluster size according to clustering method. At thresholds of 4.5% and 6.5%, PhyloPart (in grey, median GD) and the Cluster Picker (dashed, maximum GD) identified nearly exactly the same number of clusters (numbers above the columns) but PhyloPart clusters were on average larger. GD genetic distance.

There was remarkable consistency in the clusters identified at maximum genetic distances between 4.5% and 7.5%, as has been previously observed [108]. We conclude that these clusters represent well-delineated epidemiological units in the UK HIV epidemic. In contrast, when the maximum genetic distance threshold was decreased to 1.5%, only half of the clusters were identified. These clusters defined by such a short distance will reflect recent transmissions and frequent samplings [192]. In contrast, the UK HIV RDB contains mostly sequences from chronically infected patients, many of whom were first sequenced long after infection, and so in order to identify relevant clusters, a threshold of 4.5%, as we have used before [108, 109], is

more appropriate. The effect of the bootstrap threshold was less evident, and so I conclude that genetic distance is the key parameter for epidemiologically relevant clustering. I stress however that the present analysis alone is not sufficient to yield a reusable definition of cluster threshold parameters, as the data set of publicly available sequences was too small for extensive testing.

Table 3.1: Time to completion (in seconds) of the Cluster Picker and PhyloPart for data sets of increasing sizes

Number of sequences	Cluster Picker (s)	PhyloPart (s)
1000	13.098	8.913
2000	36.137	44.151
3000	68.772	112.729
4000	115.618	672.085
5000	173.584	1447.047
6000	244.290	1713.749
7000	328.651	2190.336
8000	419.369	1081.785
9000	526.070	1043.838
10000	658.607	2321.955
11000	769.469	2343.197
12000	911.086	3061.134
13000	1059.509	2851.417
14000	1228.151	2078.609
15000	1383.366	2625.491
16000	1581.351	2797.329
17000	1775.639	3047.713
18000	1990.372	NA

Notes: Both programs were run on a Windows desktop computer with an Intel Core i5-2400 3.10GHz with 4 processors, reserving 1.5G of heap space. PhyloPart did not complete on the desktop computer with n=18,000 sequences as heap space could not be increased. Settings were left as default in the Cluster Picker and set at t=0.05 in PhyloPart. For 10,000 sequences, program specific RAM usage was 265,000K for PhyloPart and 100,000K for the CP. Computational complexity for this data set approximates O (n2).

The CP uses maximum within cluster genetic distance, while PhyloPart, another recently released sequence clustering tool, uses the median. In previous studies, the Leigh Brown group has identified clusters in trees based on mean within cluster distance [109, 140]. However, it was decided to use maximum genetic distance in the CP for three reasons. First, maximum genetic distance (as well as median genetic distance) is less affected by the number of sequences within a cluster (which can be the result of more or less intensive population sampling and contact tracing). When the mean is used, the distance is normalized by the total number of sequences in the cluster, potentially leading to clusters in which most of the sequences are very close together but one sequence is only distantly related to the group. In agreement with this prediction, in a longitudinal analysis the genetic distance threshold did not have to be increased in 2007 to capture most 2005 clusters despite the additional of a large number of sequences. Second, maximum genetic distance is a metric more comparable to the time depth used to identify clusters in BEAST [140]. Third, maximum genetic distance is faster to compute, improving program efficiency. Another difference between the programs is that distances are calculated de-novo from the sequences in the CP, while in PhyloPart, the patristic distances are used. The use of patristic distances is probably preferable (because distances but the CP calculates distances from sequences to do automatically do what was previously done manually by our group. In this way its results are consistent with previously established thresholds and results [108]. Cluster definition in PhyloPart is a function of the whole tree: a subtree is classified as a cluster if its median genetic distance is smaller than a percentage of the whole tree. However, the user-specified genetic distance threshold in the CP allows external information to be incorporated into the definition, such as the average observed distance within transmission pairs if that is available. This strategy was chosen because it is the most widely used definition; in fact, previous studies have demonstrated epidemiologically related viral sequences had less than 4.8% nucleotide substitutions between them [108]. Similarly, because studies vary in the bootstraps they use for support of clusters, we left this as a flexible option for the user to choose.

For data sets containing up to 17,000 sequences, both PhyloPart and CP yield results on a desktop in reasonable time. Theoretically, PhyloPart will slow down in large

datasets, as it calculates all pairwise distances then stores them, so they can be accessed each time they are needed. This is an advantage for smaller datasets and speeds up processing, but for large trees, the time to generate matrices of all pairwise distances increases as a polynomial function of the number of sequences n (n (n-1)/2 computations). The CP calculates pairwise genetic distances within a potential cluster as required even if those distances were already calculated when the parent node was tested (and rejected). Nevertheless, the CP was not slower than PhyloPart on small datasets and in fact completed on average three times more rapidly. On large trees, it becomes faster to calculate subsets of pairwise genetic distances only within potential clusters, even if this must be repeated several times. Another alternative, not explored here, is the single-linkage approach proposed by Wertheim *et al.* [99], which does not require a phylogenetic tree and calculates pairwise distances only once. The absence of a requirement for building a tree means that the single-linkage method will perform much more rapidly than either Cluster Picker or PhyloPart on large datasets. With expanding sizes of HIV-1 data sets and other fast evolving pathogens, there is increasing need for new and faster algorithms.

Our longitudinal cluster analysis demonstrated differences in cluster growth between clusters that were confined to single UK locations in 2005, and those that already contained sequences from several locations across the UK. If confirmed, these results suggest that targeting interventions on individuals within multiple origin clusters to prevent onward transmission would yield disproportionate results. Such real-time analyses are made possible by the CP and CM. As the purpose of this Chapter was to demonstrate the functionality of the CP and CM, a simple example was used. We hope that others will use the tools in more elaborate ways to truly provide insight into the dynamics of HIV transmission, as well as other infectious diseases. Concerning cluster dynamics, we note that new sequences added to clusters do not necessarily reflect new infections: they could reflect new diagnoses within the time frame, and one potential explanation of the observed cluster growth may indeed be referral-based testing.

The automation of cluster picking and matching with epidemiological information is a necessary advance as pathogen sequence databases have become too large to analyse

manually. The *pol* region of HIV is routinely sequenced for clinical purposes, and several European countries have created central repositories for the sequences. These data, combined with the tools we have made available, offer opportunities for the real-time surveillance of the HIV epidemic. I hope that by providing strategies for cluster identification and description, these user-friendly tools will facilitate comparisons of epidemics between studies and countries. In Chapter 1, I use these tools and automated pipeline to compare the UK and Swiss epidemics, without the two countries having to share data.

4 A DIRECT COMPARISON OF TWO DENSELY SAMPLED HIV EPIDEMICS: THE UK AND SWITZERLAND

4.1 Abstract

The UK and Swiss HIV epidemics have both historically been driven by transmission of subtype B among MSM. The Swiss population is 1/8 the size of the UK and HIV prevalence in Switzerland is nearly double. Both epidemics are densely sampled, by the UK HIV RDB and the Swiss HIV Cohort Study respectively. Previous independent analyses have suggested dramatically different epidemic dynamics.

Coordinated analyses using a common bioinformatics pipeline to compare HIV transmission patterns were performed. Sequence clusters were identified in maximum likelihood phylogenetic trees of subtype A1, B and C *pol* sequences against a background of global sequences at a range of bootstrap (70%-95%) and genetic

distance (1.5% and 4.5%) thresholds. Degree distributions were generated for each risk group and I compared distributions between countries using KS tests, DDQC and bootstrapping. Univariate and multivariate logistic regression was used to predict cluster membership based on country, sampling date, risk group, ethnicity and sex.

Over 8000 subtype B sequences from Switzerland and >30000 from the UK were analysed. A genetic distance of 1.5% yielded mainly pairs in both. After adjusting for sample dates, the Swiss epidemic was more clustered at 1.5%, but degree distributions did not differ significantly between the two countries. At 4.5%, the UK was more clustered and the degree distributions for MSM, heterosexuals and the population as whole differed significantly by the KS test. Only the heterosexual distributions varied based on the DDQC test, and dropping high degree heterosexuals clustered exclusively with MSM eliminated this difference. Because the KS test is sensitive to variation in scale between networks compared, I tested whether the differences were due to different epidemic sizes by jackknifing the UK epidemic and showed the underlying degree distributions were the same.

Despite differences in risk group composition, Swiss and UK clustering patterns are similar. Differences observed are largely explainable by the distinct sizes of two epidemics. The underlying epidemic dynamics driving new infections are thus likely to be similar.

4.2 Introduction: The UK and Swiss epidemics

The UK and Swiss epidemics share a similar history. HIV was introduced into UK MSM around 1980 [89], probably from the USA [90]. Until the early 1990s the UK epidemic was dominated by the transmission of subtype B among MSM and PWID (see Introduction 1.1.2 and 1.4.2). In the 1990s, cases among heterosexuals became more common [93] and these were often linked to immigration [90]. Between 2002 and 2010, 48% of HIV diagnoses were among heterosexuals [93]. Since the 1990s, non-B subtypes have been increasing within the UK [94, 95]. For some time the UK epidemic was compartmentalised by subtype and risk group, with B circulating among MSM and non-B among heterosexuals. Yet recently a small but significant proportion

of non-B subtypes have been acquired through sex between men [93] (see Chapter 5). 100,000 people are estimated to be living with HIV in the UK (0.15% prevalence), a quarter of whom are unaware of their infection [6]. HAART became available to UK residents in 1996 and universally available (regardless of citizenship or immigration status) in 2012. HIV positive people on successful treatment in the UK have a normal life expectancy [45].

Switzerland was the European country with the highest HIV prevalence at the beginning of the HIV epidemic in the 1980s [193]. Initially HIV was spread through MSM and PWID [194] with heterosexual transmission only starting to play a role after the mid-1980s. The number of new diagnoses steadily declined in the nineties, mainly owing to the introduction of needle exchange programs in PWID, heightened awareness, wide scale HIV testing, and the introduction of HAART. The number of new HIV diagnoses in Switzerland has fluctuated since 2000 with no clear time trends. The epidemic is currently thought to be fuelled through MSM transmission with fewer heterosexuals and almost no PWID cases [97, 105] (less than 3 per year in the last 5 years [195]).

The two countries also have two of the largest and most extensive HIV sequence databases in the world: the UK HIV RDB (see Methods 2.1.2) and the Swiss HIV Cohort Study (SHCS; http://www.shcs.ch/; see Methods 2.1.4), which contain *pol* sequences from at least 60% of diagnosed patients within each country. In the early 2000s, less than half of diagnoses entered the databases, while since 2008, this has reached 80-90%. Based on estimates of total numbers of infected people (including undiagnosed), the databases currently contain around 50% of sequences from all alive infected individuals within each country. These databases have allowed Switzerland and the UK to lead the way in the field of population-level HIV research, including HIV phylogenetics.

One important focus of phylogenetic analyses is on clusters: groups of sequences within the tree that are more related to each other than to the rest of the tree (see section 3.2). Clustered sequences are likely to be represent epidemiologically linked infections with short durations of time between transmissions and thus clusters represent the

leading edge of the epidemic. High levels of clustering indicate rapid transmission and that transmission partners are being captured.

However, these analyses have suggested these two countries' epidemics have different structures (see section 1.4.3 for details), including different proportions of sequences in clusters. In the UK, 24%, 40% and 22% of patients infected with HIV-1 subtypes B [108], A1 and C [109] respectively, have been found in clusters whereas the corresponding numbers for Switzerland are 55%, 21% and 16% [97, 105]. Clearly these different results arise, at least in part, due to the different cluster definitions (see section 3.2) used by the two teams: the Swiss cohort defines a cluster as a group of ≥10 sequences supported by a bootstrap-value >80% [105], while the UK looks at clusters containing at least three sequences with a within cluster genetic distance ≤4.5% and supported by bootstraps ≥90% [108]. Another reason for the disparity between the two countries could be differences in sampling procedures. However it is also possible that the contact and transmission processes between Switzerland and the UK differ because of the difference in size between the two countries and their geography.

Access to the data from these national cohorts is subject to restrictions. Even though it is possible to submit proposals to carry out analyses, one limitation is that national data must not leave the country. UK data must be analysed in the UK and Swiss data must be analysed in Switzerland. Because analyses have been conducted according to in-house bioinformatics pipelines, the differences between the two epidemics have never been elucidated. Here, I present an analysis of the two epidemics conducted using a bioinformatics pipeline applied in parallel in the two countries. In brief, phylogenies were reconstructed in each country independently (using the same LANL background sequences) and processed using the CP and CM. I wrote python and R scripts to merge and modify the files output by the CP and CM (see Appendix 3). The final output was a spreadsheet containing a cluster on each line with details on number of sequences (national and LANL), risk group, ethnic group and gender composition. These files were exchanged and data from both countries were analysed independently in each country. I generated degree distributions and compared those across countries

(including KS tests, DDQC and bootstrapping), while Mohaned Shilaih constructed the linear regression model. I wrote the manuscript and created all the figures.

4.3 Methods

4.3.1 Data

4.3.1.1 Switzerland (CH)

9232 HIV *pol* sequences were retrieved from the Swiss HIV Cohort Study resistance database (SHCS RDB, 2014).The SHCS RDB aggregates all HIV resistance tests performed across Switzerland by all laboratories involved in HIV resistance testing in Switzerland. SmartGene is responsible for sequences and meta-data storage and management (http://www.smartgene.com). The RDB is part of the SHCS, which is an ongoing national clinical cohort of HIV patients of age 16 and above with biannual follow up (http://www.shcs.ch). The sequences were assigned subtypes using the REGA algorithm [152, 196]: subtype B (8390 sequences, ~91%), A1 (435 sequences, ~5%), and C (419 sequences, ~5%). The earliest available sequence for each individual was used in the present analysis.

4.3.1.2 UK

63,065 HIV *pol* sequences were obtained from the UK HIV RDB 2014 download (see Methods 2.1.2). Subtypes B (31,649 sequences, 48.6%), C (15,864, 24.4%), and A1 (2512, 4.0%) were analysed here.

4.3.1.3 Background sequences

LANL background sequences were selected as described previously (see Methods 2.1.3). For this step, UK sequences were removed from LANL alignments before the UK Viroblast run, and Swiss LANL sequences were removed before the Swiss run.

4.3.2 Tree Building and Cluster Picking

Duplicate sequences were removed from all datasets. Maximum likelihood phylogenetic trees were constructed for each country and subtype separately (six trees in total) using FastTree version 2.0 [163] with 100 bootstrap replicates. A range of

cluster definitions were used. Clusters were picked if they were supported by bootstrap thresholds of 70%, 80%, 90% and 95% and maximum genetic distance (GD) of 1.5% or 4.5% (8 thresholds total) [197]. In addition, clusters in the Swiss tree were selected to contain at least 80% SHCS sequences, and clusters in the UK trees at least 80% UK sequences. In a separate analysis, the 80% criterion was removed (i.e. all clusters with least one UK or Swiss sequence were examined) to investigate mixing between national and foreign sequences. The automated pipeline included analysis with the CP and CM as well as processing through python and R scripts (see Appendix 3). The CP was modified to recognise IUPAC nucleotide ambiguity codes as matches, increasing clustering by around 15% in both datasets. From this output file, I generated degree distributions (the number of links for each node). As output files contained the risk group composition for each cluster, it was possible to break down the degree distribution by risk group. Statistical frameworks exist to formally compare degree distributions but not cluster size distributions. In addition degree distributions can be bootstrapped to test the robustness of conclusions. Bootstrapping was performed to simulate the effect of the sampling process. Nodes were sampled with replacement from the network with information on their cluster membership. Nodes sampled with the same cluster membership were linked together in each bootstrapped network, so that clusters sometimes increased in size, sometimes decreased in size or otherwise disappeared, and degree distribution was re-estimated each time. Jack-knife resampling where the number of nodes sampled was smaller than the full network size was also performed.

4.3.3 Statistical analysis

The number of sequences clustering at different thresholds between the two epidemics was compared using Fisher's exact test with Bonferroni correction for the number of tests. Degree distributions were compared using KS tests [182] and the DDQC algorithm [183] (see Methods 2.2.5.4). The DDCQ corrects for differences in population size when comparing degree distributions while the KS test does not.

The UK subtype B dataset as a whole was 3.75 times larger than the Swiss dataset (Table 4.1). The UK MSM dataset was 5.7 times larger and the UK HET dataset was

1.3 times larger. To investigate the effect of this difference in size (the pool of possible infectors), the UK degree distribution was jack-knife sampled to match the size of the Swiss epidemic. One hundred jack-knife replicates were generated, and in each replicate the degree distribution was re-estimated based on the links present in the sample.

Mohaned Shilaih used a logistic regression model to characterise the factors influencing clustering in the two countries. The model was applied with cluster membership as the outcome variable and with the country of origin (UK or Switzerland) as the main exposure variable. Sampling dates, risk group, sex, and ethnicity were adjusted for.

All statistical analyses were conducted in R [154].

Table 4.1: Baseline demographics of the two datasets

		UK			CH		
		A1	**B**	**C**	**A1**	**B**	**C**
Sex	Female	1442 (57.5%)	2327 (7.4%)	3869 (62.2%)	269 (62%)	1812 (22%)	237 (57%)
Risk	MSM	170 (6.8%)	22157 (70.4%)	660 (4.2%)	21 (5%)	3914 (47%)	25 (6%)
	HET	1718 (68.5%)	3399 (10.8%)	11893 (75.2%)	366 (84%)	2626 (31%)	357 (85%)
	PWID	92 (3.7%)	873 (2.8%)	128 (0.8%)	12 (3%)	1553 (19%)	9 (2%)
	Other/ NA	527 (21%)	5021 (16%)	3134 (19.8%)	36 (8%)	287 (3%)	28 (7%)
Ethnicity	White	494 (19.7%)	22724 (72.3%)	1724 (10.9%)	194 (45%)	7333 (87%)	118 (28%)
Total		**2507**	**31450**	**15815**	**435**	**8390**	**419**

4.4 Results

4.4.1 Comparison of the two HIV+ populations (Baseline demographics)

In the final analyses, the Swiss datasets contained 1374 subtype A1 sequences (435 Swiss, 939 LANL), 15043 B (8390 Swiss, 6653 LANL) and 1571 C (419 Swiss, 1152 LANL). The UK datasets comprised 4421 A1 (2507 UK, 1914 LANL), 38863 B (31450 UK, 7413 LANL) and 22027 C (15815 UK, 6212 LANL). The Swiss epidemic comprised more PWID, fewer self-identified heterosexuals and more patients of White ethnicity (Table 4.1). These differences were in part a result of the different subtype composition across the two countries, but even within subtype B there were notable differences (Table 4.1). The subsequent analysis mainly focused on subtype B.

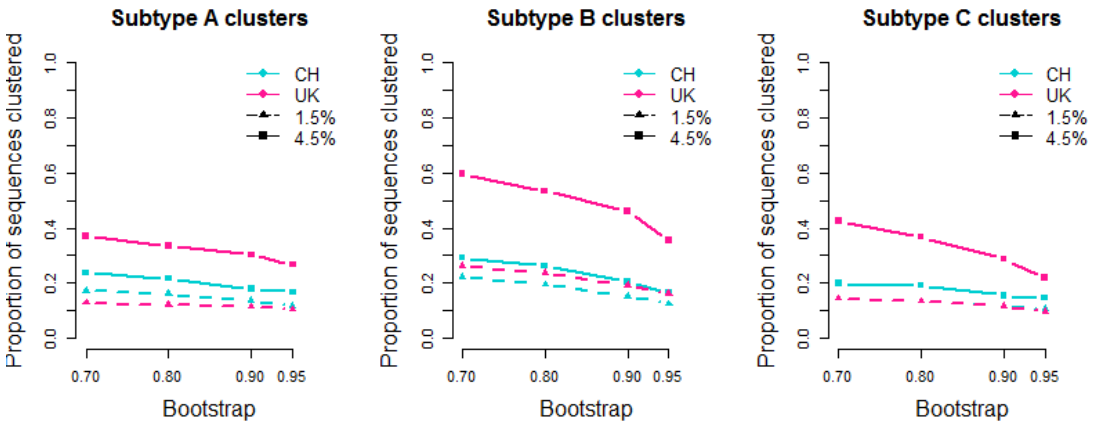


Figure 4.1: Proportion of UK (pink) and Swiss (blue) sequences in clusters. At different genetic distance (1.5% and 4.5%) and bootstrap (70%, 80%, 90%, 95%) thresholds. CH Switzerland.

In the UK, a partial *pol* sequence was available for at least 30% of HIV diagnoses from 2003 and for over 80% of diagnoses from 2007 onwards. In the SHCS, >60% of the HIV+ population has an HIV sequence available. Sequence dates go back to 1995 owing to retrospective sequencing of samples from the SHCS Bio-bank. Accordingly, sequences are available for 79% of patients in the SHCS diagnosed with HIV after 1995. In both Switzerland and the UK individuals for whom a sequence and epidemiological data were available broadly matched the characteristics of the HIV

diagnosed population as a whole, in terms of risk group, sex, ethnicity and age distribution [148, 198].

Table 4.2: Proportion of sequences clustering at different cluster thresholds for the UK and Switzerland (CH) subtypes A1, B and C.

Subtype	Bootstrap	Genetic Distance	Proportion clustered (CH)	Proportion clustered (UK)	p.Bonferroni
A1	0.7	0.015	0.17	0.13	0.53
A1	0.8	0.015	0.16	0.12	0.55
A1	0.9	0.015	0.14	0.11	1
A1	0.95	0.015	0.12	0.1	1
A1	0.7	0.045	0.24	0.37	<0.0001
A1	0.8	0.045	0.22	0.33	<0.0001
A1	0.9	0.045	0.18	0.3	<0.0001
A1	0.95	0.045	0.17	0.27	<0.0001
B	0.7	0.015	0.22	0.26	<0.0001
B	0.8	0.015	0.19	0.24	<0.0001
B	0.9	0.015	0.15	0.19	<0.0001
B	0.95	0.015	0.12	0.16	<0.0001
B	0.7	0.045	0.29	0.6	<0.0001
B	0.8	0.045	0.26	0.53	<0.0001
B	0.9	0.045	0.2	0.46	<0.0001
B	0.95	0.045	0.16	0.36	<0.0001
C	0.7	0.015	0.14	0.14	1
C	0.8	0.015	0.13	0.13	1
C	0.9	0.015	0.12	0.11	1
C	0.95	0.015	0.11	0.09	1
C	0.7	0.045	0.2	0.42	<0.0001
C	0.8	0.045	0.19	0.37	<0.0001
C	0.9	0.045	0.16	0.29	<0.0001
C	0.95	0.045	0.15	0.22	<0.001

Note: p-values shown are from Fisher's exact test, Bonferroni- corrected for multiple comparisons.

4.4.2 Difference in clustering

A larger proportion of subtype B sequences were clustered in the UK than in Switzerland and this difference was statistically significant at all thresholds in the uncorrected analysis (Table 4.2, Figure 4.1). Concordantly, the UK was more clustered in the univariate analysis (Table 4.3) with odds for being in a cluster 4 times higher at 4.5%. At a threshold of 4.5%, subtypes C and A1 were also more clustered in the UK, but there was no significant difference at 1.5%. The proportion of sequences clustering was much higher in the UK than in Switzerland at 4.5% (averaged across bootstrap thresholds) for subtypes B (49% vs 23%) and C (32% vs 17%).

Because of the difference in sampling time distributions and demographics between the two subtype-B datasets, we considered the logistic regression adjusting for those variables. More recent samples were much more likely to cluster than older samples (Table 4.4) and when the model was adjusted for sample date, the Swiss epidemic was more clustered than the UK epidemic at 1.5% (and all bootstrap cut-offs). At 4.5%, clustering remained higher for the UK than for Switzerland but the strength of the association was halved compared with the univariate model.

Table 4.3: Unadjusted logistic regression predicting cluster membership for the UK and Switzerland (Subtype B).

Bootstrap	Genetic Distance	Covariates	OR	2.5%	97.5%
0.7	0.015	UK	1.316	1.243	1.394
0.8	0.015	UK	1.335	1.258	1.418
0.9	0.015	UK	1.396	1.308	1.492
0.95	0.015	UK	1.444	1.345	1.551
0.7	0.045	UK	5.001	4.745	5.272
0.8	0.045	UK	4.119	3.904	4.347
0.9	0.045	UK	3.942	3.722	4.176
0.95	0.045	UK	3.176	2.986	3.381

OR: odds-ratio

Table 4.4: Logistic regression predicting cluster membership in the UK and Switzerland, corrected for sample year (Subtype B).

Bootstrap	Genetic Distance	Covariates	OR	2.5%	97.5%
0.7	0.015	UK	0.66	0.617	0.705
		Sample year	1.138	1.131	1.145
0.8	0.015	UK	0.665	0.621	0.713
		Sample year	1.142	1.134	1.149
0.9	0.015	UK	0.694	0.645	0.749
		Sample year	1.146	1.138	1.154
0.95	0.015	UK	0.706	0.651	0.765
		Sample year	1.153	1.144	1.162
0.7	0.045	UK	2.692	2.538	2.856
		Sample year	1.134	1.128	1.14
0.8	0.045	UK	2.12	1.996	2.252
		Sample year	1.145	1.139	1.151
0.9	0.045	UK	1.968	1.846	2.099
		Sample year	1.156	1.15	1.163
0.95	0.045	UK	1.551	1.448	1.662
		Sample year	1.161	1.154	1.168

OR: odds-ratio

4.4.3 Degree distributions

Degree distributions for each country and risk group were generated based on cluster size distributions and compositions. For example a cluster containing 3 heterosexuals is equivalent to 3 heterosexuals each with degree 2. Degree distributions for the populations as a whole and for each risk group were then compared between the two countries using the KS and DDQC tests. Based on the KS test, there was no difference between the two countries at 1.5%. At 4.5%, distributions differed significantly for

HET, MSM and the population taken as a whole but not for PWID (Table 4.5). The difference appeared to be driven by the longer tail of the UK distributions (Figure 4.2), indicating the existence of larger clusters in the UK.

Table 4.5: Results of the KS test for comparing degree distributions between the UK and Switzerland (Bonferroni-corrected).

Bootstrap	Genetic Distance	Risk group	p.Bonferroni
0.9	0.045	all	0
0.9	0.045	MSM	0
0.9	0.045	HET	<0.0001
0.9	0.045	PWID	1

Note: p values are shown only for bootstrap=0.9 but significance was consistent across bootstraps.

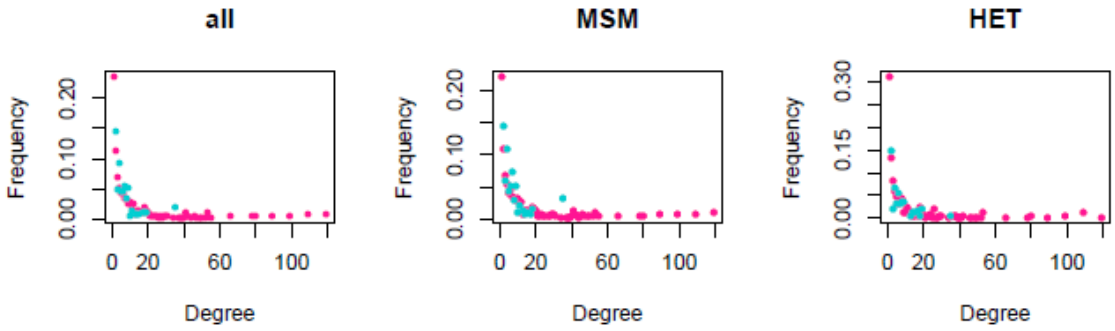


Figure 4.2: Degree distributions of the UK (pink) and Swiss (blue) subtype B epidemics CH Switzerland, MSM men who have sex with men, HET heterosexuals. Cluster definition was: genetic distance = 4.5% and bootstrap = 90%. Note that the proportion of individuals of each degree is shown rather than the absolute number of individuals of each degree. The number of clustered individuals was much larger in the UK than in Switzerland.

Because the KS test is sensitive to differences in scale between networks, we then applied the DDQC. The DDQC measures the distance between networks based on features extracted from their degree distribution. This metric has been devised deliberately to be robust to differences in scale between networks. However, it does

not indicate whether distances calculated are significant or not. In order to generate a null distribution for the DDQC, the UK and Swiss degree distributions were compared to themselves through bootstrapping. The UK and Swiss degree distributions were each bootstrapped 100 times (see Methods) and the DDQC distance calculated between the true data and each bootstrap replicate (Figure 4.3). Between country DDQC values were considered significant if they exceeded the 95% percentile of the within country DDQC values (one-sided test). The DDQC distance was higher than expected only for HET (95th percentile DDQC=0.97, observed HET DDQC=1.22).

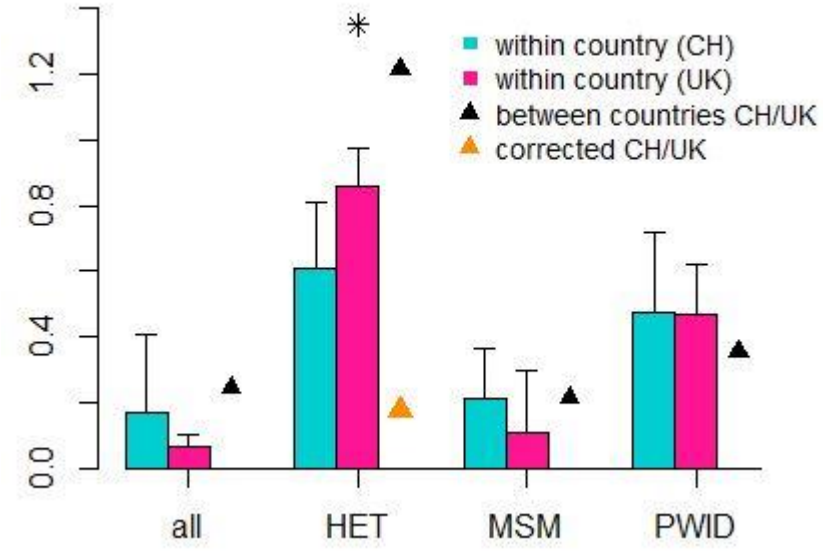


Figure 4.3: DDQC distances within and between countries. CH Switzerland, MSM men who have sex with men, HET heterosexuals. Cluster definition was: genetic distance =4.5% and bootstrap=90%. In order to generate null distributions of the expected values for the DDQC, we bootstrapped the Swiss and UK distributions and calculated DDQC values comparing the original datasets to their bootstrap samples (in blue and pink). The top of the coloured bars represent the mean distance of within country comparisons and the whiskers represent the 95% percentiles. The DDQC distance was then calculated between the UK and Swiss degree distributions (black triangles). The distance between countries was considered significant if it exceeded the 95% percentile from the

simulated values, which was the case only for HET at 4.5% (indicated by *). When we removed heterosexuals who were likely to have been infected through sex with men from the UK degree distributions, the DDQC distance between the UK and Swiss HET degree distribution fell within the simulated null distribution (orange triangle).

We hypothesised that the difference highlighted by the KS test at 4.5% might be the result of a difference in scale between the two epidemics. Although HIV prevalence is lower overall in the UK, the UK population (and the HIV+ population) is much larger than that of Switzerland and so the pool of partners available within the population is bigger. Thus the UK subtype B dataset (n=31450) used here was 3.75 times larger than the Swiss dataset (n=8390), the UK MSM dataset was 5.7 times larger and the UK heterosexual dataset was 1.3 times larger (Table 4.1). To examine the effect of epidemic size on clustering and degree distribution, we down-sampled the UK subtype B datasets to match the size of the Swiss datasets. In parallel, the Swiss dataset was bootstrap sampled with replacement. This was repeated 100 times for the population as a whole and for MSM and HET. In each replicate, we re-estimated the degree distribution based on the links remaining in the sample. When these equal-sized resampled datasets were compared, the UK and Swiss degree distributions overlapped for the population as a whole and for the MSM population, but not for HET (Figure 4).

In the true and the jackknife sampled UK HET population, individuals with high degree (>20) were present, who were not observed in the Swiss data (Figure 4.4). The largest exclusively heterosexual risk group in the UK comprised 27 individuals (bootstrap=0.9, GD=4.5%), all HETs with higher degree were in clusters that were dominated by MSM. When we dropped HETs with degree >26 from the UK HET distribution, the DDQC distance between the two networks was below the 95% percentile of the within country DDQC values (DDQC = 0.17, Figure 4.3).

4.4.4 Cross border transmission

We investigated intermingling between national and foreign sequences by removing the 80% national criterion. We used a tight genetic distance threshold (1.5%, 70% bootstrap) to be more likely to capture close transmission partners. At this threshold, Swiss sequences clustered with 162 non-Swiss sequences and UK sequences clustered with 353 non-UK sequences. For Switzerland, Western European countries provided over 75% of the links. For the UK, 50% of close links were with other European countries and 20% originated from other Anglophone countries: Australia, Canada and the USA (Figure 4.5).

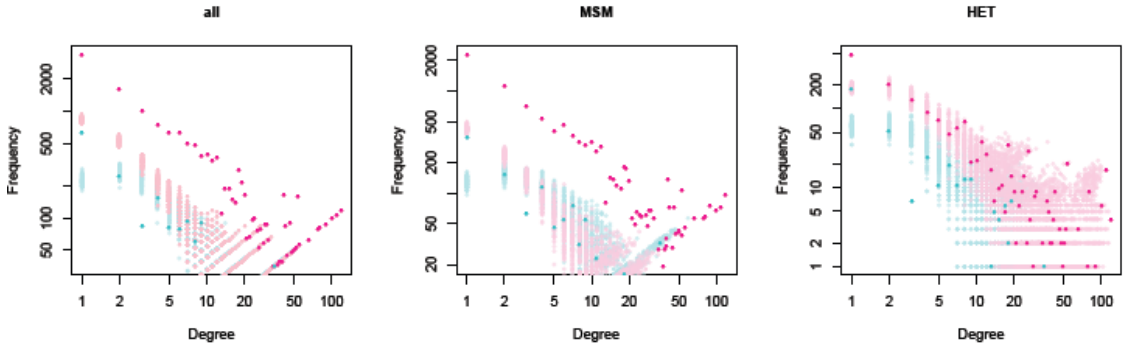


Figure 4.4: Jack-knife and bootstrap sampled degree distributions of the UK (pink) and Swiss (blue) epidemics. UK subtype B degree distributions for men who have sex with men (MSM), heterosexuals (HET) and the population as whole (all) were jack-knife sampled 100 times to match the size of the Swiss epidemics (in light pink). The Swiss epidemic was bootstrapped 100 times to its full size (in light blue). Degree distributions are shown on a double-logged scale. Samples overlapped for MSM and the dataset as a whole, but not for HET. Where Swiss replicates cannot be seen they are covered by the UK replicates.

4.5 Discussion

The aim of this study was to compare clustering patterns between the two most densely sampled HIV epidemics, the UK and Switzerland, while adhering to data governance procedures. Because a statistical framework for comparing cluster distributions is lacking, we generated degree distributions based on cluster sizes and composition and

compared them through formal statistical tests: the KS test, the DDQC and bootstrapping. Based on the KS test, there were differences between the UK and Swiss subtype B degree distributions at 4.5%. However, downsampling the UK dataset to the size of the Swiss dataset rectified this difference in all risk groups except for heterosexuals. In parallel, only heterosexuals showed a significant difference based on the DDQC test, which corrects for network size.

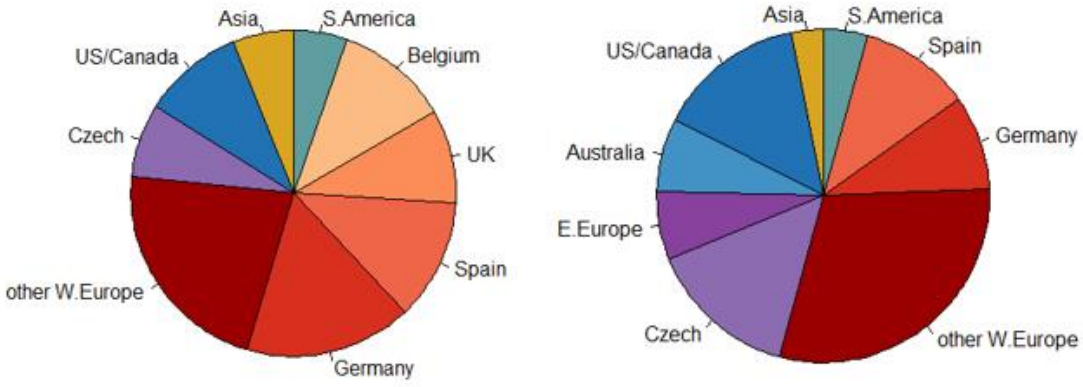


Figure 4.5: Origin of close linkages for Switzerland (left) and the UK (right). Swiss sequences clustered closely (1.5% genetic distance, 70% bootstrap) with 162 non-Swiss sequences and UK sequences clustered with 353.

The degree distribution of UK heterosexuals had a long tail representing male heterosexuals clustered exclusively with MSM. Previous analyses of the UK HIV RDB have demonstrated that a proportion of self-reported male heterosexuals are likely to have been infected through sex with men [106] and it is likely to be the case for the heterosexuals identified here. This result highlights the limitations of our approach in which self-reported heterosexuals are identified as heterosexuals even though they cluster only with MSM. Another way of classifying patients could be to do so based on the characteristics of the other individuals they cluster with [199]. In this example, these self-reported heterosexuals would instead be classified as MSM. When those high degree heterosexuals were removed from the UK HET dataset, the UK and Swiss degree distributions no longer differed. Male heterosexuals who have sex with men are likely to also have sex with women and provide a bridge between

MSM and heterosexual epidemics. This is a likely route for the spread of non-B subtypes among MSM in the UK (Chapter 1). More detailed analyses of the Swiss epidemic have found little overlap between the MSM and HET epidemics [105]; however, the Swiss heterosexual with the highest degree was similarly part of a HET/MSM cluster comprising 36 individuals, while the largest exclusively heterosexual cluster contained only 9. In fact, 47% of UK and 38% of Swiss heterosexuals were in HET/MSM clusters.

A more obvious difference was that 23% of Swiss heterosexuals clustered with PWID, as opposed to 12% of UK heterosexuals. However, PWID degree distributions showed no difference between countries, although this could be due to sample size. More importantly, the stemming of the heterosexual epidemic through PWID in Switzerland is likely to be an old process [105] while the potential bridging between HET and MSM in the UK is likely to be ongoing [106] (Chapter 1).

All our findings were consistent across bootstrap thresholds, and so bootstrap cut-off is not as useful in delineating clusters as genetic distance. At 1.5%, we found no difference between the degree distributions of the Swiss and UK epidemics. At such a tight threshold, mostly pairs and recently infected patients are likely to be captured and these groupings are similar across the two countries. At 1.5%, the UK epidemic initially appeared more highly clustered than the Swiss epidemic, but when sample date was taken into account, the Swiss epidemic became more clustered than the UK epidemic, indicating that close partners were more likely to be captured by the SHCS. The time between a new diagnosis and that patient's sequence being included in the database is shorter in the SCHS, and it is possible that Swiss coverage is slightly higher.

At 4.5%, the UK was more clustered and so the UK HIV RDB is more likely to capture larger transmission chains. However, the downsampled UK epidemic degree distributions overlapped with the Swiss degree distributions. While this does not change the fact that the proportion of individuals in clusters in the UK is higher, it means that the difference is seemingly due to its greater epidemic size rather than because of differences in the contact or transmission processes between the two

countries. The UK HIV epidemic is not particular insular but because the UK HIV RDB is so large the impact of foreign sequences in phylogenies is less dramatic. Both countries are integrated into global unsampled epidemics, and this study underlines the importance of HIV public health interventions at the European and global scales.

Transmission between European countries has been analysed in more depth elsewhere [98]. In agreement with that analysis, we found Spain to be a major mixing partner for both Switzerland and the UK. Germany and the Czech Republic were also identified as important. In their phylogeographic analysis, Paraskevis *et al.* noted significant migration of HIV from the UK to Germany [98] and this may be what we are capturing here. We did not conduct a phylogeographic analysis to determine the direction of these transmission events, but our use of a tight genetic distance threshold increases the likelihood that links captured are direct. We also found increased linkage between the UK and other Anglophone countries. In Switzerland strong segregation has been observed between German and French-speaking regions [105] and this language-dependency of HIV transmission warrants investigation at the global scale.

Both countries have noted the subtype diversification of their respective epidemics [93, 200], yet the difference in size between the UK and Swiss subtype A1 and C datasets (18,000 vs 900, respectively) meant that a detailed analysis was not worthwhile. In Switzerland the most recent analysis concluded that fewer than 25% of non-B infections were acquired in the country [97] whereas in the UK over 50% of infections in individuals born abroad are thought to have occurred in the UK [95]. Local non-B non-heterosexual transmission appears far more extensive in the UK (Chapter 5).

Although degree distributions are a blunt tool for elucidating the dynamics of an epidemic [123], they allowed us to apply statistically robust methods (which are not affected by our degree distribution definition) to compare the two epidemics without the need for exchanging sensitive data. The data being bootstrapped are not fully independent because a) sequences are sampled from clusters and cluster membership of sequences depends on other sequences in the cluster and b) sequences in the tree are all related to each other through the phylogeny. However, the aim of bootstrapping

and jackknifing tests was to determine whether sampling and population size might explain the differences in degree distributions observed between the two countries. The non-independence of the data would cause bootstrap and jackknife samples to look more similar to each other and to the original data, meaning that the difference between the UK and Swiss degree distributions would not be minimised. Thus the finding that the difference was no longer significant after jackknife sampling and this finding should not be affected by the non-independence of the data. A second issue, the distributions of sample dates differing between the two cohorts, arose because of extensive retrospective sequencing by the SHCS of patients diagnosed early on in the epidemic and for whom samples had been stored. The SHCS bias towards older samples explains in part the lower clustering observed in Switzerland. However, clustering remained significantly higher in the UK at 4.5% after sample date was adjusted for. While our aim was that the analysis conducted be as similar as possible across the two countries, subtypes were assigned using REGA in Switzerland and SCUEAL in the UK. The impact of this should be limited as the analysis focused on pure subtypes and mostly on subtype B. Next, our analysis is limited by the fact that we were unable to directly use each other's data and analysed only meta-data generated from each other's cluster analyses. We were not able to build trees containing sequences from both the SHCS and the UK HIV RDB, but the epidemics are likely to be highly intertwined. This can be concluded from the cross-border analysis as the UK was an important mixing partner for Swiss sequences. The reverse was not seen (i.e. Switzerland appearing as a major mixing partner for the UK), likely to be due to the ratio of LANL UK sequences available to SHCS sequences in comparison to the ratio of Swiss LANL sequences to UK HIV RDB sequences. A future study might be facilitated by the same person conducting both sets of analyses. This would allow them to get a better feel for each set of data, even if data were not amalgamated. Finally, international linkages are biased in favour of countries that have deposited sequences in public databases. We suggest the apparently important contribution of the Czech Republic to both epidemics may arise from recent submission of large numbers of sequences from that country.

In conclusion, by using a highly coordinated approach we showed that apparent major differences in clustering patterns between the UK and Switzerland subtype B epidemics can be explained by differences in size. This is the first study leveraging the vast amounts of data available in more than one national HIV database. We have shown how to make use of such data without breaching data governance procedures. Current transmission of subtype B in the UK and Switzerland is likely to be driven by similar underlying factors.

5 TRANSMISSION OF NON-B HIV SUBTYPES IS DRIVEN BY LARGE NON-HETEROSEXUAL CLUSTERS

5.1 Abstract

In the UK HIV subtypes continue to be characterised by risk group composition. MSM represent over 85% of subtype B infections, while over 85% of non-B subtype infections occur among self-identified heterosexuals. However, between 2002 and 2010 the proportion of non-B diagnoses among MSM increased from 5.4% to 17%, which may indicate that the transmission dynamics of non-B subtypes are changing.

Over 14,000 subtype A1, C, D and G sequences from the UK HIV RDB were analysed. Transmission clusters were defined by a maximum genetic distance of 4.5% and 90% bootstrap support, and I investigated relative patterns of growth by risk group between 2007 and 2009.

Of 1148 clusters of these 4 subtypes which contained at least two sequences in 2007, >75% were pairs and >90% were heterosexual. Most clusters (71.4%) did not grow during the study period, and pairs were significantly less likely to grow than larger clusters (75.5%, $p<0.0001$). In comparison with simulated trees for the same time period, cluster growth was lower than expected for small clusters, and higher than expected for clusters larger than 6. Growth was higher than expected for clusters comprising sequences from MSM and PWID and lower than expected for all other risk groups. The number of new diagnoses in clusters containing both heterosexuals and MSM was much higher than expected for subtype C. Both risk group ($p<0.0001$) and original cluster size ($p<0.0001$) were predictive of cluster growth in a generalized linear model.

These results show that despite the increase in non-B subtypes historically associated with heterosexual transmission, MSM and PWID continue to be most at risk of ongoing transmission. Crossover of subtype C from heterosexuals to MSM has led to the expansion of this subtype within the UK.

5.2 Introduction: non-B subtype transmission in the UK

The global HIV-1 epidemic is characterised by extremely high genetic diversity: nine subtypes (A-D, F-H, J and K) circulate along with numerous recombinant forms. While globally subtype C predominates, accounting for 50% of infections [11], non-B subtypes were rare in the USA and Europe until recently. In the UK from the 1980s to around 1995, the epidemic was dominated by the transmission of subtype B among MSM [107]. Subtype B remains individually the most prevalent subtype in Europe (>80% of infections) [201] and in the UK (~40% of diagnoses) [93]. However, non-B subtypes increased in prevalence in the UK from <25% of diagnoses in the early 1990s [202] to 60% in 2010 [93] (Figure 1.3). Rises have been seen in other European countries [97, 203, 204]. In the USA, only 3% of samples sent to the national laboratory between 2004 and 2010 were non-B subtypes, but the portion increased from 0% to 4% during that period [205].

Generalised HIV epidemics (HIV prevalence > 1% in the general population [206]) are characterised by non-B subtypes transmitted among heterosexuals [201, 207]. The increase in non-B subtypes in Western countries between 1990 and 2003 corresponds to the rise in new HIV diagnoses among heterosexuals born abroad. However, the number of new HIV diagnoses among heterosexuals born abroad decreased from 4426 in 2005 to 2,688 in 2013 due to changing patterns of migration [6]. In 2010, heterosexually acquired infections represented around half of all UK diagnoses [6] and 85% were classed as non-B subtypes [93]. This association between subtype and risk group has led to the frequent use of subtype as a proxy for transmission route in phylogenetic analyses of HIV [108, 109]. However, by 2007, the prevalence of non-B subtype HIV infections among UK-born MSM was 5.4% [208] and by 2010, this proportion had risen to 17% (415 infections) [93].

Several studies have set out to measure the relative contribution of migration and domestic HIV transmission to the increase in non-B subtype prevalence. In the UK, significant clustering of subtypes A1 and C has been observed [109] including among MSM [209] indicating some local transmission. However, clustering alone doesn't reveal the proportion of non-B subtype infections being acquired within the country. In Switzerland, von Wyl *et al.* (2011) estimated the proportion of new infections arising within Swiss-specific non-B subtype clusters (subtrees including ≥80% Swiss sequences) to be fewer than 25% [97]. Brand *et al.* (2014) examined clustering of sequences sampled from recently acquired infections (based on the results of the BED enzyme immunoassay [210]) concluding that at least 20% of non-B subtype infections had been acquired in France. The authors noted that while most non-B subtype diagnoses in France are made among heterosexuals, MSM were involved in the majority of clusters (defined as >90% bootstrap and <1.5% average genetic distance) [100].

In the UK, Aggarwal *et al.* investigated medical record and laboratory diagnostics for a small cohort to determine for each patient whether they had been infected before or after arrival into the UK [94] but such an intensive approach is infeasible at the national level. Rice *et al.* explored the origin of infections among heterosexuals born abroad by

calculating their estimated year of infection based on a detailed analysis of CD4+ decline [95]. In combination with their date of entry in the country, this was used to determine whether each patient was more likely to have been infected before or after arrival into the UK. These authors estimated that 73% of HIV+ heterosexuals born abroad diagnosed in 2002 were infected outside the UK , but by 2011 over 50% had acquired HIV within the UK [95]. This analysis was performed at the national level; however, it does not discriminate the sub-epidemics associated with different subtypes, and neither Aggarwal [94] nor Rice [95] can distinguish infections acquired through travel following domicile in the UK.

In order to analyse the recent transmission dynamics among non-B HIV subtypes in the UK, I used tools recently developed by our group (Chapter 0) [197] which facilitate the analysis of large pathogen sequence datasets and link them to epidemiological data. I applied these approaches to the analysis of sequences from the UK HIV RDB. As in previous work international databases were used to eliminate clustering that does not reflect UK-based transmission [109]. Taking this approach I have quantified the transmission dynamics of non-B subtypes across risk groups in the UK.

5.3 Methods

5.3.1 Data

43,002 partial HIV *pol* sequences were obtained from the UK HIV RDB 2010 download (see section 2.1.2). Subtypes C (10872, 25.3%), A1 (2083, 4.9%), G (965, 2.2%) and D (815, 1.9%) were the most common after subtype B and are analysed here.

5.3.2 HIV cluster dynamics

5.3.2.1 Cluster identification

The CP and CM (Chapter 0) [197] were used to analyse transmission dynamics. Phylogenetic trees were built in RaxML [162] for each subtype separately against a background of global sequences from LANL (see section 2.1.3). In all four trees, clusters were picked for analysis if they contained at least 2 sequences, 80% or more

of sequences were from the UK, bootstrap support exceeded 90% and maximum pairwise genetic distance was below 4.5% [108].

5.3.2.2 Cluster dynamics

Clusters were sorted into risk groups based on the self-identified risk group information associated with each sequence in the cluster. Proportions were calculated based on all sequences in a cluster, including those for which risk group was unavailable. I used the following hierarchical rules to classify clusters:

- If >25% of the sequences in a cluster came from PWID, the cluster was classified as PWID;
- If both HET and MSM each accounted for >10% of sequences, the cluster was classified as "crossover";
- If MSM accounted for >50% of sequences the cluster was classified as MSM, and;
- If HET accounted for >50% of sequences, the cluster was classed as HET.
- Otherwise, the cluster was classified as unknown (not available, NA).

Two other sets of rules tested the extremes of this risk group classification. According to the majority risk group definition, the risk group of the cluster was that of the majority of the sequences in the cluster. If two risk groups each accounted for 50% of sequences, both risk groups were used and growth was divided proportionally between them (or attributed to the crossover risk group in the case of HET-MSM clusters). According to the minority cluster definition, the risk group of any sequence in the cluster entered the risk group classification and growth rate was split between risk groups as for the majority definition. In both cases, clusters containing 50% or more sequences with unknown risk group were classified as NA. For clusters that contained at least one UK sequence collected up to December 2007, I counted the number of UK sequences in the cluster collected prior to December 2007 (old sequences) and the number of UK sequences added to the cluster after 2007 (new sequences). Cluster growth was calculated as the number of "new" sequences divided by the number of "old" sequences, and expressed as a percentage increment based on initial cluster size [211]. Cluster growth between clusters of different starting sizes and of different risk groups were compared using the Kruskal-Wallis test.

5.3.3 Simulations

Expected distributions for cluster growth were simulated according to a model in which all tips in the tree are equally likely to transmit. The distribution of bootstrap support values and branch lengths were obtained from the full trees (excluding global background sequences), following which sequences collected after 2007 were stripped. Removed sequences were then added to a random branch in the tree to simulate new infections. The probability of any branch being selected was proportional to its length. The branch length and bootstrap support for the new bifurcation were drawn at random from the branch length and bootstrap distributions of the full tree containing all sequences up to end of 2009 (see Appendix 3). For pseudocode, see Appendix 6. Sequences were simulated along the resulting phylogenies under a GTR substitution model with nucleotide frequencies from the original data using SeqGen [212], ensuring that mean genetic distance among simulated sequences was equivalent to that of the true sequences. This was repeated 1000 times for each tree. Clusters ($\geq$90% bootstrap and $\leq$4.5% genetic distance) were picked in the simulated trees, and cluster growth calculated as in the original trees, in order to generate expected cluster growth distributions for each subtype. Cluster growth was compared between clusters of different starting sizes and different risk groups.

In parallel, we conducted a permutation test, permuting the risk groups across the tips of the trees. Each time risk group was re-estimated according to the rules in section 5.3.2.2, and we compared cluster growth according to risk group.

5.3.4 Generalised linear model

A generalised linear model (GLM) was constructed to express the number of new sequences as a variable dependent on the number of old sequences, risk group and subtype. The distribution of the dependent variable was tested against Poisson, negative binomial and gamma distributions to verify that a GLM would be appropriate. The distribution of the number of new sequences was adequately fitted by a negative binomial distribution ($X2$ test, p=0.1) with no other model offering a better fit, and so the Negative Binomial GLM available in R [154] was used.

5.4 Results

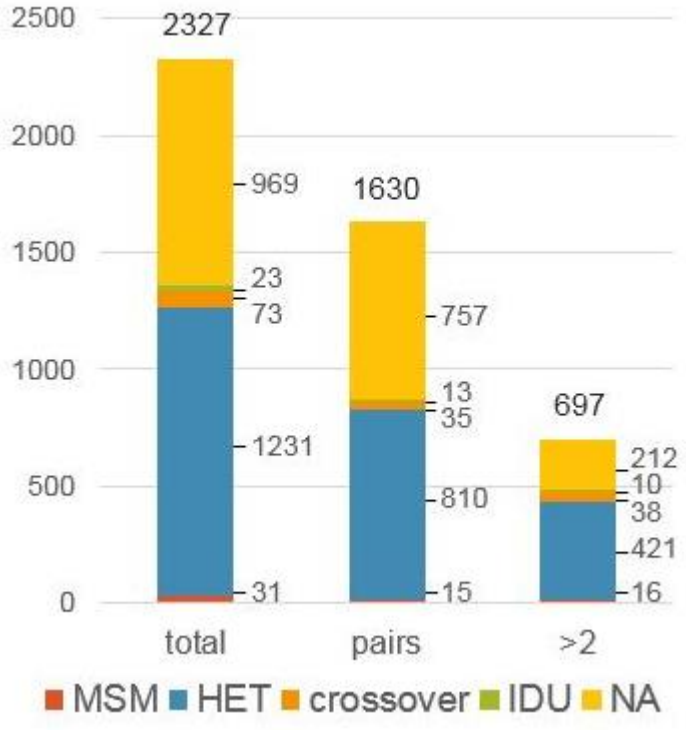


Figure 5.1: Risk group classification of clusters (2009). 2327 clusters (304 A1, 1785 C, 113 D, 125 G) contained at least two sequences: 1630 clusters were pairs and 681 comprised more than two sequences. MSM men who have sex with men, HET heterosexual, PWID people who inject drugs, NA not available.

I investigated the dynamics of transmission clusters of non-B subtype HIV sequences between January 2007 and December 2009 to determine the drivers of new diagnoses during this time period. After exclusion of any duplicates, the 2010 UK HIV RDB subtype A1 dataset contained 2083 sequences, of which 630 were collected after December 2006 and the subtype C dataset contained 10830 sequences, including 4852 collected after 2006. Of 815 subtype D sequences, 279 were collected after 2006; and of 965 subtype G sequences, 472 were collected after 2006. Thus while the UK HIV RDB subtype A1 and D datasets grew by ~50% during the study period, the subtype C and G datasets grew by over 80%. To these were added over 7000 non-UK sequences selected by Viroblast to distinguish between UK and non-UK transmission.

Phylogenetic trees were reconstructed separately for the four subtypes and clusters were identified.

A total of 2327 clusters (304 A1, 1827 C, 112 D and 125 G) meeting the above criteria were identified in the four trees. 101 clusters were excluded because they comprised <80% UK sequences, of which only 26 contained more than one UK sequence. 5999 UK sequences (41%) clustered with at least one other UK sequence used in the phylogenetic analysis. Over half of linked sequences were linked to only one other ("pairs"; 3260 of 5999 sequences and 1630 of 2327 clusters). Of 2327 clusters, 969 (42%) could not be assigned a risk group. Of the clusters that could be classified, the great majority were HET (1231/1358; 91%). A further 31 clusters were classed as MSM (2.3%), 23 as PWID (1.7%) and 73 as crossover clusters containing sequences from MSM and HET (5.4%; Figure 5.1).

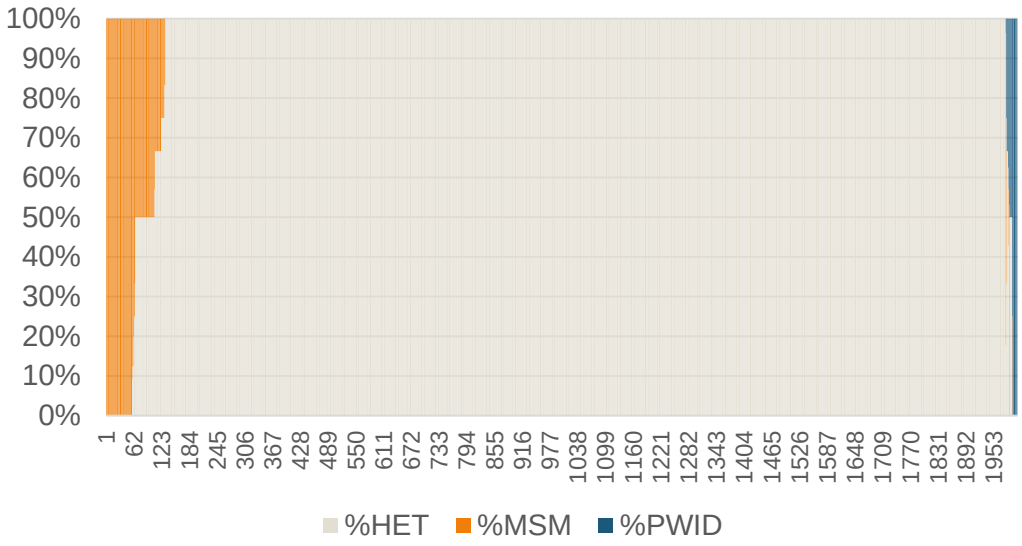


Figure 5.2: Risk group composition of each cluster. Clusters were sorted by proportion of heterosexuals (HET), men who have sex with men (MSM) and people who inject drugs (PWID). Sequences which did not have risk group information are not shown and clusters which were not assigned a risk group are not shown.

Risk group cluster classifications were mostly "clean" (Figure 5.2) with 82.4% of clusters containing a single risk group. Only four clusters out of 2327 contained a mix

of heterosexuals, MSM and PWID. Crossover clusters were clear mixes of MSM and heterosexual sequences.

5.4.1 Cluster growth depends on initial cluster size

Of the 1148 clusters that existed (pairs or larger clusters) in January 2007, only a minority changed during the study period (328/1148, 28.6%; Table 5.1). This effect was more pronounced in subtypes A1, D and G where fewer than 20% of clusters changed, compared to 33.5% in C (Table 5.1, Fisher's exact test $p<0.0001$). Pairs were significantly less likely to grow than larger clusters ($p<0.0001$). Among pairs only 215/879 (24.5%) changed during the study period. Overall, of 6233 new non-B subtype sequences added to the database after January 2007, 1457 (23.4%) linked to sequences already in the database before that date. This proportion was higher among subtypes A1 (24.6%) and C (25.0%) than among D (11.1%) and G (12.7%; $p<0.0001$).

Table 5.1: Proportion of clusters showing growth between 2007 and 2009.

Subtype		Total	Growth
A1, C, D, G	all clusters*	1148	328 (28.6%)
A1, C, D, G	Pairs	879 (76.6%)[#]	215 (24.5%)
A1	all clusters*	190	37 (19.5%)
A1	Pairs	147 (77.4%)[#]	24 (16.3%)
C	all clusters*	792	265 (33.5%)
C	Pairs	614 (77.5%)[#]	175 (28.5%)
D	all clusters*	87	12 (13.8%)
D	Pairs	63 (72.4%)[#]	8 (12.7%)
G	all clusters*	79	14 (17.7%)
G	Pairs	55 (69.6%)[#]	832 (14.5%)

Notes: *containing at least one sequence collected before December 2006; [#] pairs as a percentage of all clusters

For clusters containing at least one sequence collected prior to January 2007, growth was calculated as the number of sequences collected after this date divided by the number of sequences already clustered. In the UK data, growth varied

significantly between clusters of different starting sizes (Kruskal-Wallis test, df=4, p<0.0001).

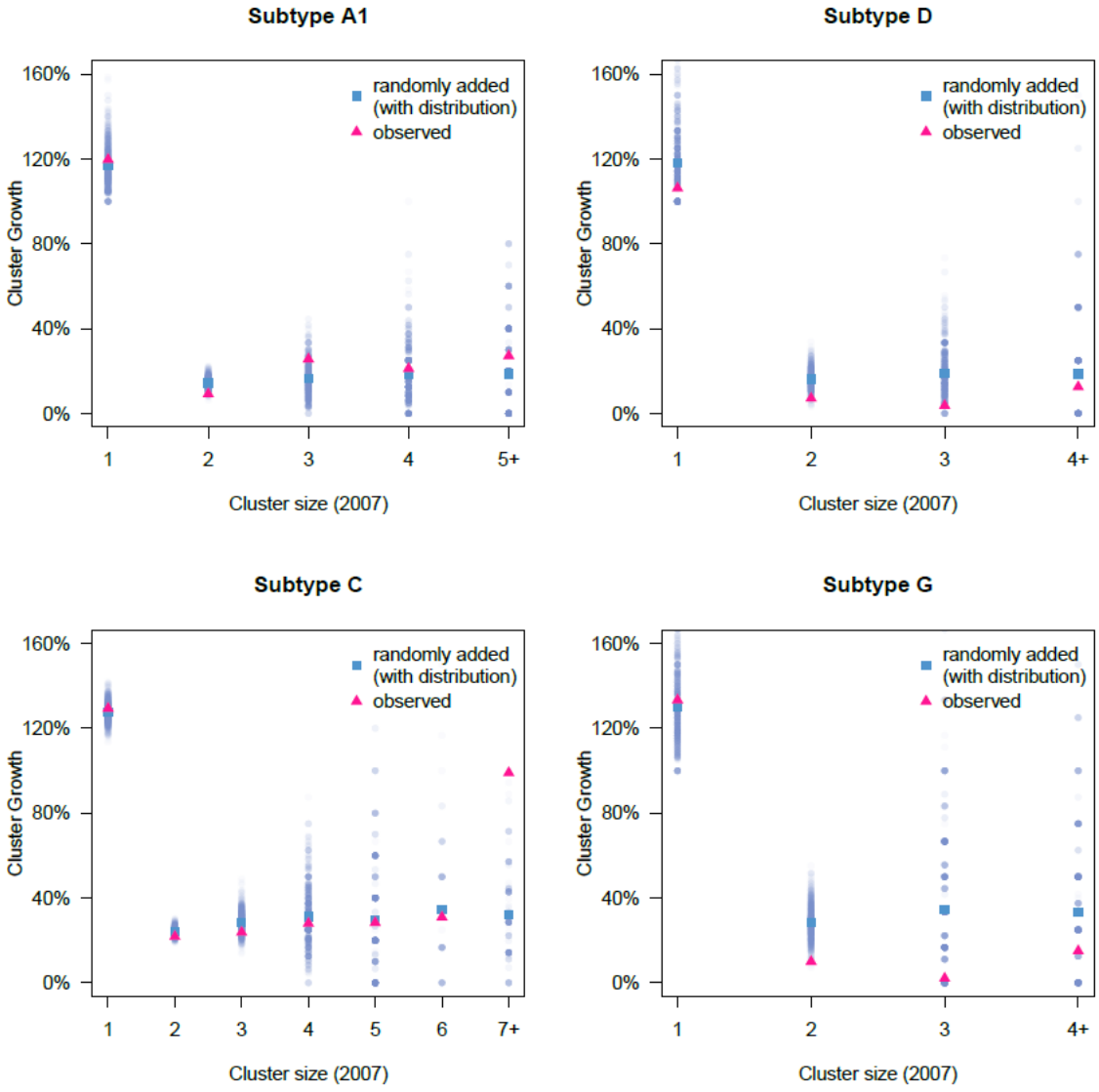


Figure 5.3: Cluster growth according to initial cluster size (2007). For each cluster in 2009 containing at least one sequence collected before 2007, cluster growth was calculated as the number of new sequences over the number of old sequences (in pink). In parallel, expected cluster growth was simulated according to initial cluster size (in blue, 1000 simulations, mean and distribution shown).

Cluster growth increasing with cluster size in 2007 for subtypes A1 and C (Figure 5.3, in pink), and this trend was significant only for subtype C (r=0.95, p<0.05). Single

sequences collected prior to 2007 that formed clusters after 2007 (670 in total) were not included.

A simulation-based approach was adopted to test whether observed cluster growth differed significantly from that expected if sequences were added to trees at random. If all individuals in a population are equally likely to transmit, the probability of a new infection linking to any specific infection from the original population is equal to one over the total size of the population. Although the entire population is not sampled, the full genetic diversity of the population is captured by the tree and a new infection unlinked to those previously sampled will fall on a branch distant to the tree's terminal nodes. The longer the length of a single branch, the more likely it is that individuals have been missed along that branch and that a newly identified infection could occur along that branch. The probability of a new infection being linked to any given cluster is proportional to the size of that cluster.

Table 5.2: Cluster growth by cluster size (2007), mean observed and expected.

Subtype	Initial cluster size (2007)	Number of clusters	Observed growth	Expected growth (2.5% quantile)	Expected growth (97.5% quantile)
A1	2	147	0.094	0.097	0.196
C	7+	8	0.990	0.000	0.857
D	2	63	0.071	0.075	0.256
G	2	55	0.100	0.145	0.439

Notes: Clusters are only shown if their observed growth did not fall within the 95% expected growth quantiles ($p<0.05$). Subtype C clusters sized 7 and above exceeded their expected growth while subtype A, D and G clusters sized 2 grew less than expected.

Sequences collected after January 2007 were stripped from the tree and added back with the probability of attachment based on the length of each branch and on the genetic distance and bootstrap distributions from the original tree (see Methods 5.3.3). In each simulated tree, clusters were picked as described previously and average cluster growth was calculated for clusters of each starting size and each risk group.

In the simulated data, average expected growth values were normally distributed (Figure 5.3, in blue). For clusters of each size, I evaluated whether the observed value of cluster growth fell within the 95% quantile estimates of the simulations (Table 5.2). For subtypes A1, D and G, growth of pairs was below the 2.5% quantile ($p<0.05$) of growth in the simulations. In 2007, there were a small number (8/792) of larger clusters (≥ 7) in subtype C, and none in the other subtypes. For this group, growth was higher than the 97.5% quantile ($p<0.05$) of simulated clusters. All other values for cluster growth fell within the 95% quantiles of the simulations.

Table 5.3: Mean observed cluster growth and expected growth by risk group.

Subtype	Risk group	Number of clusters*	Observed growth	Expected growth (2.5% quantile)	Expected growth (97.5% quantile)
A1	MSM#	9	0.421	0	0.417
A1	PWID#	2	0.833	0	0.750
C	MSM#	10	0.640	0.050	0.550
C	PWID#	5	1.015	0	0.875
C	crossover#	36	0.550	0.073	0.500
D	NA	31	0.075	0.358	1
G	non-HET	6	0.056	0.130	0.462

Notes: *containing at least two sequences collected before January 2007. Clusters are only shown if their observed growth did not fall within the 95% expected growth quantiles ($p<0.05$). Clusters which grew more than expected are indicated by #.
MSM men who have sex with men; PWID people who inject drugs; NA not available; non-HET non-heterosexual.

5.4.2 Cluster growth is higher for non-heterosexual risk groups

Analysis by risk group was performed on clusters containing at least 2 sequences in 2007 (n=1148). Growth differed significantly between risk groups when all subtypes were analysed together (Kruskal-Wallis, $p<0.005$, df=4). Although MSM (mean growth=0.54), PWID (mean growth=0.64) and crossover (mean growth=0.44) clusters grew significantly more than HET clusters (mean growth=0.18), this was also observed in the simulated data ($p<0.0001$, df=4) due to the differing initial cluster size between risk groups ($p<0.0001$, df=4). For subtypes A and C, cluster growth was

higher in the observed data than in the simulations for PWID, MSM and crossover clusters (Table 5.3, Figure 5.4), indicating initial cluster size was not responsible on its own.

In the permutation test (Figure 5.5), growth followed the same patterns as in the simulation except for subtype A1 and C clusters where crossover clusters where growth was higher than in the observed data. This can be explained by the fact that the majority of MSM clusters became crossover clusters after the permutations because of the over-representation of heterosexuals among the new tips,

Changing the rules of cluster risk group classification changed the risk group of only 11/1148 clusters for the minority definition and 43/1148 clusters for the majority definition (of which 29 were crossover clusters which became either HET or MSM). Differences in growth rates between risk groups were unchanged other than crossover growth rate dropping in the majority definition because most crossovers clusters were relabelled as either MSM or HET (Figure 5.6).

5.4.3 Cluster size and risk group act independently on cluster growth

To determine whether the effect of cluster size remained significant when risk group was accounted for, I constructed a generalized linear model (GLM) expressing the number of new sequences as a variable dependent on the number of old sequences, risk group and subtype. Risk group and number of old sequences were added to the model sequentially, with risk group highly predictive ($p<0.0001$) and the addition of cluster size significantly improving the fit of the model (ANOVA likelihood ratio test, $p<0.0001$). The difference in effect between subtypes was highly significant ($p<0.0001$). There was no significant interaction between risk group and cluster size ($p=0.108$). In terms of effect, the model explained 24.75% of the variance in growth: risk group accounted for 12.31%, initial cluster size for 9.32% and subtype for 3.12%. Cluster growth was higher for crossover ($p<0.0001$), PWID ($p<0.05$) and MSM ($p<0.01$) clusters, for subtype C clusters ($p<0.01$) and increased with initial cluster size ($p<0.0001$).

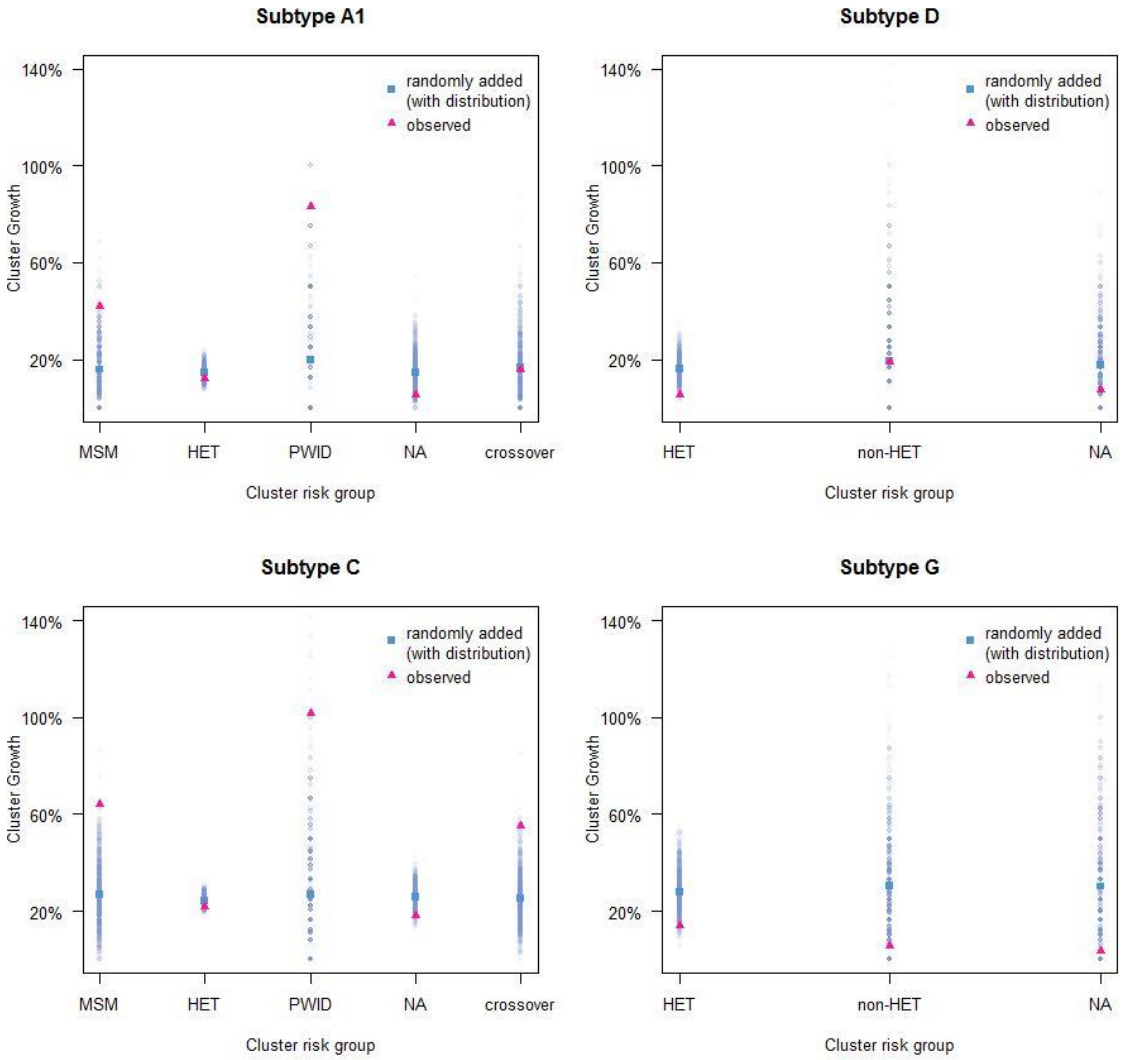


Figure 5.4: Cluster growth according to risk group (simulation). Observed cluster growth is shown in pink and mean expected growth and distributions from 1000 simulations are shown in blue. For subtypes D and G, numbers of non-heterosexual clusters were small and so these were amalgamated for analysis. MSM men who have sex with men; HET heterosexual; PWID people who inject drugs; NA not available; non-HET non-heterosexual.

Because the data contained a large number of clusters which did not increase in size at all (dependent variable=0), I also ran a zero-inflated model which estimated coefficients separately for growth and non-growth (pscl library in R [213]). Initial cluster size was included as an offset because we expect larger clusters to grow more.

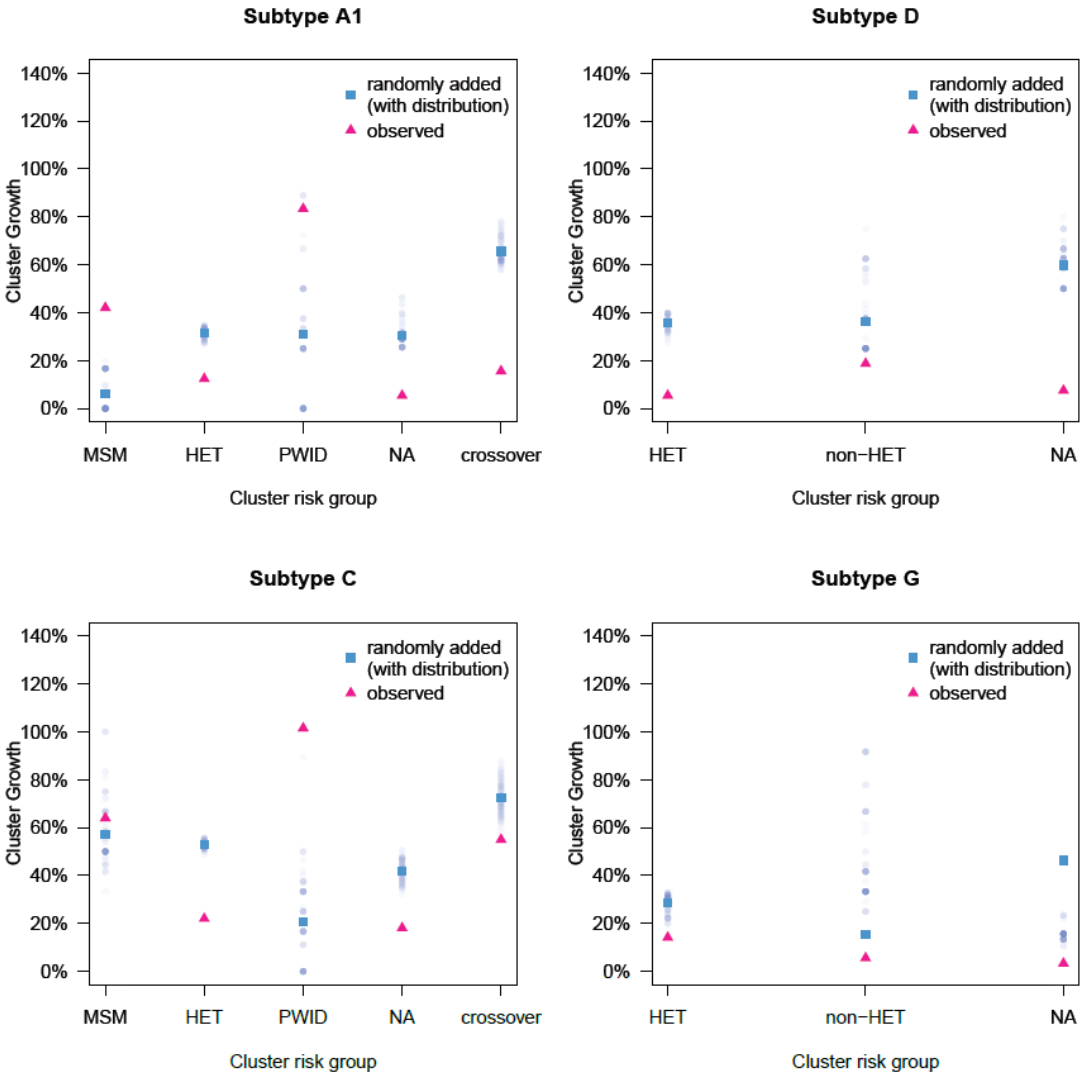


Figure 5.5: Cluster growth according to risk group (permutation). Observed cluster growth is shown in pink and mean expected growth and distributions from 100 permutations are shown in blue. For subtypes D and G, numbers of non-heterosexual clusters were small and so these were amalgamated for analysis. MSM men who have sex with men; HET heterosexual; PWID people who inject drugs; NA not available; non-HET non-heterosexual.

None of the zero-inflation model coefficients (risk group and subtype) were significant (all p>0.6). The effect of risk group and subtype followed the same trend as in the original glm, but the effects were not all significant. Heterosexual clusters grew significantly less than crossover clusters (p <0.05) and subtype C clusters grew more

than other subtypes (p <0.01), but the effect of PWID, MSM and subtypes D and G were not significant.

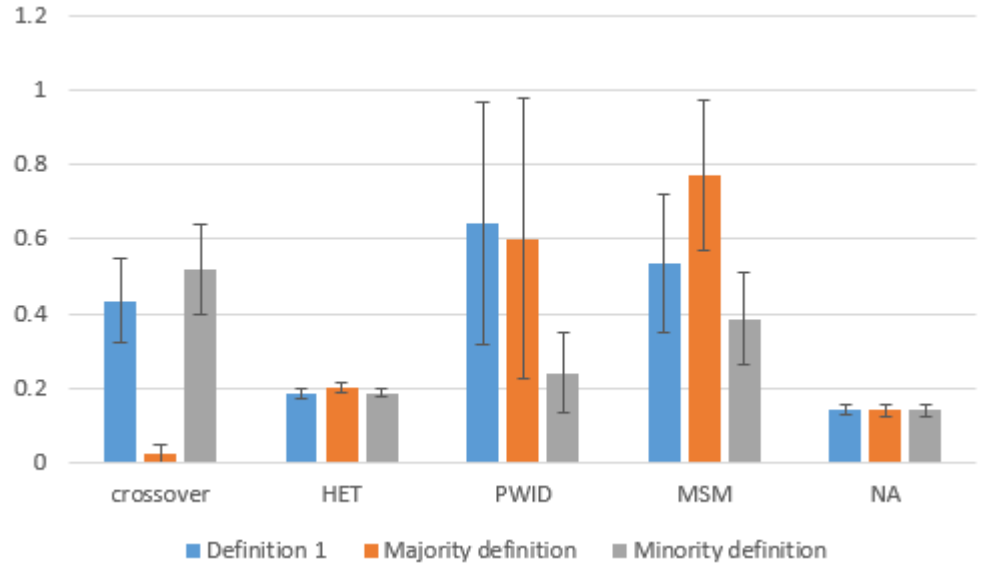


Figure 5.6: Growth rate according to risk group under three definitions. Two risk group classification procedures (minority and majority definition) were tested in addition to the classification used in the paper. Cluster growth rates are according to risk group did not change significantly as shown by the overlap in standard error bars, except for among crossover clusters as most of these disappeared based on the majority definition. MSM: men who have sex with men, HET: heterosexual, PWID: people who inject drugs; NA not available.

5.4.4 C is increasingly acquired in the UK while D and G are imported.

An increase in the ratio of clustering to non-clustering sequences in the database over time would indicate a rise in the proportion of local transmissions. When all subtypes were analysed together, this ratio did increase (Figure 5.7) but the change was not significant (Cochran-Armitage test across years [214, 215], p=0.5). Broken down by subtype, the subtype C clustering ratio rose from 0.69 to 0.75 (p=0.01), indicating an increasing proportion of infections acquired within the UK over time. In contrast, the clustering ratio decreased for D (from 0.41 to 0.33; p<0.01) and G (from 0.46 to 0.33; p<0.0001), signifying that most new sequences were unlinked to those already in the UK and are more likely to be the result of migration. However, overall numbers were

small, with only 279 subtype D and 965 subtype G diagnoses after 2006. The change was not significant for A1 (p=0.7).

For all clusters in 2009 containing at least one sequence in 2007, I counted the number of new sequences clustering in crossover and MSM clusters (Table 5.4). When I looked at the proportion of new diagnoses linking to crossover and MSM clusters (2-3% of new diagnoses) as opposed to HET clusters, proportions were not different across subtypes (Fisher's exact test, p=0.38), although numbers were small for subtypes D and G, with only 90 new diagnoses linking to subtype D and G clusters.

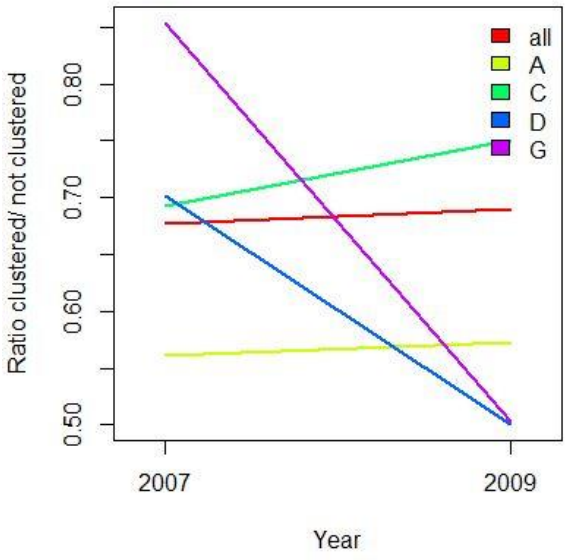


Figure 5.7: Change in clustering ratio between 2007 and 2009. The ratio of clustered to non-clustered sequences increased for subtype C (p=0.01), decreased for D (p<0.01) and G (p<0.0001) and did not change for A1 (p=0.7) or when all clusters were analysed together (p=0.5).

5.5 Discussion

The interpretation of cluster distributions and cluster change has proved difficult in the past because of the lack of a meaningful statistical framework to compare observed and null distributions. Here, I developed a novel approach to generate null patterns of cluster growth with which to compare observed growth. Sequences were added to the tree according to the branch length and bootstrap distributions of the observed

phylogenetic tree, but on branches chosen at random. This simulation method has the added benefit of not requiring all the infections to be sampled because the full genetic diversity of the population is still captured by the phylogenetic tree.

Table 5.4: Number of sequences added to clusters between 2007 and 2009

Initial cluster size	crossover	HET	PWID	MSM	NA	Grand Total
A1	7 (1.11%)	76 (12.1%)	7 (1.11%)	14 (2.22%)	51 (8.1%)	155 (24.6%)
1	2	41	2	2	44	91
2+	5	35	5	12	7	64
C	81 (1.67%)	620 (12.78%)	38 (0.78%)	40 (0.82%)	432 (8.9%)	1211 (24.96%)
1	15	362	5	7	320	709
2+	66	258	33	33	112	502
D	2 (0.72%)	14 (5.02%)	1 (0.36%)	1 (0.36%)	13 (4.66%)	31 (11.11%)
1	0	8	1	1	7	17
2+	2	6	0	0	6	14
G	9 (1.91%)	30 (6.36%)	3 (0.64%)	0 (0%)	18 (3.81%)	60 (12.71%)
1	9	14	2	0	15	40
2+	0	16	1	0	3	20
All subtypes	26 (1.59%)	740 (11.87%)	49 (0.79%)	55 (0.88%)	514 (8.25%)	1457 (23.38%)
1	26	425	10	10	386	857
2+	73	315	39	45	128	600

Notes: According to risk group (crossover; HET heterosexual, PWID people who inject drugs, MSM men who have sex with men, NA not available) and cluster size in 2007 (single sequence = 1, or cluster ≥ 2). Percentages are calculated as the number of sequences joining clusters of each risk group out of the total number of new sequences for each subtype (A1 = 630, C=4852, D=279, G=472).

At least 25% of non-B subtype infections diagnosed between 2007 and 2009 were linked to previously diagnosed infections and are likely to have been acquired within the UK. Among pre-existing clusters, cluster size and non-heterosexual risk group

each independently predicted higher cluster growth. The main conclusion of this analysis is that crossover to non-heterosexual risk groups has resulted in increased UK-based transmission of subtype C in particular, as has been observed in France [100]. However, the majority of clusters did not grow and there were important differences between the subtypes analysed.

Subtypes C, A1, D and G have established generalised heterosexual epidemics in sub-Saharan Africa [216] and are the most common subtypes in the UK after subtype B, where they also circulate predominantly among heterosexuals [109]. Around half of diagnoses in the UK every year are now non-B subtypes, with subtype C alone accounting for one third and A1, D and G on the rise [93]. This distinguishes the UK from the rest of Europe, where subtype B still accounts for >80% of infections [201]. Subtype AE was excluded because although similar in frequency to D (~2%), subtype AE shows a different pattern of spread as it is found equally in MSM and heterosexuals.

The most striking patterns observed here were the differences in cluster growth between risk groups. The overwhelming majority of clusters in this study were heterosexual, as would be expected. However, clusters classified as MSM or PWID were significantly more likely to link to new infections despite their numbers being small. Despite the success of needle exchange programs in the UK and the overall low incidence of HIV among PWID [217], these results confirm how rapidly HIV can spread in PWID and raise questions about the effectiveness of harm reduction efforts in these particular groups. I observed a number of crossover clusters containing sequences from both heterosexuals and MSM, which were also more likely to grow. In particular for subtype C, the number of newly clustering sequences in crossover clusters vastly exceeded that expected based on the simulations. Although the dynamics of D and G remain dominated by imports, similar patterns seem to be emerging. Sexual linkage between risk groups leads to the introduction of new subtypes. Crossover of subtype B from MSM into heterosexuals has been demonstrated in the UK [106], where Black-African men who clustered exclusively with MSM were more likely to self-identify as heterosexuals compared to other

ethnicities. It is likely that this same pattern is acting as a driver for spread of non-B subtypes among MSM. In order to reduce the spread of non-B subtypes among MSM, it would be important to encourage self-identified heterosexual men to disclose whether they are having sex with men (through publicity campaigns or during consultations) and if so educate them on HIV transmission. MSM could also be informed as to the increased prevalence of HIV among Black-African MSM who also have sex with women. In contrast, the epidemic growth of non-B subtypes among heterosexuals in the UK appears limited, in agreement with models suggesting that the prevalence of HIV among heterosexuals in the Netherlands was hardly affected by immigration [218].

It is important to note that most pre-existing clusters did not link to any new infections (>70%). Subtype A1, D and G pairs grew less than expected, with 85% not changing at all. The tendency of larger clusters to act as a driving force for epidemic growth has not been formally tested previously and is consistent with the hypothesis of preferential attachment [140], whereby already highly connected individuals tend to make proportionally more contacts over time. This was emphasised by the GLM analysis which showed that cluster size remained significant in the model even when risk group was taken into account.

Public Health England data indicate that the UK epidemic is growing, with 6000 new cases diagnosed every year [6], half of which are non-B subtypes [93]. These new diagnoses could be imports, new infections or new diagnoses of old infections. If they were all imports, none of their sequences would attach to pre-existing clusters. If they were new diagnoses but actually old infections, they could attach to pre-existing clusters without reflecting new infections within that cluster. Clusters would "grow" by the definition implemented in the present analysis, but the epidemic would be unchanging. It would be possible to detect a difference between this latter scenario and one of true growth (new infections linked to old clusters) by looking at the distribution of branch lengths of new diagnoses. New infections are likely to be connected by short branch lengths to existing clusters while old infections will be more distantly related (due to the greater opportunity for within host evolution). In time-resolved trees, new

infections would also be more likely to converge close to the tips of clusters while old infections would converge deeper into the cluster. We are currently performing time-resolved analyses of these data.

The relative frequencies of A1, C, D and G in the UK reflect their respective global prevalence [11], but I observed significant differences in their patterns of growth. Even though, proportionally, the UK subtype G epidemic grew as much as subtype C (over 80% during the study period), the majority of those subtype G infections did not link to previous UK infections. The ratio of clustering to non-clustering sequences for D and G decreased, so D and G appear not to be being transmitted as much as C within the UK. Meanwhile, subtype C cluster growth was much higher than expected. Several large, non-heterosexual clusters displayed cluster growth more similar to subtype B [108], and numerous crossing over events were observed. In parallel, I found that D and G have for the most part not crossed over into non-heterosexual risk groups. As such, I conclude that the spread of non-B subtypes appears to exhibit threshold behaviour, whereby rapid growth follows transmission into non-heterosexuals or groups of high-risk heterosexuals.

The figure of 25% is an underestimate of the proportion of non-B strains acquired within the UK. Only clusters that contained at least one sequence collected prior to 2007 were analysed because there is a better chance that subsequent infections will have been acquired in the UK. However some infections acquired within the UK will be missed, because of undiagnosed patients and because of the choice of clustering threshold. In the latter case, evolution in one or both patients can cause sequences to diverge such that either genetic distance will exceed the cut-off or bootstrap support will be too low [219]. It is difficult to compare these results to the published estimate of 50% heterosexuals born abroad having acquired their infection in the UK [95], because I am looking at subtype acquired rather than country of birth, and risk groups other than heterosexuals are included. However, non-UK born individuals are more likely to be infected with non-B subtypes [207] and MSM are more likely than heterosexuals to have acquired their infection in the UK [6] so a higher proportion of non-B subtype infections acquired in the UK would be expected. Lower rates of

diagnosis among heterosexuals (31% undiagnosed compared to 16% among MSM [6]) may play a part in explaining this result. Sequences from MSM are more likely to be captured by the UK HIV RDB. Nevertheless, as the UK HIV RDB is generated through population surveillance, not from a cohort, over time it will accumulate HIV positive patients from participating centres as they transition from undiagnosed into care, and linkages will increase.

There are limitations to the analysis. Importantly, the number of diagnoses in heterosexuals born abroad varies in parallel to patterns of migration from HIV endemic countries into the UK and both have recently decreased. Therefore the rise in non-heterosexual non-B subtype diagnoses could be interpreted as a relative rise and not an absolute one. However these cluster analyses focus on infections that link to previous diagnoses and cluster growth rates will not be impacted by changing patterns of migration. It is true that that the increase in the clustering ratio observed for subtype C could have occurred a result of decreased immigration. However, as this pattern was unique to subtype C and is concordant with the extensive spread of subtype C among MSM, these results nonetheless point toward subtype C being increasingly acquired within the UK.

The cut-off of January 2007 was selected to equalise the number of sequences available on either side. In time-resolved trees of the UK non-B epidemic, the distribution of branch lengths is known to be continuous indicating that sequences were added to clusters linearly with time [109] and so no particular cut-off would be expected to affect these findings. Similarly, although a bias could be introduced by the risk group classification of clusters, the density plots indicated that the majority of clusters comprised a single risk group and the sensitivity analysis demonstrated the robustness of the main conclusions.

Finally, simulations of tree growth are simplifications of the epidemiological process. Formal methods linking phylogenetic reconstructions and epidemiological models remain difficult to apply to large datasets. However, the use of simulations as a null model to compare data to has provided a formal test. Non-heterosexual cluster growth rates diverge dramatically from a null model in which all individuals in the population

are equally likely to transmit. In addition, modelling cluster growth based on the phylogenetic tree instead of using an epidemiological model has the advantage of implicitly accounting for new infections from unsampled sources without having to predetermine the population sampling coverage.

Overall, I found multiple subgroups within the UK non-B epidemic. The majority of clusters did not grow and those that did were more likely to already be larger and to contain non-heterosexuals. There were differences between the subtypes analysed, with subtype C infections increasingly acquired within the UK while the dynamics of subtypes D and G were dominated by imports. One possible explanation for the difference in observed cluster growth rates could be that infections in heterosexuals occur in sequential chains of transmissions [109] while among MSM, infections occur in rapid bursts [108] within a more densely connected contact network. In Chapter 1, I investigate differences in transmission network structure between risk groups.

I conclude that crossover into non-heterosexual risk groups has led to rapid non-B expansion within the UK recently. This study underlines the importance of continuing efforts to prevent HIV transmission among all risk groups despite the UK epidemic changing from being MSM and subtype B dominated to being more diverse in terms of risk group and subtypes.

6 CHARACTERISING UK HIV TRANSMISSION NETWORKS BY RISK GROUP

6.1 Introduction: degree distributions of networks and targeting interventions

The assumption of random mixing frequently made by standard epidemic theory is violated in the case of sexually transmitted infections (STIs) such as HIV. HIV can only be transmitted to a number of close contacts which is much smaller than the total population size. In addition, the number of sexual partnerships is far from being normally distributed. Instead, most individuals have few sexual contacts while a small number have many [220]. This heterogeneity in number of contacts is observed in many real life networks (neuronal networks, power grids, social networks [116]) and is thought to be central to the propagation of STIs including HIV. Without taking into account the variation in number of sexual partnerships in models, it is difficult to recreate observed HIV incidence [221].

The degree distribution of sexual partnerships of a population varying in number of contacts has a heavy tail and frequently follows a power law [128]. The variance of the degree distribution is determined by the exponent of that power law. Importantly, for certain values of the exponent, the variance of the distribution becomes infinite and the epidemic threshold disappears [110] which means that untargeted interventions will not stop the epidemic (see section 1.5.5). Although this theoretical result only holds true in infinite populations, the exponent of the power law can give us an indication of the extent to which interventions should targeted. Appropriate characterisation of the degree distribution can be used to assess whether a network adheres to a particular probability model [127], providing information on the process that led to its formation.

Studies of sexual networks have found contradicting results. Some studies have found power law distributions with infinite variance to fit Swedish [128], British, US and Zimbabwean sexual contact networks [129], but reanalysis of these same data have shown otherwise [123, 124, 126]. Importantly, some of the best fit models make HIV predictions which are inconsistent with the observed epidemic patterns [123], underlining our lack of understanding of the contact networks themselves and their effect on the spread of disease.

In reality, it is very difficult to obtain reliable contact network information (see section 1.5.6). It is impractical and expensive for large populations and requires personal information which in the case of sexual contacts (and in view of the criminalisation of HIV transmission see reference [145] and section 7.4), participants may not readily volunteer. In the case of HIV, the time lapse between infection and diagnosis also leads to problems of recall, and patients often do not know who infected them [139].

Recent years have seen the development of a new method for reconstructing HIV transmission networks: inference from viral sequences [140] (see section 2.2.4.2). Sequencing HIV for detecting drug resistance has led to the accumulation of 100,000s of sequences in databases globally, and so data are readily available (see section 2.1.1). In addition, networks reconstructed from sequences have the advantage of including only contacts that are relevant for understanding HIV transmission.

In their 2011 analysis, Leigh Brown *et al.* found that the Waring distribution offered the best fit to UK subtype B transmission networks constructed from viral sequences [140]. Waring is a power law distribution that results from a preferential attachment model.

In this chapter, I further analyse the UK degree distributions by subtype (subtypes A1, B and C) and risk group (heterosexuals, MSM and PWID) and by sex. This distinction is important because previous studies have found different distributions to fit male versus female sexual contact networks [123, 124, 126, 129] as well as heterosexual versus MSM networks [129]. The network reconstruction method from viral sequences has not been extensively validated until now, so in this chapter its performance is evaluated against simulated epidemics. In addition, the method is known to reconstruct far more links between nodes than the number of transmission events so I develop and evaluate thinning algorithms which lead to a better estimation of the true transmission network.

6.2 Methods

6.2.1 HIV Epidemic Simulation

An HIV epidemic was simulated for 57 years using the HIV-specific individual-based DSPS [175]. In the underlying SIR model the R category represented AIDS and death (individuals are removed from the pool of infectors). Birth and (AIDS-unrelated) death were set at 0.1 and 0.01, respectively, and progression to AIDS (removal) was set at 0.8. The growth of the population was set to "stable" to ensure only two individuals would populate each deme at any given time point.

The population simulated was fully heterosexual (Figure 6.1). The majority of demes represented households containing a heterosexual couple. One deme contained a cohort of sex workers. The average number of sexual contacts was set at 100 per year and households were classed as either low risk or high risk. For high risk individuals, 50% of contacts were within the household, 30% were with other high risk individuals and 20% were with low risk individuals. For low risk individuals, 80% of contacts were within the household, 15% were with high risk individuals and 5% were with

other low risk individuals. The sex worker deme contained 600 female hosts, each with 600 contacts per year on average. All sex worker contacts are with males and so the sex workers cannot infect each other. The simulation was initiated with the infection of a single sex worker to maximise the likelihood of the epidemic starting.

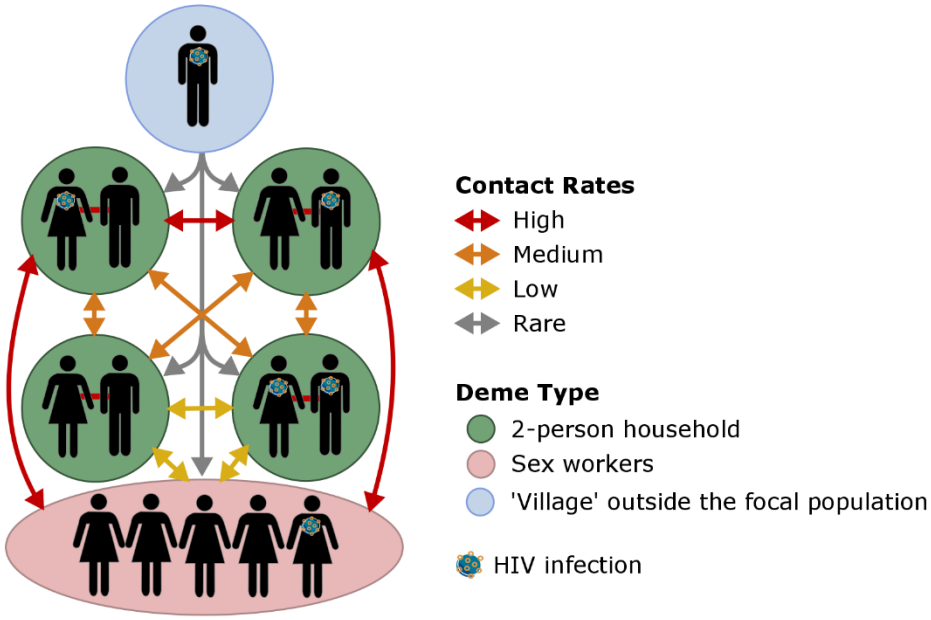


Figure 6.1: Contact network used in the DSPS HIV epidemic simulation © Emma Hodcroft [176]

True phylogenies were generated using the VirusTreeSimulator (see Methods 2.2.3.3). Full genome sequences were simulated along the true phylogenies using piBUSS [177] according to a General Time Reversible model. Rate of evolution was partitioned by gene (gag/ pol: $2.5.10^{-3}$ substitutions per site per year, env: 5.10^{-3}).

This model was developed by Emma Hodcroft as part of the PANGEA_HIV modelling initiative (http://www.pangea-hiv.org/).

6.2.2 Network reconstruction in simulated data

6.2.2.1 Reconstruction of time-resolved phylogenies

Full genome HIV sequences were simulated but this analysis focused on the *pol* region for consistency with previous work and comparison with currently available patient data. Phylogenetic trees were reconstructed independently for each of the three sets of

sampled sequences (100% coverage, 60% and 20%) in RaxML with 100 bootstrap replicates [162]. Trees were transformed into time-scaled trees using Least Squares Dating (http://www.atgc-montpellier.fr/LSD) [173]. LSD runs rapidly on datasets containing tens of thousands of sequences (see Methods 2.2.2.2).

6.2.2.2 Receiver Operating Characteristic (ROC) analysis (see Methods 2.2.5.5)

True times to most recent common ancestors (tMRCA) for each pair of viruses were retrieved from the true phylogeny as well as information on whether the host of those viruses were connected by a direct transmission link. In parallel, reconstructed tMRCAs were retrieved from the time-scaled phylogenies. The method for reconstructing networks from time-scaled phylogenies has been described previously [140] (and see Methods 2.2.4.2). Networks were reconstructed at time depths of 1-10 years. In brief, for each cut-off *x*, two nodes were linked to each other in the network if their tMRCA was below *x*. The objective was then to use ROC analysis to measure our ability to correctly infer whether two sequences were linked or not.

6.2.2.3 True positives (TP)

Ideally, the true reconstructed network should capture only the true sequence of transmission, which in a network containing 4662 nodes, would mean 4661 links only. However, our method of establishing linkage is based on tMRCA so if two sequences share a common ancestor within the prescribed amount of time, our method will link them whether or not that link is direct. Therefore, I conducted two sets of analyses with two different sets of true positives: 1) comparing the links reconstructed to the true direct links (this is the true transmission network) and 2) comparing the links reconstructed to those expected based on the true tMRCAs of each pair of sequences.

6.2.2.4 True negatives (TN)

Our network is dominated by a large number of negative instances: of >10 million possible links in the network, only 4661 are true transmission links (and >10 million are incorrect). This huge number of true negatives (negatives that will be identified correctly as negatives in the reconstruction) will make our method look better than it is (high specificity). As such, I decided to include in our count only links between pairs

of nodes that had a tMRCA below 15 years. This reduced the total number of links (including all TP, TN, false positives (FP) and false negatives (FN)) to 40657 in the dataset of 4662 sequences.

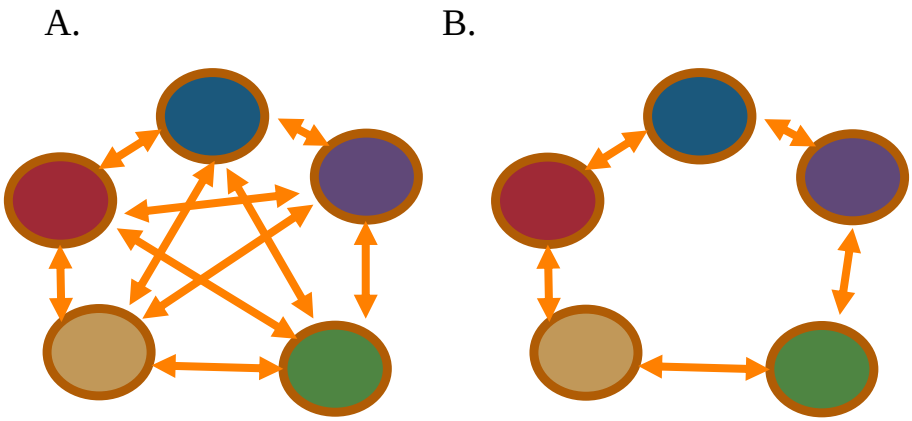


Figure 6.2: Network cliques (A.) and cycles (B.)

6.2.2.5 Network thinning

As stated above, the current network reconstruction method links nodes in the network based solely on tMRCA. The number of links in reconstructed networks by far exceeds the true number of transmission links. In many cases, cliques (Figure 6.2A, where all nodes in a cluster are linked to each other) and cycles (Figure 6.2B, where groups of nodes are connected through a closed walk) remain, both of which reflect the incorrect assignment of links between nodes. Consequently, I implemented thinning algorithms which selectively dropped links within cliques and cycles based on pairwise genetic distance and tMRCA (see Appendix 3). In brief, all cliques are identified within the network [222] and in each clique, the largest edge is eliminated if its length (genetic distance or tMRCA) is greater than the smallest edge length in the clique. Because cliques are not independent from each other, all cliques in the network are then re-estimated before targeting the next edge for deletion. Once all cliques have been eliminated, all cycles are identified within the network and processed in the same way based on the distance of edges. For a more detailed explanation see Appendix 4. For pseudocode, see Appendix 6. I then tested whether thinning improved the precision of our reconstructions (in terms of identifying direct links only).

6.2.2.6 Degree distribution

As well as the ability to reconstruct the true transmission network, I was interested in the ability of our method to correctly estimate the degree distribution of the network. The R [154] package degreenet [126] was used to fit degree distributions (and confidence intervals for best fit parameters) to the Pareto distribution, the Negative Binomial (NB) distribution, the Yule distribution and the Waring distribution. The discrete Pareto is a power law distribution. The other distributions tested are indicative of specific processes of network formation. The NB assumes that individuals in the population acquire partners at a fixed but heterogeneous rate. The Yule [130] and Waring [131] distributions both result from preferential attachment models. They incorporate a probability proportional to k that a new link is made to a person of degree k, as well as a constant probability that a new link is made to a person of degree 0 (previously sexually inactive). In addition, the Waring incorporates a rate for non-preferential attachments, so that a partnership may form at random between two people, regardless of degree [124, 131]. Importantly, the NB always has finite variance, while the Yule and the Waring have infinite variance when their exponent value is between 2 and 3. Infinite variance of a degree distribution indicates that interventions targeted towards the network's high degree nodes will be necessary (see section 1.5.5).

6.2.3 True data

In addition to the simulated data, real HIV *pol* sequences were obtained from the UK HIV RDB (2014 download; Methods 2.1.2). In order to distinguish between UK and non-UK transmission clusters, trees were reconstructed with a background of sequences from LANL, as described previously (Methods 2.1.3). 2507 UK subtype A1, 31450 subtype B and 15815 subtype C sequences were analysed along with 1414 subtype A1, 7413 subtype B and 6212 subtype C LANL sequences.

6.2.3.1 Phylogenetic analysis

For each of the three subtypes, phylogenetic trees were reconstructed in RaxML with 100 bootstrap replicates. Clusters were picked if they were supported by a bootstrap ≥90% and had a maximum genetic distance ≤4.5% [108].

For time resolved analysis, clusters were pooled for analysis in BEAST [76] because of the size of the datasets. Clusters from the ML tree were sorted into pools of up to 300 sequences so that all sequences from the same cluster were analysed together [140]. In the final reconstructed networks (see below), we ensured that no sequences from different clusters has been linked together based on the tMRCAs estimated in BEAST. In a subset of data, I compared the constant, exponential, lognormal, Bayesian skyline plot and Bayesian skyride models. The best model was selected as the skyride by means of its Bayes Factor, as estimated through path sampling. Chains were run for 100,000,000 generations and sampled every 1000 with an SRD06 genetic model and uncorrelated lognormal clock. Chains were run in duplicate, checked for convergence in Tracer and combined. Maximum clade credibility trees were generated in Tree Annotator.

6.2.3.2 From phylogeny to network

In the maximum clade credibility trees, networks were generated as above by linking nodes if their estimated tMRCA was below 5 years. Networks for subtypes A1, B and C were amalgamated for analysis. Thinning was performed as explained above based on pairwise tMRCA and genetic distances within cliques and cycles. Nodes were linked to epidemiological information available (sex and risk group) and degree distributions were generated for the network as whole and for each sex, each risk group and each subtype. Degree distributions were fitted using degreenet [126].

6.3 Results from simulated data

6.3.1 Epidemic dynamics

Sampling was set to begin in year 43 of the simulation (by which time the epidemic had stabilised) to last until the end of the simulation (year 57; Figure 6.3). Based on

the timed series of transmission events (or transmission tree), the true phylogeny of all viruses in hosts alive and infected during the sampling period (4662 in total) was generated. The dataset was down sampled to 60% (2798 nodes) and 20% (933 nodes).

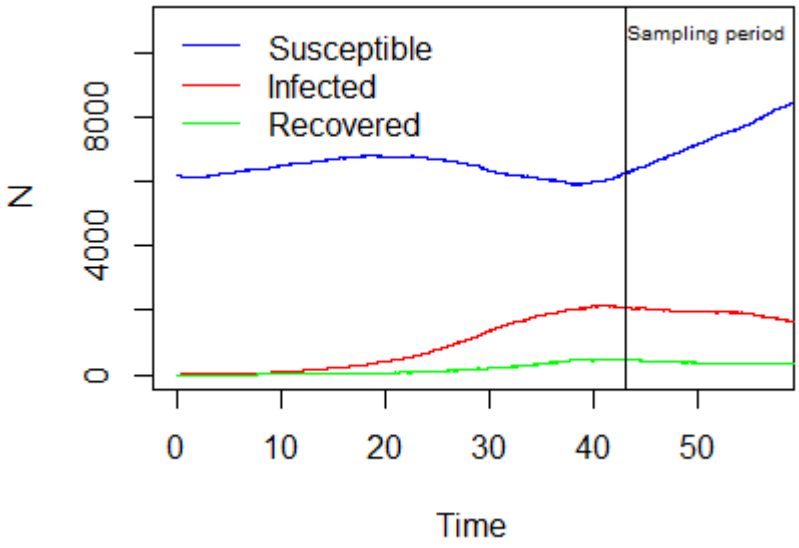


Figure 6.3: Simulation epidemic dynamics

6.3.2 The true and reconstructed tMRCAs are highly correlated

Trees were reconstructed at all three sampling depths independently. These trees were then processed using LSD (see Methods 2.2.2.2) in order to convert branch lengths into calendar time. For each pair of nodes, I then extracted the tMRCA from these reconstructed phylogenies, and in parallel extracted their true tMRCA from the transmission trees. The true tMRCA and reconstructed tMRCA were highly correlated ($r-=0.96$, $p<0.0001$), with reconstructed tMRCAs estimates slightly above true values (Figure 6.4). The correlation stayed significant and did not drop at lower sampling depths. Based on the true transmission network, I was able to distinguish between pairs of nodes that had directly transmitted to each other (direct links) and those who hadn't. The correlation and its significance were not affected by whether direct links only were analysed as opposed to all links.

6.3.3 Sensitivity and specificity of the network reconstruction method are high

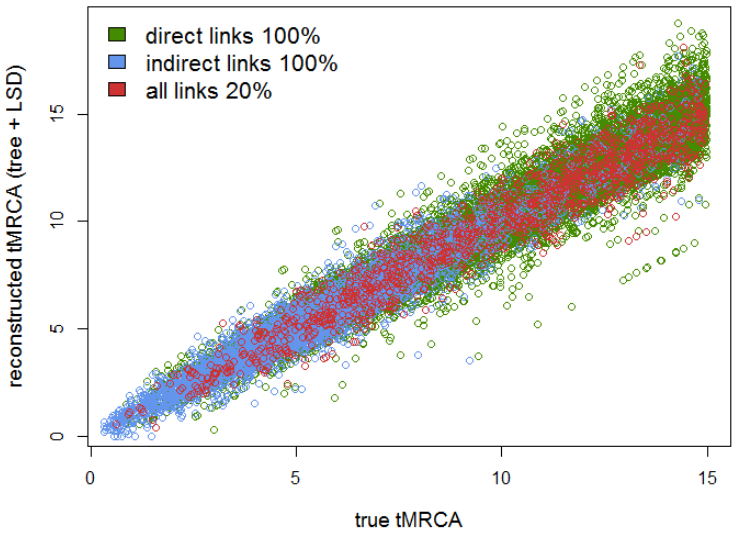


Figure 6.4: Correlation between true tMRCA and reconstructed tMRCA (r=0.96 and p<0.0001 for all sampling depths). Correlation coefficients were estimated between the true tMRCA and the tMRCA for each pair of sequences in reconstructed trees with 100% sampling, 60% sampling and 20% sampling; and between true tMRCAs of all sequence pairs in the tree (all links) and only those pairs linked by direct transmission (direct links).

For a cut-off of 5 years, for example, two sequences would be considered linked (true state = positive) if their true tMRCA ≤ 5 years (Figure 6.5). If their tMRCA> 5 years, the sequences would be unlinked (true state = negative). The true state of the edge is thus categorical. The predictor variable (reconstructed tMRCA), meanwhile, is continuous. ROC analysis tests all the different cut-offs for the predictor variable, and each time calculates the number of TP, FP, TN and FN (Figure 6.5). Based on these numbers, the sensitivity and specificity are calculated (Methods 2.2.5.5). ROC curves plot sensitivity against 1- specificity for every cut-off (Appendix 5, left hand column). Both measures can also be plotted against the cut-off used (Appendix 5, right hand column) in order to identify the cut-off which maximises sensitivity and specificity (best cut-off).

		Reconstruction (all cut-offs tested)	
		Linked	**Unlinked**
True state (e.g. at 5 years)	**Linked (true TMRCA ≤ 5 years)**	True Positive	False Negative
	Unlinked (true TMRCA > 5 years)	False Positive	True Negative

Figure 6.5: ROC analysis: comparing the reconstructed and true state of edges.

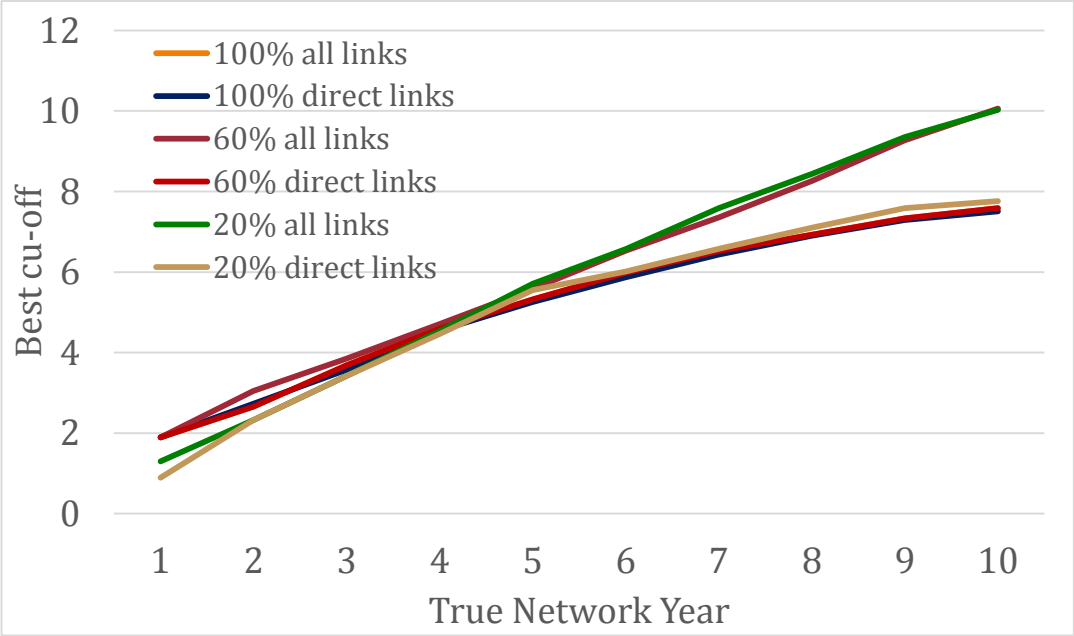


Figure 6.6: Best cut-offs for each network year based on sampling proportion (100%, 60%, and 20%) and whether links were direct or indirect.

The area under the curve (AUC) of ROC curves measures how well the method as a whole is doing at predicting the true state of each edge. For a perfect predictor, the AUC would be equal to 1. In the analyses, the AUC never fell below 0.9, in any of the sample fractions or whether I looked at all links or at direct links only. The performance of the classifier as a whole decreased as network years increased, but AUC stayed above 98% up to 5 years at all thresholds and whether all links were

classed as true, or direct links only. When classifying only direct links as true, the AUC decreased, as did the best cut-off (Appendix 5).

Table 6.1: Accuracy of reconstruction of (direct and time-based) links at three sampling depths. True links are nodes with a true tMRCA ≤ 5 years and a reconstructed tMRCA of 5.38 was used for all

Sampling	positives	TP	FP	TN	FN	sensitivity	specificity	precision
100%	All Links	3260	1031	36173	193	0.94	0.97	0.76
100%	Direct Links	1473	2818	36292	74	0.95	0.93	0.34
60%	All Links	1117	406	12944	72	0.94	0.97	0.73
60%	Direct Links	493	1030	12983	33	0.94	0.93	0.32
20%	All Links	136	47	1631	9	0.94	0.97	0.74
20%	Direct Links	68	115	1636	4	0.94	0.93	0.37

For a true network depth of 5 years, the best selected cut-off remained stable (5.26-5.71, mean=5.45) at all sampling depths (100%, 60%, 20%) and whether I classified all links as true or only direct links (Figure 6.6). The best cut-off for identifying direct links ranged from 5.26 to 5.55 (mean = 5.38). Networks were analysed at a true depth of 5 years in all subsequent analyses with a reconstructed tMRCA of 5.38 years.

At a network depth of 5 years and using a cut-off 5.38 years, the sensitivity and specificity were above 90% (Table 6.1). When all links with a tMRCA ≤ 5 years were considered positive, precision however was ~74%, meaning that of the reconstructed links, ¾ were correct. When only direct links with a tMRCA ≤ 5 years were considered positives, precision was only ~34%, meaning that only 1/3 of reconstructed links were in fact true links. Of note, precision did not drop at lower sampling depth as compared to 100% sampling coverage.

Table 6.2: Accuracy of reconstruction of direct links after thinning at three sampling depths. Thinning was performed by dropping links within cliques and cycles based on genetic distance and tMRCA. True links are nodes with a true tMRCA ≤ 5 years and a reconstructed tMRCA of 5.38 was used for all.

Sampling	positives	thinning	TP	FP	TN	FN	sensitivity	specificity	precision
100%	direct	Distance, no cliques	1249	925	38185	298	0.81	0.98	0.57
100%	direct	Distance, no cliques, no cycles	1246	830	38280	301	0.81	0.98	0.6
100%	direct	tMRCA, no cliques	1237	990	38120	310	0.8	0.97	0.56
100%	direct	tMRCA, no cliques, no cycles	1231	845	38265	316	0.8	0.98	0.59
60%	direct	Distance, no cliques	447	489	13524	79	0.85	0.97	0.48
60%	direct	Distance, no cliques, no cycles	446	465	13548	80	0.85	0.97	0.49
60%	direct	tMRCA, no cliques	450	507	13506	76	0.86	0.96	0.47
60%	direct	tMRCA, no cliques, no cycles	449	462	13551	77	0.85	0.97	0.49
20%	direct	Distance, no cliques	64	82	1669	8	0.89	0.95	0.44
20%	direct	Distance, no cliques, no cycles	64	82	1669	8	0.89	0.95	0.44
20%	direct	tMRCA, no cliques	65	82	1669	7	0.9	0.95	0.44
20%	direct	tMRCA, no cliques, no cycles	65	81	1670	7	0.9	0.95	0.45

6.3.4 Thinning improves precision with little cost to sensitivity and specificity

The aim of the thinning algorithms was to eliminate FP links to improve the precision of the network reconstruction method as the number of FP called was so high. In this case the reconstructed links were counted only against the true direct links in the network. Thinning improved the precision of the reconstruction method, but with some costs to its sensitivity (Table 6.2). At high sample fractions, precision improved the most, but at low sample fraction thinning had the smallest impact on sensitivity. Removing links from cliques and cycles was slightly better than cliques only. There was no difference between distance-based thinning and tMRCA-based thinning. Overall, around 10% of TP links and 52% of FP links were removed by thinning (Figure 6.7).

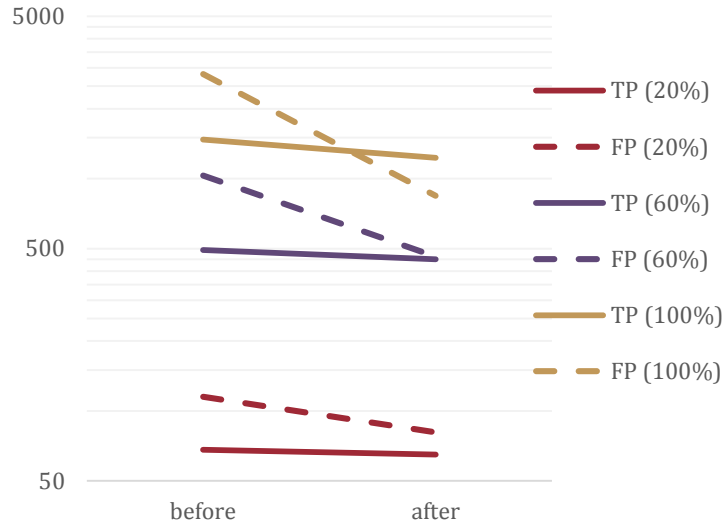


Figure 6.7 Change in the number of true positives (TP) and false positives (FP) after thinning (tMRCA-based with both cliques and cycles thinned).

Thinning removed 37/183 links at 20% sampling (20%), 612/1523 links at 60% (40%) and 2215/4291 links at 100% (52%). We compared the thinning algorithm to an approach where the same proportion of links were dropped at random from the network, (Figure 6.8). We repeated the random thinning 100 times at each sampling depth and each time measured the sensitivity, specificity and precision in the networks

generated. Precision was significantly higher in the networks with targeted thinning, as was sensitivity at 60% and 100% sampling depth.

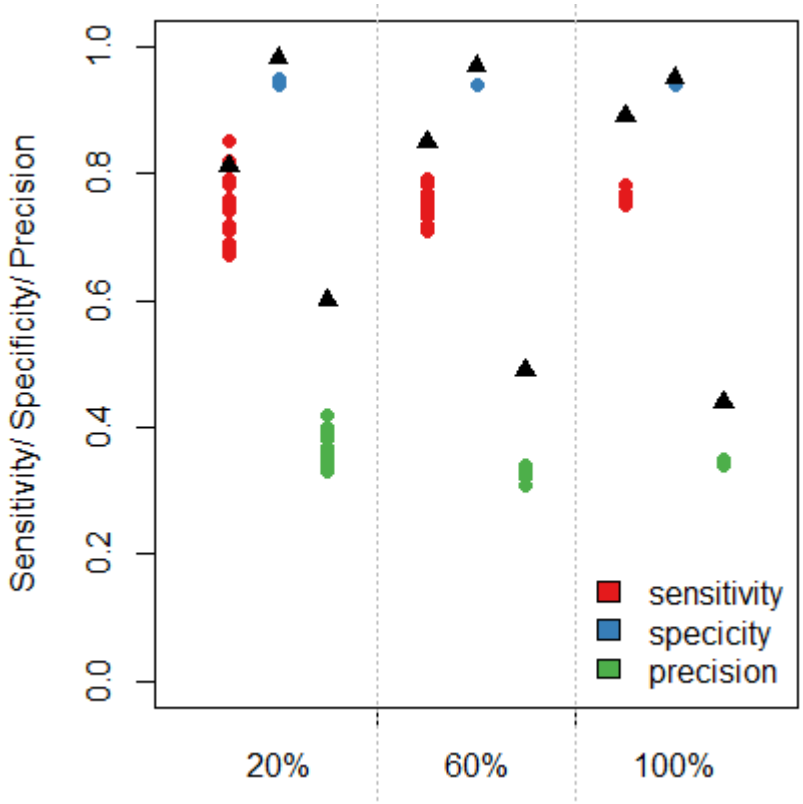


Figure 6.8: Sensitivity, specificity and precision in randomly thinned networks compared to targeted thinning. Coloured distributions show results from 100 replicates, black triangles represent the results for targeted thinning.

6.3.5 Degree distributions

At 100%, mean degree in the thinned networks was closer to that in the true network than mean degree in the unthinned network, but the difference was not significant (overlapping error bars, Figure 6.10). Maximum degree was much higher in the unthinned 100% network than in the true data (23 vs 14).

The true transmission network was best fit by a Waring distribution with an exponent α=28.91. I reconstructed the degree distribution for networks at all three sampling depths, unthinned and with four kinds of thinning (distance-thinning with no cliques, distance-based thinning with no cliques or cycles, tMRCA-thinning with no cliques,

tMRCA-based thinning with no cliques or cycles; Figure 6.9). Each time I evaluated whether the best fit distribution matched that of the true transmission network. Waring was the best fit distribution at 100% sampling and 60% sampling for the unthinned networks and for the distance-based thinned networks (Figure 6.11). The tMRCA-thinned distributions were best fit by NB distributions. The 20% sampled unthinned network and tMRCA-thinned networks were best fit by the NB and the distance-thinned networks were best fit by a Yule distribution. In all cases, the exponent α was above 3. The 95% confidence intervals for α overlapped with the 95% confidence intervals in the true network in all reconstructions except for the unthinned networks at 100% and 60% (Figure 6.11). In all networks analysed α was within the range of finite variance, as was the case in the true transmission network.

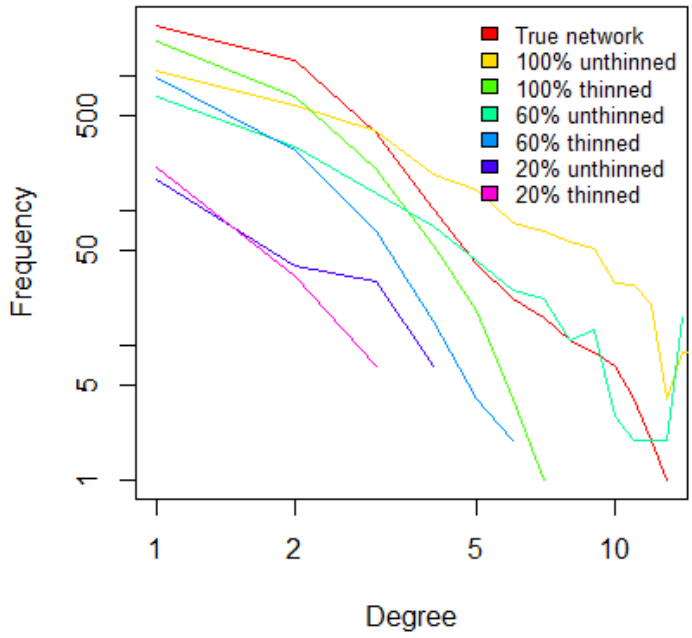


Figure 6.9: Degree distributions of the true network and the unthinned and thinned reconstructed networks (distance-based, no clique, no cycle).

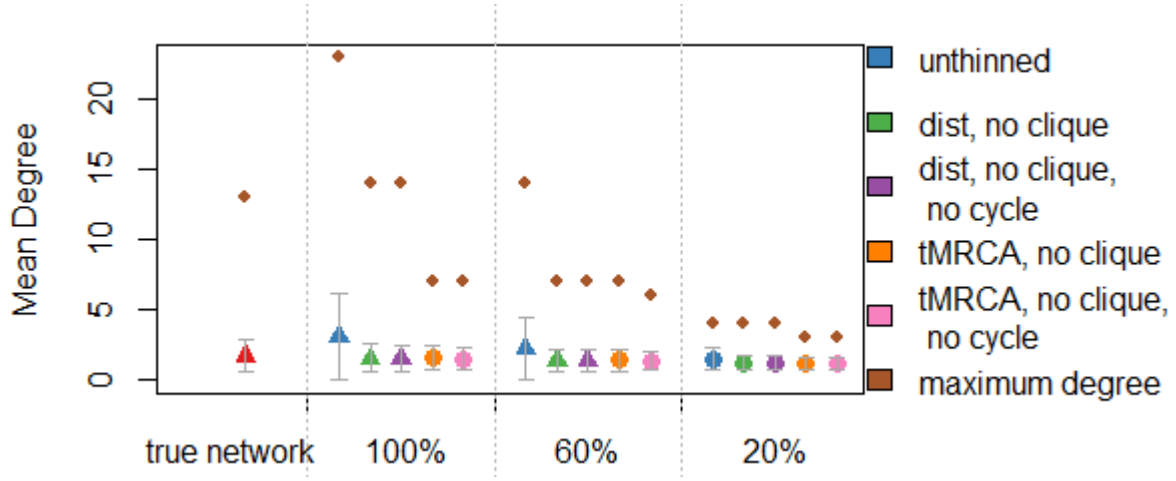


Figure 6.10: Mean and maximum degree in the simulated networks.

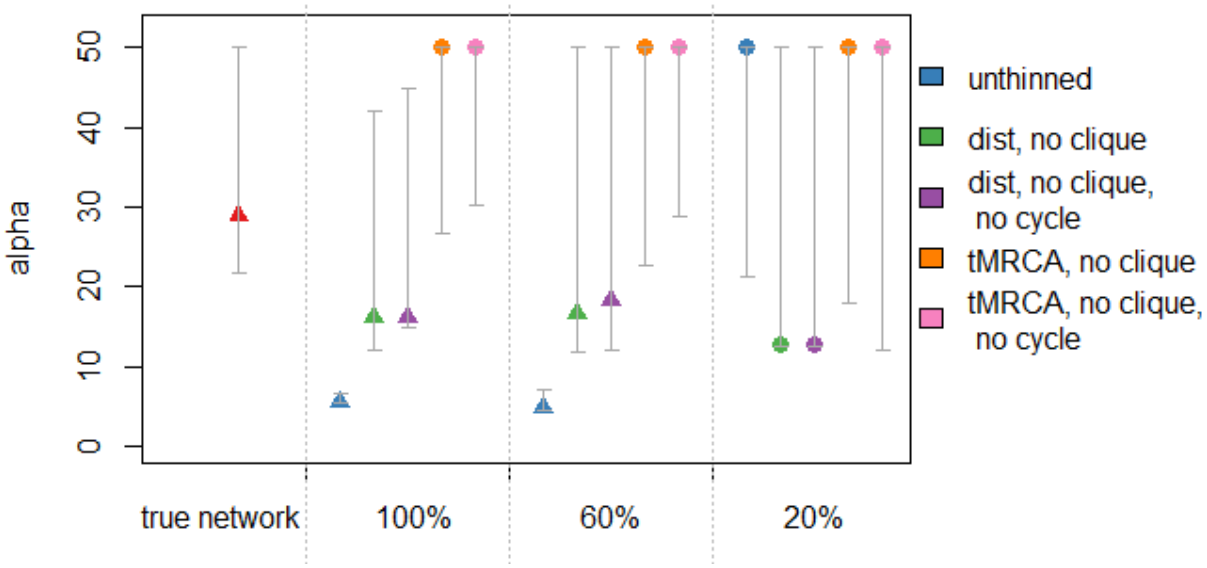


Figure 6.11: Exponent α (Waring distribution) in the true transmission network and according to sampling proportion and thinning method. Triangles represent datasets where Waring was selected as the best fit. Bars represent 95% confidence intervals and were estimated through bootstrapping.

6.4 Degree distributions of the UK HIV epidemic

Network membership for sequences from the UK HIV RDB was established in two stages. Initially, clusters were picked in a ML tree based on genetic distance (≤4.5%)

and bootstrap support (≥90%) for the cluster. All clusters containing at least two UK HIV RDB sequences were selected for analysis in BEAST (Table 6.3). Maximum clade credibility trees generated from BEAST runs were processed using an R script (see Appendix 3) to return the tMRCA and genetic distance for pairs of sequences with tMRCA ≤5 years. The returned tables are network edge lists where each line contains the names of the two linked nodes as well as their genetic distance and tMRCA. Networks for all subtypes and pools were combined for analysis. The breakdown by subtype, risk group and sex for the population as whole is shown in Methods 2.1.2. The breakdown for sequences which stayed linked to at least one other in the final network is shown in Table 6.4. 30% of subtype B sequences remained linked to at least one other compared to only 10.5% of subtype A1 sequences and 12.3% of subtype C sequences (Table 6.3).

Networks were thinned, as described previously. For unthinned and thinned networks, degree distributions were generated for each subtype, risk group and sex, totalling nine categories: females (F), males (M), MSM, PWID, heterosexuals (HET), heterosexual females (HET F), heterosexual males (HET M), subtype B, and subtypes A1 and C.

Table 6.3: Number of sequences and clusters analysed at each stage

	A1	B	C
Total number of sequences	4421 (2507 UK, 1414 LANL)	38863 (31450 UK, 7413 LANL)	22027 (15815 UK, 6212 LANL)
Clustered sequences (% of total)*	993 (22.5%)	16544 (42.6%)	4827 (21.9%)
Number of clusters	315	3519	1731
Sequences linked at 5 years (% of total)*	462 (10.5%)	11750 (30.2%)	2712 (12.3%)
Number of clusters	191	3547	1184

*these numbers include LANL sequences

In the unthinned network, Waring was the best fit distribution for all categories of nodes other than HET F and subtype A1 and C nodes. Subtype B nodes were best fit by a Waring distribution with $2 < \alpha < 3$, as shown previously [140] (Table 6.5, Figure

6.12). This effect was particularly driven by PWID (α=2.73), while for MSM, α was slightly above 3. However, in the thinned networks, only subtype B nodes, PWID and MSM were still best fit by the Waring, and in all cases α exceeded 3 (Table 6.5, Figure 6.12).

Table 6.4 Sex and risk group of patients whose sequences linked to at least one another in a network with a cut-off of 5 years.

		A1	B	C
Sex	F	168 (41.5%)	589 (5.2%)	1124 (42%)
	M	164 (40.5%)	9165 (80.9%)	976 (36.5%)
	NA	73 (18%)	1572 (13.9%)	574 (21.5%)
Risk group	HET	242 (59.8%)	1083 (9.6%)	1759 (65.8%)
	PWID	19 (4.7%)	186 (1.6%)	38 (1.4%)
	MSM	58 (14.3%)	8153 (72%)	241 (9%)
	Other/NA	86 (21.2%)	1904 (16.8%)	636 (23.8%)
Total*	14925	405	11326	2674

*These numbers represent only UK HIV RDB sequences. Females (F), males (M), men who have sex with men (MSM), people who inject drugs (PWID), heterosexuals (HET).

6.5 Discussion

The aim of this study was to evaluate and improve the performance of the published network reconstruction method [140]. Because the method infers far more links than those which could have existed in the true transmission network. I tested thinning algorithms which eliminate links based on pairwise tMRCAs and genetic distance. I then reconstructed thinned UK HIV networks to estimate whether their variance was finite, given the implications of degree variance on intervention strategies (see section 1.5.5).

As before, the unthinned subtype B network was best fit a by a Waring distribution with $2 < \alpha < 3$, indicating infinite variance [140]. Broken down by risk group, α for PWID alone remained <3, while for MSM it was actually slightly above 3. Heterosexuals were also best fit by a Waring but α was well above 3. When

heterosexual men and women were split up, Waring was the best fit for heterosexual men, but the NB was selected as the best model for women. Distinct distributions of sexual partnerships are commonly found for heterosexual men and women [123, 129]. In the UK HIV RDB, one possible explanation is that a proportion of heterosexuals may be misclassified MSM [106] (Chapters 1 and 5). After thinning, the mean and maximum degree for every category of nodes decreased but the change in means was not significant. For females, heterosexuals and heterosexual males who previously had been best fit by a Waring distribution, the NB was the best fit in the thinned networks. For subtype B, males, PWID and MSM, Waring was still the best fit but in all cases α exceeded 3, indicating finite variance. Nonetheless, α remained lowest for PWID. While α may not lie in the range indicating the necessity of targeted interventions, the value for α among PWID still indicates high variance, so targeted interventions may be more efficient. This is an impressive finding given the overall small number of PWIDs in the dataset: 243 only, as compared to 8452 MSM. Taken together, these findings are consistent with our understanding of the difference in epidemic dynamics between risk groups in the UK [108, 109], and with the results presented in Chapter 5.

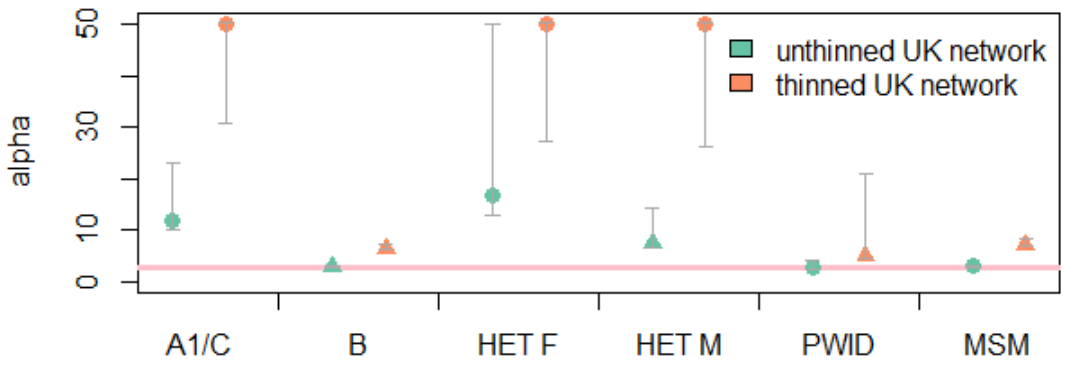


Figure 6.12: Exponent α (Waring distribution) in the thinned and unthinned UK networks. Triangles represent datasets where Waring was selected as the best fit. Bars represent 95% confidence intervals (estimated through bootstrapping). The pink band indicates the range of α signifying infinite variance. Results for thinned networks are from distance-based no clique, no cycle networks. Subtypes A1 and C (A1/C), subtype B (B), heterosexual females (HET F), heterosexual males (HET M), people who inject drugs (PWID), men who have sex with men (MSM).

Table 6.5: Best fit degree distribution for each category of nodes

	Best fit degree distribution and Waring distribution parameter α (95% confidence intervals)	
Node type	Unthinned network	Thinned network
Subtypes A1/C	NB	NB
Subtype B	Waring: 2.93 (2.93-3.16)	Waring: 6.47 (6.3-7.4)
F	Waring: 9.65 (8.4-17.3)	NB
M	Waring: 3.29 (3.28-3.5)	Waring: 10.51 (10.1- 13.4)
HET	Waring: 10.21 (9.34-16.1)	NB
HET F	NB	NB
HET M	Waring: 7.45 (6.42-11.76)	NB
PWID	Waring: 2.73 (2.7-4.1)	Waring: 5.02 (4.7-14.9)
MSM	Waring: 3.09 (3.03-3.3)	Waring: 7.10 (6.8-8.1)

Note: Results for thinned networks are from distance-based no clique, no cycle networks. Females (F), males (M), men who have sex with men (MSM), people who inject drugs (PWID), heterosexuals (HET).

In the simulated data, the unthinned networks contained numerous very high degree nodes (Figure 6.9), leading to an incorrectly estimated very low α. Thinning networks eliminated 50% of FP links, improved our recovery of mean and maximum degree, and generated values for α closer to its true value. However, there were high numbers of FP links in the thinned networks (precision never exceeded 60% even at 100% sampling), and 10% of true positive links were lost through thinning. Therefore while some kind of thinning appears necessary, the exact thinning algorithms developed here may not be best suited to the problem.

I am currently examining how systematic misclassifications of links are. For example, as time between infection and sampling increases, direct transmission links will become more difficult to identify because of within patient evolution, but it would be possible for an algorithm to take additional information into account before deciding whether to link two nodes or not. Some predictors (such as the time between infection and sampling) will never be available in true epidemics, but others like the date of ART initiation or whether or not patients were sampled in early infection (inferred

through CD4+ counts, viral load, genetic diversity or RITA result), can help determine the likelihood or direction of a link.

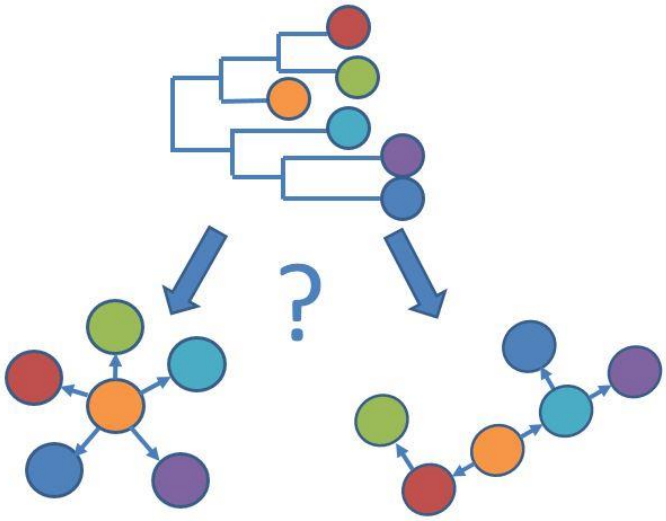


Figure 6.13: Resolution of the transmission chain from a phylogenetic cluster. The phylogeny is compatible with both transmission diagrams, but the transmission diagrams entail different interventions.

The aim of thinning is to better resolve the true chains of transmission. Reconstruction of exact HIV transmission chains through phylogenetic analysis has remained elusive. For large clusters of closely related infections, it is not possible to determine the sequence of transmission events (Figure 6.13). In terms of interventions, the two scenarios represented would engender dramatically different responses. In the first case, an intervention would need to focus on individuals with many contacts. In the second case, meeting places or patterns should be targeted. Here, I wished to improve the estimation of the degree distribution of the true transmission network, because degree distribution is indicative of the intervention strategies suited to the population as whole. Other metrics that may be important in understanding the HIV spread include network clustering coefficients (see section 1.5.3) and assortativity (the tendency of linked nodes to share characteristics). It will be important to evaluate how thinning affects our estimates of those metrics.

Another aim of thinning is to identify high degree nodes that should be targeted with interventions (for example early initiation of treatment). In the same vein, numerous studies have focused on identifying correlates of transmission by identifying clustering on the tree of factors such as the presence or absence of co-infections, including STIs, stage of infection, and drug treatment and compliance [59]. In those types of analyses, all clustered sequences are treated as high transmitters, despite the fact that many of them will be recipients who have not necessarily transmitted the virus onwards. Because a person can only be infected once, if they are linked to many other nodes in a transmission network, all but one are onward transmissions. Therefore, characteristics associated with high degree nodes will be more pertinent to high transmitters than those associated with clustered individuals. We are currently in the process of testing whether we are able to correctly identify high transmitters in the simulated data. In parallel, we have been epidemiologically characterising high degree nodes in the UK data.

One criticism of methods that reconstruct networks from phylogenies is that the phylogeny contains all the information and that some links are artefactually created based on tMRCA that would not be possible based on the phylogeny. However, this assumes that the phylogeny is correctly resolved. It is likely not possible to know the direction of transmission very well from phylogenies based on sequences alone. There are too many uncertainties due to sequence similarity, as demonstrated by the low observed bootstrap support values within clusters. (Full length sequences may however improve the resolution of branching order within clusters, see section 7.8.3). Networks which incorporate this uncertainty can be annotated with additional information (such as RITA result, date of initiation of ART etc.) to eliminate improbable links. Novel methods which improve the resolution of the chain of transmission directly from the phylogeny (section 7.7) will likely perform better than the methods presented here; however at this point they cannot be run on datasets as large as this one. In parallel, approximations such as the one presented here as still useful, especially on such large datasets.

Another limitation of the network reconstruction method is that it appears not to perform well at low sampling densities. In the simulated epidemic sampled at 20% the Waring distribution was not selected as the best fit, and α varied dramatically between the thinned and unthinned networks. In the UK, sampling is around 60% so this will not be an issue, but it will limit the applicability of these methods (in their present form) to low sampled epidemics, such as those of sub-Saharan Africa [223]. It is possible that the method is not limited directly by sampling proportion but by sample size, and in order to test that, I will analyse low sampling proportions (5-20%) from much larger simulated epidemics (50-100,000 nodes). At low sampling densities, exact transmission links will not be captured, however by using more relaxed tMRCA thresholds for linkages, it may be possible to generate degree distributions which approximate the degree distribution of the true transmission network. Another important next step will be to test how well we are able to estimate α in epidemics simulated with $2 < \alpha < 3$.

In this chapter, I used LSD to reconstruct tMRCA in the simulated sequences and BEAST in the UK HIV RDB data. The LSD tMRCA reconstruction performed extremely well in the simulated data (Figure 6.4) but this underlines a problem with the simulated sequences: they were simulated under a GTR model and phylogenies were reconstructed under a GTR model, leading to over-fitting. BEAST may not have improved the tMRCA reconstruction any further. However in the true HIV data, within patient and between patient evolution combined with selection and heterogeneity in time between transmissions mean that a Bayesian genealogical framework will improve reconstructions [224]. With any simulation, there is a risk of over-fitting to that simulation rather than to true data. For that reason, I did not use the 5.38 year cut-off selected in the simulations in the true as it is unlikely to be any more informative that using a 5 year cut-off. In the future, I will evaluate the performance of selected thresholds in an independently simulated epidemic.

In conclusion, the network reconstruction performs well in simulated data at high sampling densities (>60%) consistent with UK coverage, but many non-existent links are inferred. Thinning improves the precision of the reconstruction. The best fit degree

distributions in thinned networks vary by risk group, with heterosexuals displaying much less variation in onward transmissions than MSM and PWID. Among MSM and PWID, the Waring distribution was selected as the best fit but parameters remained within those indicating finite variance. Nonetheless, α was very low among PWID suggesting that interventions targeted towards high degree nodes may be more efficient to curve HIV spread in this group.

7 DISCUSSION

7.1 Summary of findings

In this thesis, I have presented automated tools for the analysis of large sequence databases and used them to compare what are currently the two most densely sampled national epidemics, the UK and Switzerland. This analysis revealed similarities between the two countries' epidemics despite clear variation in risk group composition and distinct proportions of sequences in clusters [97, 105, 108, 109]. I conclude that the current underlying subtype B transmission dynamics in the UK and Switzerland are likely to be similar, including the integration of both countries' epidemics into the European and global epidemics. For this analysis, I focused on subtype B because numbers of non-B subtypes were so low in Switzerland. In contrast, non-B subtypes now represent >50% of diagnoses in the UK [93]. In Chapter 5, I examined the transmission of non-B subtypes within the UK. An increasing proportion of non-B subtypes have previously been shown to have been acquired within the UK [95]. I found that spread was most rapid when non-B subtypes have been introduced into high risk groups (MSM and PWID). This pattern is consistent with observations in France [100] and appears to be to be emerging in Switzerland [97]. Two initially distinct

epidemics, the MSM subtype B, and heterosexual non-B have become connected. Men self-reporting as heterosexuals who have been infected through sex with men are likely to drive the crossing over of subtypes between risk groups [106] (Chapters 4 and 5). In Chapter 6, I estimated the degree distribution for different risk groups in the UK. Differences in fit were in agreement with findings from previous chapters. Importantly, the best fit of the Waring distribution for MSM and PWID networks underlines the importance of identifying the most highly connected individuals within these groups.

7.2 HIV in the UK

Recent years have seen a surge in the number of HIV infections acquired through heterosexual sex, heterosexuals accounted 45% of diagnoses in 2012 [198]. However, as immigration from sub-Saharan Africa has declined, so have diagnoses among heterosexuals [6]. I found that for the most part, onward transmission from this group is very limited (Chapter 5). Heterosexual transmission network structure is consistent with a small number of transmissions per person, indicating that random interventions can be successful in curbing the heterosexual epidemic. The UK epidemic continues to be concentrated within high risk groups. However, heterosexuals, especially men, tend to be diagnosed much later, underlining the need for better testing and increased awareness in this group. Another issue highlighted in this thesis is the possible importance of male heterosexuals being infected through sex with men [106] (Chapters 4 and 5). Sexual questionnaires should be rephrased to recognise such men to accurately estimate their risk factors. These men may play a role in bridging previously disconnected epidemics, and future network analysis will focus on further characterising such nodes. Previous work has already shown that they are more likely to be black African [106]. Similar overlap has been observed in Switzerland [97] and in the USA [225] although to a lesser extent, and bridging may be even more extensive in countries where homosexuality is illegal or more stigmatised than in the UK. MSM are much less likely to openly identify as MSM and are likely to also have female sexual partners [226, 227].

Another finding of Hué *et al.* was the high proportion (42%) of heterosexual clusters containing only women [106]. Female only transmission clusters are likely to represent a bias against sampling heterosexual men, because heterosexual men tend to present later. This issue of "missing men" is widespread in databases from the USA (Joel Wertheim, personal communication) and African countries, where proportionately more women are recruited to many HIV studies (Marcia Kalish, personal communication). Methods to estimate the likelihood of unsampled transmitters within clusters are not yet applicable to datasets of the size of the UK HIV RDB [228]. However, representation in the database has increased over time and if this interpretation is correct, analyses should start to capture these "missing men" retrospectively. Future work will include reconstructing UK HIV RDB clusters through time to investigate whether men appear in clusters (or in their genetic neighbourhood) over time.

In Chapter 5, I found evidence of recent rapid transmission of HIV among PWID despite excellent harm reduction in this group in the UK. This result has instigated a reanalysis of PWID sequence data at the national level to quantify transmission dynamics within this group as such an analysis has never been carried out in the UK. In parallel, the Leigh Brown group is collaborating with local health authorities (who have additional data associated with each sequence) affected to elucidate the cause of these recent outbreaks. One possibility is that some PWID infections could have taken place in prisons, where clean needles are not provided. Another is that infections may be among Eastern Europeans not accessing health services.

This analysis underlines that the UK HIV epidemic continues to be led by MSM, but that even within this group there is high variability in the number of onward transmissions. A recent modelling study found undiagnosed and asymptomatic high sexual activity aged below 35 years to be the key group sustaining the HIV epidemic [229]. The Leigh Brown group are working on characterising highly connected nodes in these networks. If nodes they are connected to represent individuals diagnosed before them, this will confirm that they were transmitting before being diagnosed. Together, these results highlight the importance of better testing among MSM and

suggest there may be a place for PreP in the UK as well. The next few years will demonstrate whether deployment of PreP in the USA has succeeded in reducing HIV incidence. It will be particularly interesting to estimate the effects of PreP at the population level, i.e. how many MSM not taking PreP have been protected by those that do. Uninfected MSM sharing the characteristics of highly connected nodes would be the prime targets for PreP.

7.3 Integration of molecular epidemiology into public health

Taken together, these results demonstrate the insights offered by integrating molecular epidemiology into public health. Some studies have gone even further. In San Diego, phylogenetic analysis was combined with contact tracing, hugely increasing the proportion of sequences clustering [187]. Such a deliberate (and non-anonymised) quest for transmission partners can offer clear opportunities for selecting individuals for interventions, for example if they are part of large active transmission clusters, or have only recently become infected (and are highly infectious). The San Diego group then reconstructed networks through time from HIV sequences and estimated a network connectivity metric for each individual [230]. They found that network connectivity was more predictive of increase in degree (or the number of onward transmissions) than any epidemiological or clinical factors associated with each node. In a follow up study they demonstrated that targeting highly connected nodes with an early intervention (for example early initiation of ART) outperformed any baseline demographics in reducing further transmissions [231]. Similarly in Beijing, targeting nodes based on network connectivity worked better than based on CD4+ counts or viral loads [232]. In Chapter 6, I discussed the possibility of identifying node characteristics associated with high degree. These characteristics could then be used either to select HIV negative individuals with those characteristics for prevention (PreP or behavioural) or to select HIV positive individuals for interventions (early initiation of ART or behavioural). In contrast, the interventions delineated above require reconstruction of an individual's (non-anonymised) network at diagnosis to determine connectivity.

Information on partners in addition to sequence data would make it possible to determine whether transmission occurred between individuals. The collection of behaviour data, such as testing frequency, last negative HIV test, change of behaviour after diagnosis, date of ART initiation and PreP usage, could further bound the timings of transmission events.

7.4 Risk to privacy and stigma

In view of the criminalisation of HIV transmission [145], the use of non-anonymised sequences combined with contact tracing raises concerns over the legal repercussions to source partners that are identified. As coverage in the UK HIV RDB increases, direct transmission links are found more and more frequently. As additional information is incorporated into models (such as dates of ART initiation or recency of infection) it will become possible to identify direction of transmission. The use of additional information can greatly improve the accuracy of reconstructions, and can enable the detailed identification of correlates of transmission [59]. Yet it is also possible that specifically identifying detailed correlates of transmission could increase the stigmatisation of individuals subsequently considered high-risk (both infected and uninfected).

Nonetheless, one can envisage a public health model which makes use of advances in molecular epidemiology methods without compromising the rights of the individual [233]. For example, public health providers could submit the sequence from a new diagnosis (potentially alongside those obtained through contact tracing) anonymously to a third party who would reconstruct phylogenies and networks with a background of fully anonymised sequences to estimate the connectivity of that individual. New diagnoses could be added to the database through a separate procedure.

As discussed in section 2.1.2, the database is anonymised and aggregated so as to prevent deductive disclosure. To some extent, genetic sequences are a form of identifiable information as they are distinct for each patient sampled. The UK HIV RDB comprises sequences from upwards of 60% of the HIV positive population, and releasing sequences into public databases could enable others to fish through those

databases to find sequences linked to their own. The Swiss and UK HIV databases have addressed this concern by making available only 10% of sequences analysed with each publication [105, 140].

7.5 Clustering methods

As explained in Chapter 3, the aim of clustering is to identify subtrees that are epidemiologically meaningful. When trees are large, this may be done for computational reasons, in order to reduce a large dataset into smaller ones. For example in Chapter 6, I preselected sequences within the large phylogenies through clustering before analysis in BEAST, which is computationally demanding. Elucidating the behaviour of clusters may the aim of the analysis in itself, because clusters represent sites of active or recent transmission. As such, the characteristics of clustered individuals are important in terms of focusing interventions. Clusters frequently stratify into risk groups [105], with some overlap between risk groups, revealing the structure of the population and the importance of crossing over between risk groups. For example the Swiss heterosexual epidemic was driven by PWID [105] early in the epidemic, and this continues to be the case in Eastern European countries such as Latvia (Denise Kuhnert, unpublished results). Currently both the Swiss and UK epidemics are more likely to be driven by crossing over from MSM into heterosexuals (Chapters 4 and 5).

However, clustering approaches have severe limitations. Firstly, a lot of information is lost from the tree. Results from the clusters, even though they represent active transmission, are not necessary generalizable to the whole of the study region's epidemic. Secondly, clustering methods rely on relatively dense sampling, or at least sampling biased towards related infections. If sampled infections are distantly related to each other, they are highly unlikely to be part of the same transmission chain and clustering is meaningless. This means that clustering analyses may not be very useful analysing low sampled epidemics in low-resource settings (see section 7.8.1). Thirdly, the choice of clustering thresholds is problematic. In my analysis (Chapter 3), clusters identified were overall consistent at a range of genetic distance thresholds (4.5% to

7.5%), indicating that clusters represented delineated epidemiological units. However there was still some change even when distance threshold was changed only a little, and there is a huge difference between clusters identified at 1.5% and at 4.5% [192]. Standard practise is to perform sensitivity analyses to evaluate the robustness of epidemiological conclusions at a range of clustering thresholds. Furthermore, clustering thresholds meaningful in one epidemic cannot necessarily be extended to another. A (quite relaxed) cut-off of 4.5% was found to delineate clusters in the UK epidemic [108], but this is in part because such a high proportion of sequences are from chronically infected individuals. Appropriate thresholds are affected by sampling fraction, the evolutionary rate within the population, the time between infection and sampling. Nearly all clustering analyses have been performed on *pol*, and cut-offs developed for *pol* are not suited to analyses of other parts of the genome.

Finally, clustering does not reveal the true genealogical relationships between sequences (see section 7.7) and do not fit into traditional epidemiological modelling (see section 7.6).

7.6 Network methods

The main advantage of reconstructing networks as well as (or instead of) phylogenies is that networks fit within a body of work on epidemiological modelling, allowing for testing of hypotheses within a rigorous statistical framework. In this thesis, I focused on estimating the degree distribution from the transmission network because of its impact on the success of different kinds of interventions. Previous analyses of sexual contact network data are limited by the fact that they are likely to miss high degree individuals. One suggestion to increase the likelihood of sampling high degree nodes has been to sample people with STIs [129]. The present method includes data only from HIV positive individuals, and so is much more likely to capture high degree nodes and to better estimate the behaviour of the tail of the distribution.

However, all individuals captured are infected and so transmission networks are being reconstructed rather than contact networks. While on one hand, there may be advantages to capturing only contacts that have led to transmissions, existing

frameworks have been developed based on the sexual contact network and it is at present unclear how the two can be reconciled. The transmission network is a result of the pathogen's behaviour on the contact network and depends on many factors including its period of infectivity and how knowledge of a person's infection status will affect their behaviour [234]. Therefore while contact networks can be used to model the transmission of a number of diseases, transmission networks inferred from pathogen phylogenies will tell us only about the pathogen under study. As sequence databases for other infectious agents grow, it will be possible to compare and contrast networks reconstructed from different pathogens.

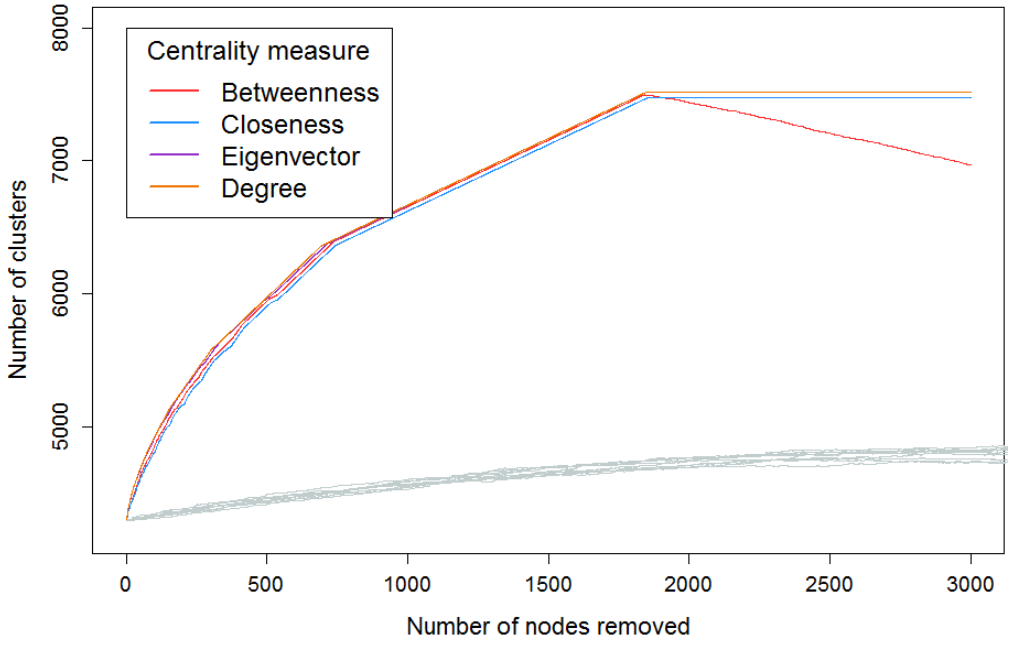


Figure 7.1: Number of distinct clusters in the network increasing as nodes are removed either at random (in grey, 5 replicates) or according to four different centrality measures.

Information on the contacts that didn't lead to transmission can be equally important for understanding disease dynamics [133]. The HIV-DSPS offers clear opportunities to reconcile transmission and sexual contact networks. One next step would be to build sexual contact networks using data from NATSAL, simulate HIV transmission and evolution along those networks and compare simulated phylogenies and transmission networks to those reconstructed from the UK HIV RDB. Another question is whether

transmission networks inferred from sequence data can be used to better calibrate sexual contact networks. Social network data, such as Twitter and Facebook, also offer new opportunities for constructing social and sexual network data, sometimes in parallel with subjects' sexual and drug using behaviours.

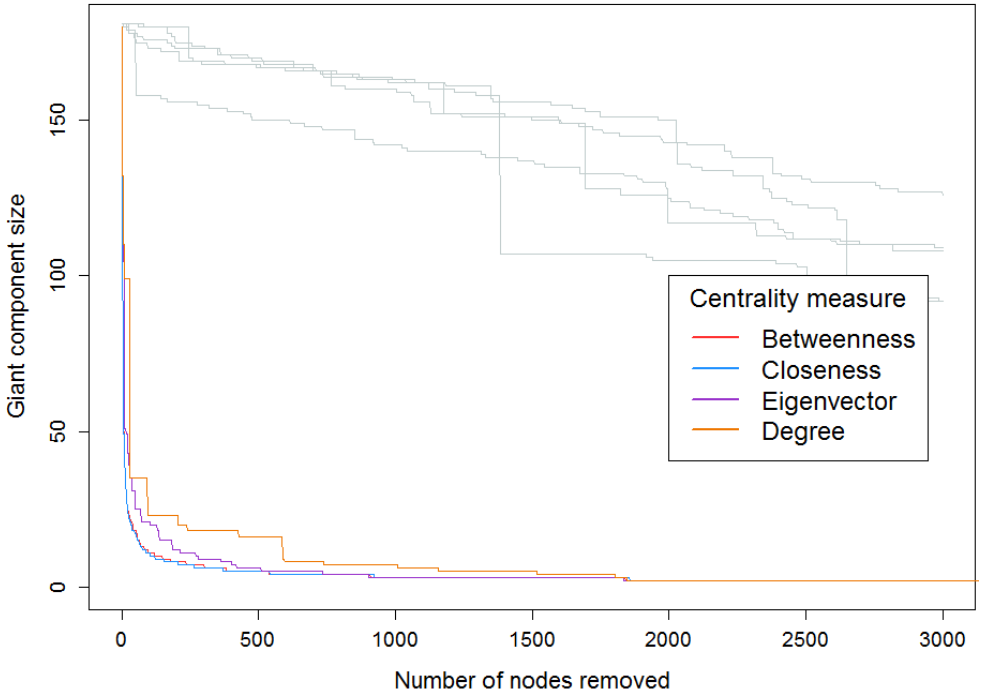


Figure 7.2: Size of the largest cluster in the network (giant component) decreasing as nodes are removed either at random (in grey, 5 replicates) or according to four different centrality measures.

As explained above, representing relationships between sequences in a network enables the isolation of individual nodes within clusters for characterisation; for example high degree nodes and bridging nodes. Many additional metrics are available to describe networks which were not investigated in this thesis. For example centrality measures quantify how important a node is in keeping a network connected. In Chapter 6, I found that targeted interventions may not be necessary in the UK, however preliminary analyses have shown that removing nodes based on centrality measures breaks up the network much more rapidly than doing so at random (Figure 7.1 and Figure 7.2). As targeted interventions are much more difficult to carry out, further

modelling should quantify optimal proportions of targeted and random interventions to maximise impact. Currently, nodes targeted in these networks represent only already infected individuals (who could be targeted with early initiation of ART) but models will eventually include uninfected individuals who could be targeted with PreP.

7.7 Resolving the true chain of transmission

The clustering and network methods developed in this thesis are limited by the fact that they do not resolve the true transmission chain and cannot detect the direction of the transmission. Novel methods have been developed to estimate the direction of transmission events on phylogenies. Overall these methods work by assuming that each internal node represents a transmission event and the branch before that, the virus in its ancestral host. Assuming complete sampling, the probability of each node in the phylogeny being the ancestral host/transmitter can be calculated to estimate the direction of each transmission event. Ancestral states can be reconstructed moving back through the tree from the tips to root.

These methods require highly sampled networks (close to 100%) with little to no introduction of new strains. As such, datasets available usually represent outbreaks of diseases among farm animals, such as foot and mouth [235-237]. Despite the very high coverage of the UK HIV RDB, and even for subtype B which is dominated by within UK transmission, imports are high and the UK epidemic is well integrated in to the global epidemic, so sampling fraction of the full epidemic is likely to be low. Nonetheless, Volz and Frost estimated for pairs of nodes in HIV phylogenies the probability of either being the transmitter [228]. While they were able to infer direction as well as calculate the probability of the transmitter missing from the phylogeny, inferences required detailed information on incidence, prevalence, sampling fraction and stage of infection.

As well as requiring high coverage, these methods are highly computationally intensive. Cottam *et al.* exhaustively mapped all possible transmission histories onto phylogenies for the 2001 UK foot and mouth disease outbreak, calculating for every node in their phylogeny the probability that it represented each host, but this would

not be possible for larger phylogenies or over multiple trees [237]. However, Morelli *et al.* implemented MCMC to sample from the probability distribution of transmission trees without having to examine every single one [236], as well as to iterate over multiple trees.

In order to improve transmission history reconstruction even further, some models now incorporate within host diversity and evolution, as well as estimate the number of unsampled cases [238]. Indeed availability of multiple sequences from patients can reveal direction of transmission and provide detail on the timing of transmission [239]. Expressing mutation rate as a rate per generation can facilitate counting of unsampled cases [240].

While lower sampling decreases the likelihood that transmission pairs will be sampled, algorithms developed to reconstruct transmission histories provide statistical frameworks for testing epidemiological scenarios on trees. Importantly, while results on direction of transmission between pairs are not generalizable, investigating wider epidemiological enquiries may be useful in understanding numerous epidemics. For example, rates of transmission between risk groups can be estimated [228, 235].

7.8 Remaining issues and opportunities

7.8.1 Applicability to low-resource settings

One important final question is whether the methods developed in this thesis can be useful in less densely sampled epidemics, in particular those of sub-Saharan Africa, where the burden of infections is greatest [223]. The network reconstruction method in Chapter 6 performed poorly at 20% sampling – and 20% sampling by far exceeds the sampling fraction currently available in sub-Saharan Africa. It remains to be tested whether low sample number, rather than low sample fraction is the issue and whether it is possible to reconstruct the properties of the true transmission networks when only small numbers of transmission partners are captured. It is possible that these methods will have to be better integrated with phylodynamic methods which use more information from the tree as a whole rather than focusing solely on related transmission

partners. Developing phylogenetic and network methods for sub-Saharan Africa is one of the objectives of the PANGEA-HIV Consortium [241].

7.8.2 Recombination

An underlying assumption of phylogenetics is that a single tree can describe the evolutionary history of all the sequences. However this assumption is violated for HIV, as recombination occurs frequently in the HIV lifecycle (section 1.2.2). Recombination will cause errors in our estimation of phylogenetic relationships and divergence times. Common practice is to look for recombination among sequences before analysis and exclude recombinants. In this thesis, I did not analyse any inter-subtype recombinants but did not eliminate intra-subtype recombinants. While extensive recombination will take place even within subtypes, Lemey *et al.* argue that this does not invalidate transmission trees inferred from HIV phylogenies as sequences analysed for transmission are sufficiently closely related to ensure phylogeny reliability [242]. In our case, analysis focused only on transmission clusters in which recombination is unlikely to have taken pace.

Recombination is likely to become more of a problem when analysing the epidemics of countries with high subtype diversity and frequent co-infection, such as Uganda, one of the counties involved in PANGEA (section 7.8.1). Even in the UK, the proportion of recombination has inarguably been increasing [93] and the recombinant epidemic cannot be ignored altogether.

One solution is to reconstruct the ancestry of sequences through networks in which sequences can have multiple ancestors [29]. To date this type of analysis has been carried out manually, but our group is currently automating the process by integrating repetitive fast single-linkage network construction [99] with a sliding-window approach to scan the entire genome across the whole population.

7.8.3 Novel sequencing methods: deep sequencing and full genome sequencing

For some pathogens, reconstructing transmission through phylogenetic analysis has only become possible because of full genome and deep sequencing. This is the case

for example for tuberculosis, because variability is so low between isolates [243]. HIV mutates so fast that *pol* alone is sufficient for reconstruction transmission trees [144]; however full genome and deep sequences will improve the precision of reconstructions.

Availability of multiple sequences from patients makes it possible to infer direction of transmission [239] based on tree structure, as will sequences obtained from deep sequencing. Estimates of timing of transmission will also be improved [238]. Deep sequencing will also disentangle co-infection. Full genome sequences improve the resolution of phylogenies and will make it easier to reconstruct the order of infections within clusters. In simulates data, HIV phylogenies reconstructed from full genome sequences more closely resemble true viral phylogenies (Gonzalo Yebra, personal communication).

Models will have to be adapted to analyse multiple sequences per patient by accounting for within host genetic diversity (see section 7.7) [238]. Phylogenies will then have to be interpreted differently: nodes will no longer represent solely transmission evets but also divergence within hosts. Full genome availability will also make it necessary to develop methods that account for recombination as longer sequences are much more likely to have undergone recombination. It has been suggested that expanding sequencing may lead to subtypes disappearing altogether [29].

7.9 Conclusion

Phylogenetic analysis can enhance traditional epidemiology by providing insights into HIV transmission patterns. The UK HIV RDB is an incredible resource for understanding the UK epidemic and the quality of the data collected is improving all the time. Better data will enable increasingly detailed analyses but come at a risk for the privacy of the individual. Researchers should be aware of these risks and in parallel have a responsibility to fight against the criminalisation of HIV transmission. While the methods developed in this thesis are useful to tailoring prevention programs in the UK, more efforts are need to develop standardised phylogenetic epidemiology

methods that can be applied as public health tools to prevent new infections in sub-Saharan Africa.

8 APPENDICES

APPENDIX 1: CENTRES CONTRIBUTING DATA TO UK HIV RDB

- **Addenbrooke's Hospital**, Cambridge (Jane Greatorex)
- Chelsea and Westminster Hospital, London (Adrian Wildfire)
- Guy's and St. Thomas' NHS Foundation Trust, London (Siobhan O'Shea, Jane Mullen)
- PHE Birmingham Public Health Laboratory (Erasmus Smit)
- **PHE London** (Tamyo Mbisa)
- Imperial College Health NHS Trust, London (Alison Cox)
- **King's College Hospital**, London (Richard Tandy)
- Leeds Teaching Hospitals NHS Trust (Tony Hale, Tracy Fawcett)
- **Liverpool Specialist Virology Centre**, Royal Liverpool University Hospital (Mark Hopkins, Lynne Ashton)
- **Manchester Specialist Virology Centre,** Central Manchester Foundation Trust (Peter Tilston)
- **Royal Free Hospital**, London (Claire Booth, Ana Garcia-Diaz)
- Royal Infirmary of Edinburgh (Jill Shepherd)
- **Royal Victoria Infirmary,** Newcastle (Matthias Schmid, Brendan Payne)
- **South Tees Hospitals NHS Trust**, Middlesbrough (David Chadwick)
- **St George's Hospital**, London (Phillip Hay, Phillip Rice, Mary Paynter)
- St Bartholomew's and The London NHS Trust (Duncan Clark, David Bibby)
- St Mary's Hospital, London (Steve Kaye)
- University College London Hospitals (Stuart Kirk)
- **West of Scotland Specialist Virology Centre,** Gartnavel General Hospital, Glasgow (Alasdair MacLean, Celia Aitken, Rory Gunson)

APPENDIX 2: FAST TREE VS RAXML CLUSTERS

Subtype C sequences from the 2010 UK HIV RDB release (see section 2.1.2) were analysed, totaling 10830 sequences. Trees were reconstructed in parallel using FastTree [163] and RaxML [162] each with 100 bootstraps. Bootstrap distributions were mostly overlapping, with RaxML showing a slight tendency towards assigning higher bootstraps (Figure 8.1). Clusters were picked in the resulting phylogenies using the Cluster Picker [197] at 90% bootstrap and 4.5% genetic distance. 1805 clusters were identified in the RaxML tree and 1565 in the FastTree phylogeny. 1547 clusters (84%) were identical across the two methods (Figure 8.2).

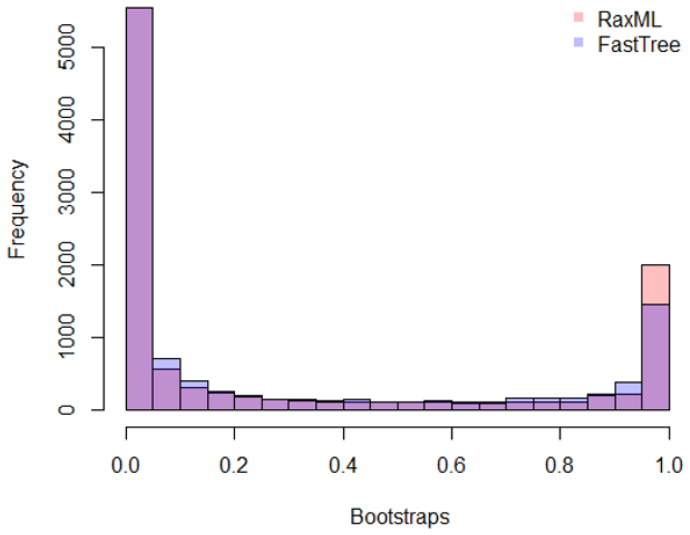


Figure 8.1: Bootstrap distributions from RaxML and FastTree phylogenies

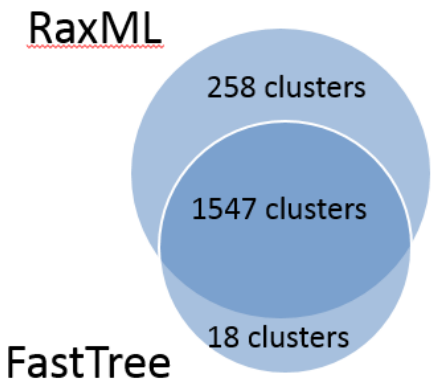


Figure 8.2: Overlap between RaxML and FastTree clusters

APPENDIX 3: CODE AVAILABLE ON CD-ROM

Chapter	Language	Code details
2	R	Remove drug resistant sites in alignments based on International AIDS Society List
3	R	Sort clusters into single origin and multiple origin, do stats, drop tips from trees (all in one script)
4	Python	Launch Cluster Picker in a loop
	Python	Process and merge CPCM output
	R	Process and merge CPCM output
	R	Cluster spreadsheets into degree distributions
	R	Jackknifing degree distributions
5	R	Simulations tree growth/cluster growth
	R	GLM
6	R	Get tMRCA from tree
	R	Thinning algorithms
	R	Fit degree distributions to networks

APPENDIX 4: THINNING ALGORITHMS

The thinning algorithm consists of two steps: thinning of cliques (fully connected clusters), then thinning of cycles (nodes connected through a closed path). Clique and cycle identification are both functions available in the sna package [222] in R.

First all cliques in the unthinned network are identified and sorted by size. The clique with the most nodes is examined first. When more than one clique contains the same maximum number of nodes, all edge lengths (genetic distances or tMRCA) from all those cliques are examined. The algorithm starts with the clique with the highest single edge length (edgeMax), sorts the edges by length and eliminates the edgeMax edge if it is higher than the minimum length edge (edgeMin). The ratio between the two can be altered, so that the edgeMax edge will only be eliminated if it is *x* times larger than edgeMin, but in this thesis we eliminated the edgeMax edge if edgeMax>edgeMin (*x* =1).

It is important to recognise that many cliques and cycles are not independent, and each time a link is removed, many other cliques in the list will be affected. Therefore, every time a link is removed, all cliques are identified again in the network as a whole before another link is selected for elimination.

If edgeMax is not greater than edgeMin in a clique examined, that clique is skipped and the next clique of the same size is examined next. If there are no more cliques of that size, the algorithm proceeds to cliques of the next size down.

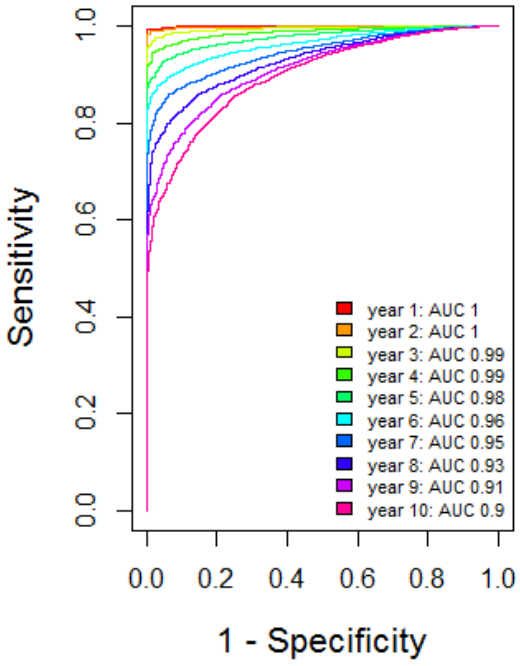
Once all cliques (where edgeMax>edgeMin) have been eliminated (the smallest being triangles), the algorithm returns the clique-thinned network for cycle thinning.

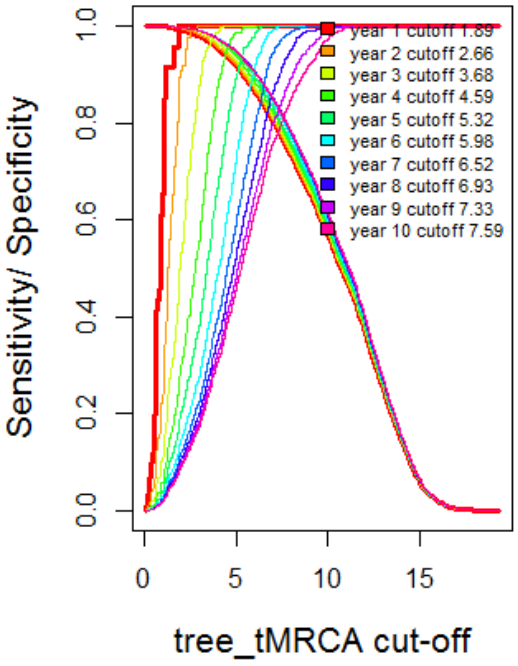
Cycle thinning proceeds in the same way, first identifying all cycles, looking at the cycle containing the most nodes first, identifying edgeMax within that cycle and eliminating the edge if edgeMax>edgeMin.

The number of nodes in the network and the number of clusters are not affected by thinning. All nodes remain in the same clusters as in the unthinned network, but those clusters have fewer links after thinning.

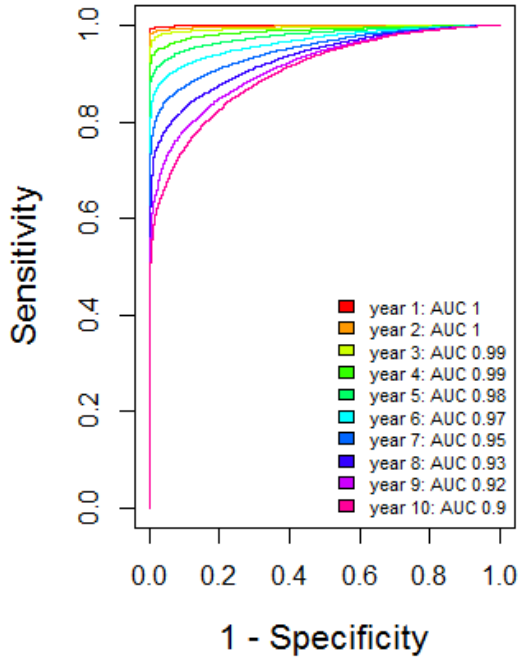
APPENDIX 5: ROC CURVES

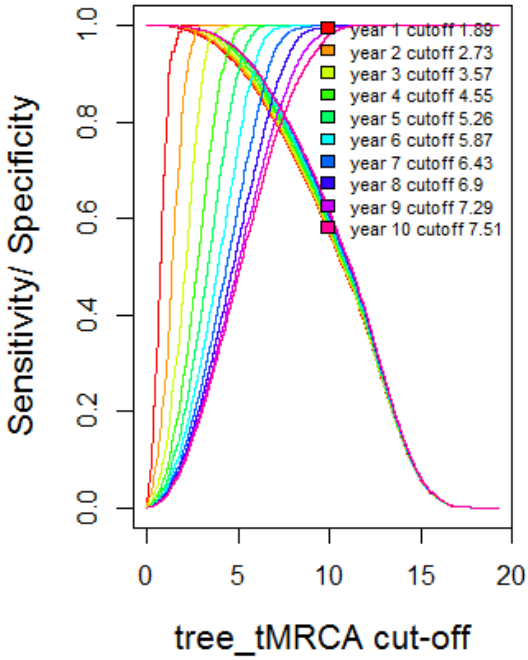
100% - all links



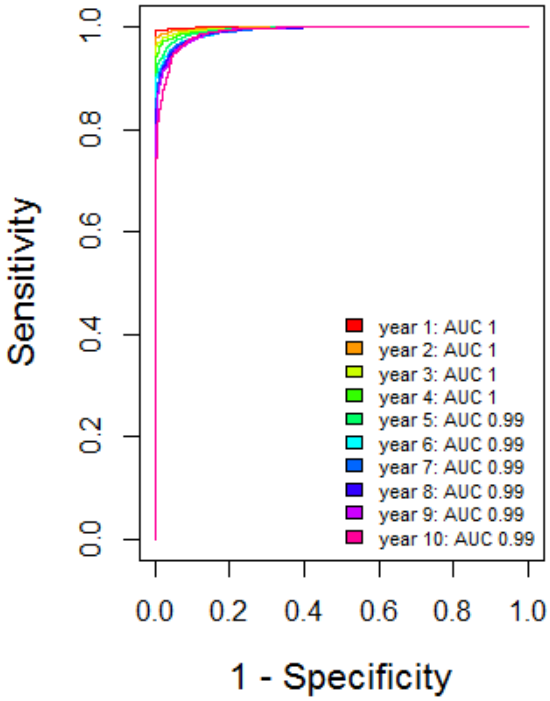


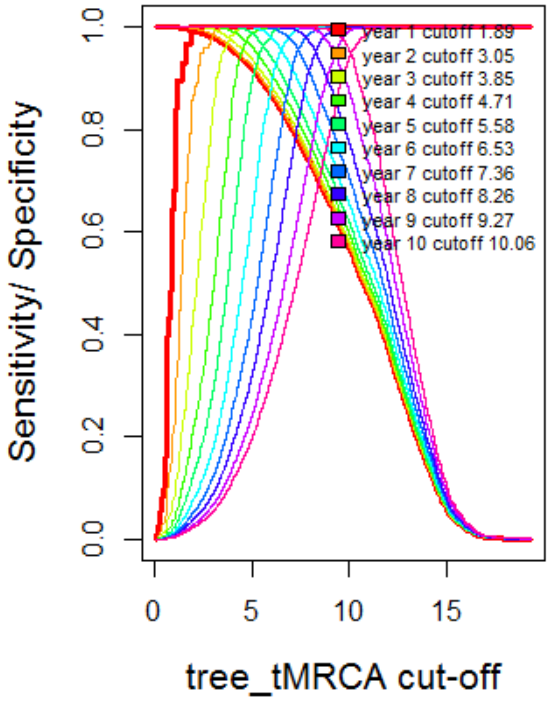
100% - direct links only



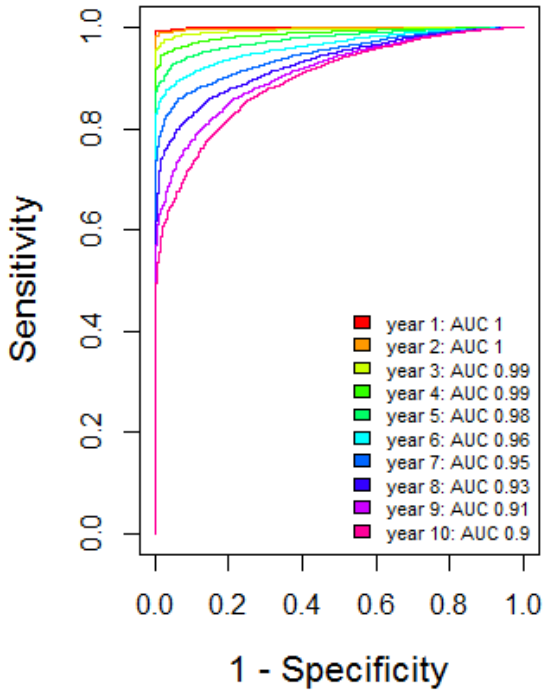


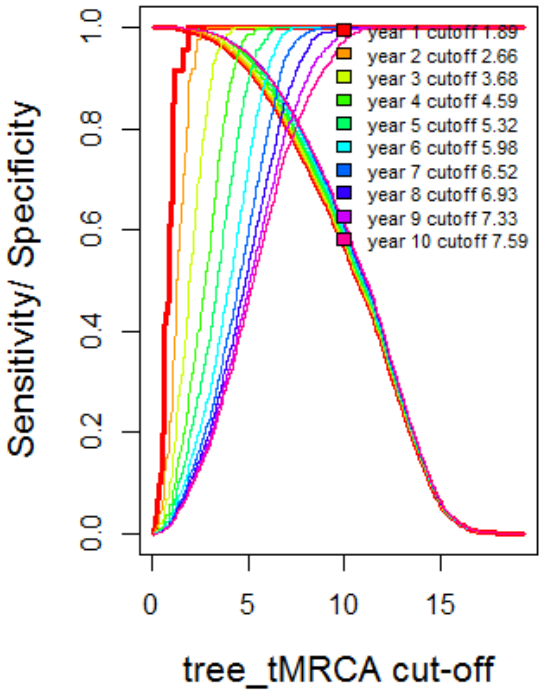
60% - all links



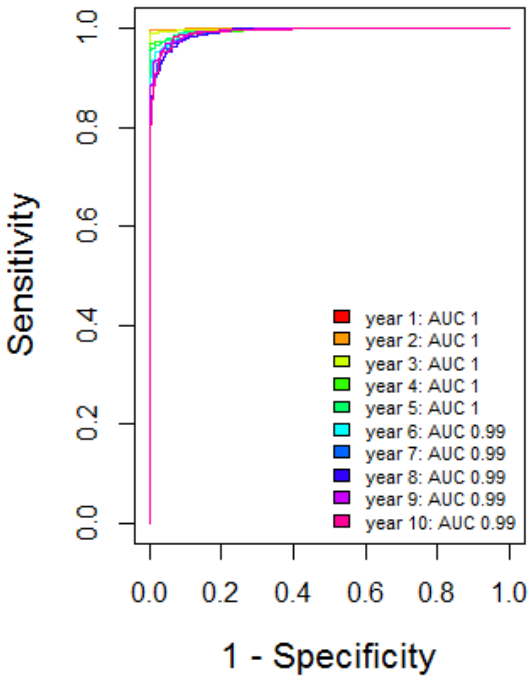


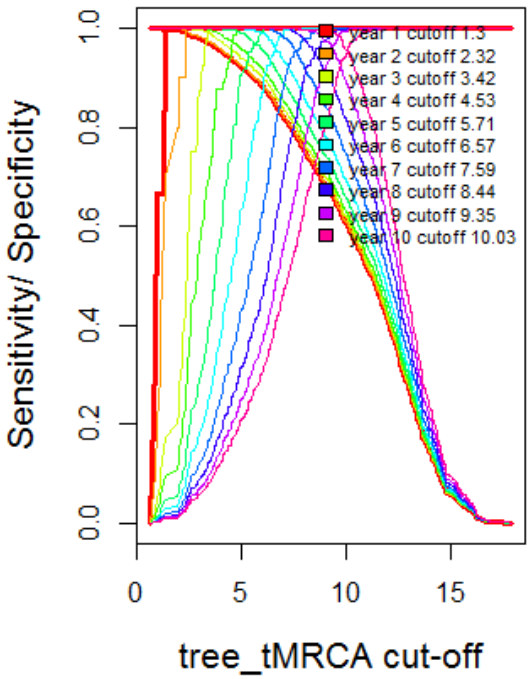
60% - direct links only



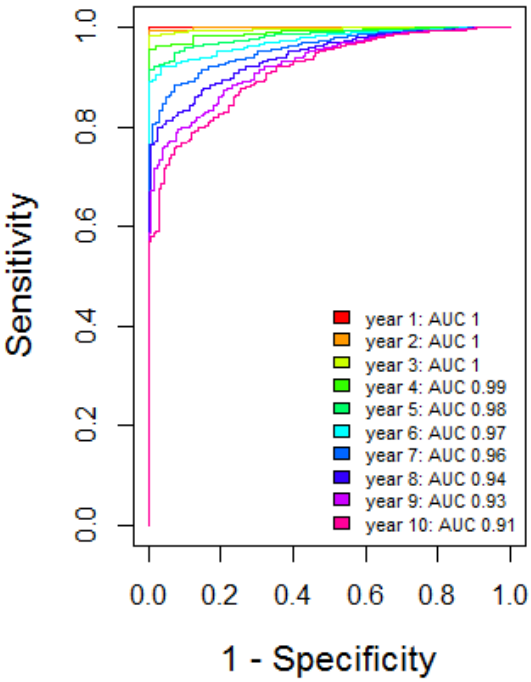


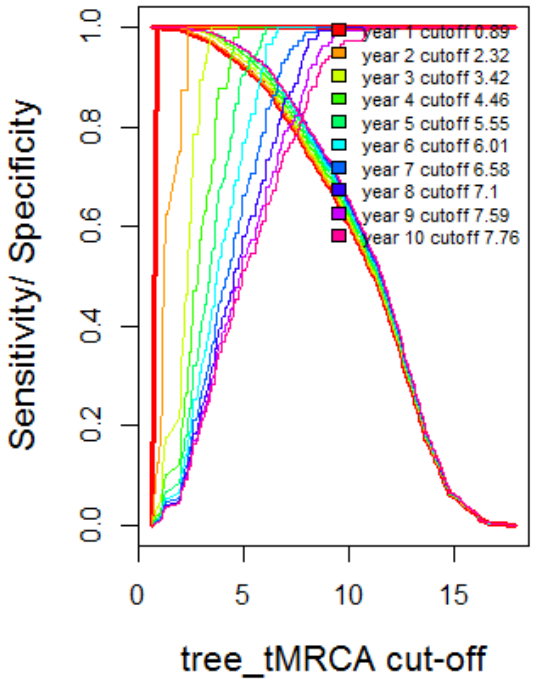
20% - all links





20% - direct links only





APPENDIX 6: PSEUDOCODE

CLUSTER PICKER (CHAPTER 3)

Initial set-up

- SET initial threshold (IT), support threshold (ST), genetic threshold (GT)
- READ sequences and tree from files
- CHECK there is a sequence for every tip (and vice versa)

Recurse through tree and LIST all subtrees with support ≥IT

For each subtree [

Calculate all pairwise genetic distances (from sequences)

Get support from subtree root? node

If (max of pairwise distances ≤ GT) and (support ≤ ST) [Subtree is marked as a cluster]

else [Get all children nodes and add these to the subtree list to be analysed as above]

]

RANDOM ADD TIPS (CHAPTER 5)

READ tree (tip names contain dates)

Get a list of all branch lengths and store in 'BranchLen'

Get a list of all bootstraps and store in 'Bootstraps'

Count number of tips with tip date >2007 (nberTips). Drop all tips with tip date >2007

For the length of nberTips [

Select a branch on the tree (the likelihood of any branch being selected can be proportional to its length or just 1 over the total number of branches)

Select a position along that branch

Sample a bootstrap from Bootstraps and a branch length from branchLen

Create a split in the tree at chosen location, supported by selected Bootstrap and add tip with selected branch length

]

THINNING (CHAPTER 6)

READ network (network contains node names and pairwise distances between each pair of nodes)

'*x*' is set by the user as the multiplier for how much bigger the maximum edge distance has to be compared to the minimum edge distance before you delete that edge (thinning the clique). *x*=1 was used in this thesis.

1. Clique thinning

Identify all cliques (nodes in a cluster all connected to each other) in network and store in 'allCliq' – they are all initially marked as 'unthinned'

This is an inbuilt function in the sna library in R

While 'unthinned' cliques remain in allCliq [

Identify all cliques in network and store in 'allCliq'

Sort cliques by size

Select the 'unthinned' clique with the largest number of nodes

(If more than one clique has the same largest number of nodes, look at all pairwise distances within those cliques and select the one with the single largest pairwise distance)

Sort pairwise distances of the clique. Identify the max (edgeMax) and min (edgeMin)

If edgeMax < (x times edgeMin), mark clique as 'thinned' (it will not be inspected again)

Else: eliminate the clique's edgeMax and repeat loop

]

2. Cycle thinning

Identify all cycles (nodes in a cluster linked up in a loop) in network and store in 'allCycl' – they are all initially marked as 'unthinned'

This is an inbuilt function in the sna library in R

While 'unthinned' cycles remain in allCycl [

Identify all cycles in network and store in 'allCycl'

Sort cycles by size

Select the 'unthinned' cycle with the largest number of nodes

(If more than one cycle has the same largest number of nodes, look at all pairwise distances within those cycles and select the one with the single largest pairwise distance)

Sort pairwise distances of the cycle. Identify the max (edgeMax) and min (edgeMin)

If edgeMax < (x times edgeMin), mark cycle as 'thinned' (it will not be inspected again)

Else: eliminate the cycle's edgeMax and repeat loop

]

9 REFERENCES

1. Centre for Disease Control. Pneumocystis pneumonia - Los Angeles. *Morbidity and Mortality Weekly Report* 1981; **30**:250-252.

2. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, xler-Blin C, Vezinet-Brun F, Rouzioux C, Rozenbaum W and Montagnier L. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 1983; **220(4599)**:868-871.

3. Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, Haynes BF, Palker TJ, Redfield R, Oleske J, Safai B and . Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science* 1984; **224(4648)**:500-503.

4. Coffin J, Haase A, Levy JA, Montagnier L, Oroszlan S, Teich N, Temin H, Toyoshima K, Varmus H, Vogt P and . Human immunodeficiency viruses. *Science* 1986; **232(4751)**:697.

5. UNAIDS. Global statistics fact sheet 2014. In: 2014.

6. Yin Z, Brown AE, Hughes G, Nardone A, Gill ON, Delpech V. HIV in the United Kingdom: 2014 Report. In. Public Health England, London: Health Protection Services; 2014.

7. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Arthur LO, Peeters M, Shaw GM, Sharp PM and Hahn BH. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 1999; **397(6718)**:436-441.

8. Hahn BH, Shaw GM, De Cock KM and Sharp PM. AIDS as a zoonosis: scientific and public health implications. *Science* 2000; **287(5453)**:607-614.

9. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH and Johnson PR. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* 1989; **339(6223)**:389-392.

10. Kuiken C, Korber B and Shafer RW. HIV sequence databases. *AIDS Rev* 2003; **5(1)**:52-61.

11. Hemelaar J, Gouws E, Ghys PD and Osmanov S. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011; **25(5)**:679-689.

12. Kuiken C, Foley B, Hahn BH, Marx P, McCutchan FE, Mellors JW, Mullins JI, Wolinsky S, Korber B. A compilation and analysis of nucleic acid and amino acid sequences. In. Theoretical Biology and Biophysics Group LANL (editor). Los Alamos, New Mexico; 1999.

13. Peeters M, Chaix ML and Delaporte E. Genetic diversity and phylogeographic distribution of SIV: how to understand the origin of HIV. *Med Sci (Paris)* 2008; **24(6-7)**:621-628.

14. Greenwood EJ, Schmidt F, Kondova I, Niphuis H, Hodara VL, Clissold L, McLay K, Guerra B, Redrobe S, Giavedoni LD, Lanford RE, Murthy KK, Rouet F and Heeney JL. Simian Immunodeficiency Virus Infection of Chimpanzees (Pan troglodytes) Shares Features of Both Pathogenic and Non-pathogenic Lentiviral Infections. *PLoS Pathog* 2015; **11(9)**:e1005146.

15. Keele BF, Jones JH, Terio KA, Estes JD, Rudicell RS, Wilson ML, Li Y, Learn GH, Beasley TM, Schumacher-Stankey J, Wroblewski E, Mosser A, Raphael J, Kamenya S, *et al.* Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature* 2009; **460(7254)**:515-519.

16. Klatt NR, Silvestri G and Hirsch V. Nonpathogenic simian immunodeficiency virus infections. *Cold Spring Harb Perspect Med* 2012; **2(1)**:a007153.

17. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JM, Kalengayi RM, Van ME, Gilbert MT and Wolinsky SM. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008; **455(7213)**:661-664.

18. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, Sun C, Grayson T, Wang S, Li H, Wei X, Jiang C, Kirchherr JL, Gao F, *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 2008; **105(21)**:7552-7557.

19. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pepin J, Posada D, Peeters M, Pybus OG and Lemey P. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 2014; **346(6205)**:56-61.

20. Berry N, Davis C, Jenkins A, Wood D, Minor P, Schild G, Bottiger M, Holmes H and Almond N. Vaccine safety. Analysis of oral polio vaccine CHAT stocks. *Nature* 2001; **410(6832)**:1046-1047.

21. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S and Bhattacharya T. Timing the ancestor of the HIV-1 pandemic strains. *Science* 2000; **288(5472)**:1789-1796.

22. Rambaut A, Robertson DL, Pybus OG, Peeters M and Holmes EC. Phylogeny and the origin of HIV-1. *Nature* 2001; **410(6832)**:1047-1048.

23. Wang GP, Ciuffi A, Leipzig J, Berry CC and Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* 2007; **17(8)**:1186-1194.

24. Siliciano RF, Greene WC. HIV latency. *Cold Spring Harb Perspect Med* 2011; **1(1)**:a007096.

25. Pantaleo G, Graziosi C and Fauci AS. New concepts in the immunopathogenesis of human immunodeficiency virus infection. *N Engl J Med* 1993; **328(5)**:327-335.

26. Pilcher CD, Joaki G, Hoffman IF, Martinson FE, Mapanje C, Stewart PW, Powers KA, Galvin S, Chilongozi D, Gama S, Price MA, Fiscus SA and Cohen MS. Amplified transmission of HIV-1: comparison of HIV-1 concentrations in semen and blood during acute and chronic infection. *AIDS* 2007; **21(13)**:1723-1730.

27. Morgan D, Mahe C, Mayanja B, Okongo JM, Lubega R and Whitworth JA. HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS* 2002; **16(4)**:597-603.

28. 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. *J Virol* 1995; **69(8)**:5087-5094.

29. Rambaut A, Posada D, Crandall KA and Holmes EC. The causes and consequences of HIV evolution. *Nat Rev Genet* 2004; **5(1)**:52-61.

30. Hu WS, Temin HM. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proc Natl Acad Sci U S A* 1990; **87(4)**:1556-1560.

31. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, Ron Y, Preston BD and Dougherty JP. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol* 2002; **76(22)**:11273-11282.

32. Robertson DL, Sharp PM, McCutchan FE and Hahn BH. Recombination in HIV-1. *Nature* 1995; **374(6518)**:124-126.

33. Robertson DL, Hahn BH and Sharp PM. Recombination in AIDS viruses. *J Mol Evol* 1995; **40(3)**:249-259.

34. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, Wong K, Stoddard JN, Raugi D, Sorensen S, Genowati I, Birditt B, McKay A, Diem K, *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J Virol* 2011.

35. Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Leigh Brown AJ and Simmonds P. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J Virol* 1993; **67(6)**:3345-3356.

36. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA and Ho DD. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993; **261(5125)**:1179-1181.

37. Kouyos RD, von Wyl V, Yerly S, Boni J, Rieder P, Joos B, Taffe P, Shah C, Burgisser P, Klimkait T, Weber R, Hirschel B, Cavassini M, Rauch A, *et al.* Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis* 2011; **52(4)**:532-539.

38. Meier UC, Klenerman P, Griffin P, James W, Koppe B, Larder B, McMichael A and Phillips R. Cytotoxic T lymphocyte lysis inhibited by viable HIV mutants. *Science* 1995; **270(5240)**:1360-1362.

39. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, Elvin JG, Rothbard JA, Bangham CR, Rizza CR and . Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* 1991; **354(6353)**:453-459.

40. Price DA, Goulder PJ, Klenerman P, Sewell AK, Easterbrook PJ, Troop M, Bangham CR and Phillips RE. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc Natl Acad Sci U S A* 1997; **94(5)**:1890-1895.

41. Frost SD, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, Paxinos E, Chappey C, Galovich J, Beauchaine J, Petropoulos CJ, Little SJ and Richman DD. Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. *Proc Natl Acad Sci U S A* 2005; **102(51)**:18514-18519.

42. Kaslow RA, Carrington M, Apple R, Park L, Munoz A, Saah AJ, Goedert JJ, Winkler C, O'Brien SJ, Rinaldo C, Detels R, Blattner W, Phair J, Erlich H, *et al.* Influence of combinations of human major histocompatibility complex genes on the course of HIV-1 infection. *Nat Med* 1996; **2(4)**:405-411.

43. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI, Walker BD, Ripke S, Brumme CJ, Pulit SL, Carrington M, Kadie CM, Carlson JM, Heckerman D, Graham RR, *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 2010; **330(6010)**:1551-1557.

44. Trachtenberg E, Korber B, Sollars C, Kepler TB, Hraber PT, Hayes E, Funkhouser R, Fugate M, Theiler J, Hsu YS, Kunstman K, Wu S, Phair J, Erlich H, *et al.* Advantage of rare HLA supertype in HIV disease progression. *Nat Med* 2003; **9(7)**:928-935.

45. May MT, Gompels M, Delpech V, Porter K, Orkin C, Kegg S, Hay P, Johnson M, Palfreeman A, Gilson R, Chadwick D, Martin F, Hill T, Walsh J, *et al.* Impact on life expectancy of HIV-1 positive individuals of CD4+ cell count and viral load response to antiretroviral therapy. *AIDS* 2014; **28(8)**:1193-1202.

46. Grant RM, Hecht FM, Warmerdam M, Liu L, Liegler T, Petropoulos CJ, Hellmann NS, Chesney M, Busch MP and Kahn JO. Time trends in primary HIV-1 drug resistance among recently infected persons. *JAMA* 2002; **288(2)**:181-188.

47. Little SJ. Is transmitted drug resistance in HIV on the rise? It seems so. *BMJ* 2001; **322(7294)**:1074-1075.

48. Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. The EuroGuidelines Group for HIV resistance. *AIDS* 2001; **15(3)**:309-320.

49. Gazzard BG. British HIV Association guidelines for the treatment of HIV-1-infected adults with antiretroviral therapy 2008. *HIV Med* 2008; **9(8)**:563-608.

50. Hirsch MS, Gunthard HF, Schapiro JM, Brun-Vezinet F, Clotet B, Hammer SM, Johnson VA, Kuritzkes DR, Mellors JW, Pillay D, Yeni PG, Jacobsen DM and Richman DD. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Clin Infect Dis* 2008; **47(2)**:266-285.

51. Patel P, Borkowf CB, Brooks JT, Lasry A, Lansky A and Mermin J. Estimating per-act HIV transmission risk: a systematic review. *AIDS* 2014; **28(10)**:1509-1519.

52. Gray RH, Wawer MJ, Brookmeyer R, Sewankambo NK, Serwadda D, Wabwire-Mangen F, Lutalo T, Li X, vanCott T and Quinn TC. Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *Lancet* 2001; **357(9263)**:1149-1153.

53. Boily MC, Baggaley RF, Wang L, Masse B, White RG, Hayes RJ and Alary M. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *Lancet Infect Dis* 2009; **9(2)**:118-129.

54. Schwarcz S, Scheer S, McFarland W, Katz M, Valleroy L, Chen S and Catania J. Prevalence of HIV infection and predictors of high-transmission sexual risk behaviors among men who have sex with men. *Am J Public Health* 2007; **97(6)**:1067-1075.

55. Weller S, Davis K. Condom effectiveness in reducing heterosexual HIV transmission. *Cochrane Database Syst Rev* 2001**(3)**:CD003255.

56. Rothenberg RB, Wasserheit JN, St Louis ME and Douglas JM. The effect of treating sexually transmitted diseases on the transmission of HIV in dually infected persons: a clinic-based estimate. Ad Hoc STD/HIV Transmission Group. *Sex Transm Dis* 2000; **27(7)**:411-416.

57. Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, Gray RH and The Rakai Project Study Group. Viral load and heterosexual transmission of Human Immunodeficiency Virus type 1. *N Engl J Med* 2000; **342(13)**:921-929.

58. Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, Laeyendecker O, Kiwanuka N, Kigozi G, Kiddugavu M, Lutalo T, Nalugoda F, Wabwire-Mangen F, Meehan MP and Quinn TC. Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. *J Infect Dis* 2005; **191(9)**:1403-1409.

59. Fisher M, Pao D, Brown AE, Sudarshi D, Gill ON, Cane P, Buckton AJ, Parry JV, Johnson AM, Sabin C and Pillay D. Determinants of HIV-1 transmission in men who have sex with men: a combined clinical, epidemiological and phylogenetic approach. *AIDS* 2010; **24(11)**:1739-1747.

60. Jin F, Jansson J, Law M, Prestage GP, Zablotska I, Imrie JC, Kippax SC, Kaldor JM, Grulich AE and Wilson DP. Per-contact probability of HIV

transmission in homosexual men in Sydney in the era of HAART. *AIDS* 2010; **24(6)**:907-913.

61. UNAIDS, UNODC. Facts about Drug Use and the Spread of HIV. In: 2010.

62. Centre for Disease Control. HIV/AIDS Surveillance Report-cases reported through December 2001. In: 2002.

63. Franceschi S, Dal ML and La VC. Trends in incidence of AIDS associated with transfusion of blood and blood products in Europe and the United States, 1985-93. *BMJ* 1995; **311(7019)**:1534-1536.

64. Rosenberg PS, Goedert JJ. Estimating the cumulative incidence of HIV infection among persons with haemophilia in the United States of America. *Stat Med* 1998; **17(2)**:155-168.

65. Chin J, Sato P and Mann JM. Projections of HIV infections and AIDS cases to the year 2000. *Bull World Health Organ* 1990; **68**:1-11.

66. Gisselquist D, Rothenberg R, Potterat J and Drucker E. Non-sexual transmission of HIV has been overlooked in developing countries. *BMJ* 2002; **324(7331)**:235.

67. Sperling RS, Shapiro DE, Coombs RW, Todd JA, Herman SA, McSherry GD, O'Sullivan MJ, Van Dyke RB, Jimenez E, Rouzioux C, Flynn PM and Sullivan JL. Maternal viral load, zidovudine treatment, and the risk of transmission of human immunodeficiency virus type 1 from mother to infant. Pediatric AIDS Clinical Trials Group Protocol 076 Study Group. *N Engl J Med* 1996; **335(22)**:1621-1629.

68. Townsend CL, Byrne L, Cortina-Borja M, Thorne C, De RA, Lyall H, Taylor GP, Peckham CS and Tookey PA. Earlier initiation of ART and further decline in mother-to-child HIV transmission rates, 2000-2011. *AIDS* 2014; **28(7)**:1049-1057.

69. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA and Holmes EC. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004; **303(5656)**:327-332.

70. Holmes EC, Nee S, Rambaut A, Garnett GP and Harvey PH. Revealing the history of infectious disease epidemics through phylogenetic trees. *Philos Trans R Soc Lond B Biol Sci* 1995; **349(1327)**:33-40.

71. Nee S, Holmes EC, Rambaut A and Harvey PH. Inferring population history from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci* 1995; **349(1327)**:25-31.

72. Ou CY, Ciesielski CA, Myers G, Bandea CI, Luo CC, Korber BT, Mullins JI, Schochetman G, Berkelman RL, Economou AN and . Molecular epidemiology of HIV transmission in a dental practice. *Science* 1992; **256(5060)**:1165-1171.

73. Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, Chen YQ, Martens CA, Ricklefs SM, Selig EJ, Porcella SF, Munshaw S, Ray SC, Piwowar-Manning E, McCauley M, *et al.* Analysis of genetic linkage of HIV from couples enrolled in the HIV Prevention Trials Network 052 trial. *J Infect Dis* 2011; **204(12)**:1918-1926.

74. Rambaut A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 2000; **16(4)**:395-399.

75. Avise JC, Arnold J, Ball JrE, Bermingham T, Lamb JE and Neigel JE. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual Review of Ecology and Systematics* 1987; **18**:489-522.

76. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007; **7**:214.

77. Taylor BS, Sobieszczyk ME, McCutchan FE and Hammer SM. The challenge of HIV-1 subtype diversity. *N Engl J Med* 2008; **358(15)**:1590-1602.

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

79. Novitsky V, Wang R, Lagakos S and Essex M. HIV-1 Subtype C Phylodynamics in the Global Epidemic. *Viruses* 2010; **2(1)**:33-54.

80. Neogi U, Bontell I, Shet A, De CA, Gupta S, Diwan V, Laishram RS, Wanchu A, Ranga U, Banerjea AC and Sonnerborg A. Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C. *PLoS ONE* 2012; **7(6)**:e39819.

81. Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE and Worobey M. The emergence of HIV/AIDS in the Americas and beyond. *Proc Natl Acad Sci U S A* 2007; **104(47)**:18566-18570.

82. Bello G, Eyer-Silva WA, Couto-Fernandez JC, Guimaraes ML, Chequer-Fernandez SL, Teixeira SL and Morgado MG. Demographic history of

HIV-1 subtypes B and F in Brazil. *Infect Genet Evol* 2007; **7(2)**:263-270.

83. Salemi M, de OT, Ciccozzi M, Rezza G and Goodenow MM. High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. *PLoS One* 2008; **3(1)**:e1390.

84. Walker PR, Pybus OG, Rambaut A and Holmes EC. Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect Genet Evol* 2005; **5(3)**:199-208.

85. Bobkova M. Current status of HIV-1 diversity and drug resistance monitoring in the former USSR. *AIDS Rev* 2013; **15(4)**:204-212.

86. Carr JK, Salminen MO, Koch C, Gotte D, Artenstein AW, Hegerich PA, St LD, Burke DS and McCutchan FE. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J Virol* 1996; **70(9)**:5935-5943.

87. Renjifo B, Fawzi W, Mwakagile D, Hunter D, Msamanga G, Spiegelman D, Garland M, Kagoma C, Kim A, Chaplin B, Hertzmark E and Essex M. Differences in perinatal transmission among human immunodeficiency virus type 1 genotypes. *J Hum Virol* 2001; **4(1)**:16-25.

88. Vasan A, Renjifo B, Hertzmark E, Chaplin B, Msamanga G, Essex M, Fawzi W and Hunter D. Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. *Clin Infect Dis* 2006; **42(6)**:843-852.

89. Pinching AJ, McManus TJ, Jeffries DJ, Moshtael O, Donaghy M, Parkin JM, Munday PE and Harris JR. Studies of cellular immunity in male homosexuals in London. *Lancet* 1983; **2(8342)**:126-130.

90. Thomson MM, Najera R. Increasing HIV-1 genetic diversity in Europe. *J Infect Dis* 2007; **196(8)**:1120-1124.

91. Brown AJ, Lobidel D, Wade CM, Rebus S, Phillips AN, Brettle RP, France AJ, Leen CS, McMenamin J, McMillan A, Maw RD, Mulcahy F, Robertson JR, Sankar KN, *et al.* The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland. *Virology* 1997; **235(1)**:166-177.

92. Op de Coul EL, Prins M, Cornelissen M, van der Schoot A, Boufassa F, Brettle RP, Hernandez-Aguado L, Schiffer V, McMenamin J, Rezza G, Robertson R, Zangerle R, Goudsmit J, Coutinho RA, *et al.* Using phylogenetic analysis to trace HIV-1 migration among western

European injecting drug users seroconverting from 1984 to 1997. *AIDS* 2001; **15(2)**:257-266.

93. The UK Collaborative Group on HIV Drug Resistance. The increasing genetic diversity of HIV-1 in the UK, 2002-2010. *AIDS* 2014; **28(5)**:773-780.

94. Aggarwal I, Smith M, Tatt ID, Murad S, Osner N, Geretti AM and Easterbrook PJ. Evidence for onward transmission of HIV-1 non-B subtype strains in the United Kingdom. *J Acquir Immune Defic Syndr* 2006; **41(2)**:201-209.

95. Rice BD, Elford J, Yin Z and Delpech VC. A new method to assign country of HIV infection among heterosexuals born abroad and diagnosed with HIV. *AIDS* 2012; **26(15)**:1961-1966.

96. Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC and Harvey PH. The epidemic behavior of the hepatitis C virus. *Science* 2001; **292(5525)**:2323-2325.

97. von Wyl V, Kouyos RD, Yerly S, Boni J, Shah C, Burgisser P, Klimkait T, Weber R, Hirschel B, Cavassini M, Staehelin C, Battegay M, Vernazza PL, Bernasconi E, *et al.* The role of migration and domestic transmission in the spread of HIV-1 non-B subtypes in Switzerland. *J Infect Dis* 2011; **204(7)**:1095-1103.

98. Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, Van de Vijver DA, Albert J, Angarano G, Asjo B, Balotta C, Boeri E, Camacho R, Chaix ML, Coughlan S, *et al.* Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology* 2009; **6**:49.

99. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM and Kosakovsky Pond SL. The Global Transmission Network of HIV-1. *J Infect Dis* 2014; **209(2)**:304-313.

100. Brand D, Moreau A, Cazein F, Lot F, Pillonel J, Brunet S, Thierry D, Le VS, Plantier JC, Semaille C and Barin F. Characteristics of patients recently infected with HIV-1 non-B subtypes in France: a nested study within the mandatory notification system for new HIV diagnoses. *J Clin Microbiol* 2014; **52(11)**:4010-4016.

101. Brenner BG, Roger M, Routy JP, Moisi D, Ntemgwa M, Matte C, Baril JG, Thomas R, Rouleau D, Bruneau J, Leblanc R, Legault M, Tremblay C, Charest H, *et al.* High rates of forward transmission events after acute/early HIV-1 infection. *J Infect Dis* 2007; **195(7)**:951-959.

102. Pao D, Fisher M, Hue S, Dean G, Murphy G, Cane PA, Sabin CA and Pillay D. Transmission of HIV-1 during primary infection: relationship to

sexual risk and sexually transmitted infections. *AIDS* 2005; **19(1)**:85-90.

103. Brown AE, Gifford RJ, Clewley JP, Kucherer C, Masquelier B, Porter K, Balotta C, Back NK, Jorgensen LB, de MC, Bhaskaran K, Gill ON, Johnson AM and Pillay D. Phylogenetic reconstruction of transmission events from individuals with acute HIV infection: toward more rigorous epidemiological definitions. *J Infect Dis* 2009; **199(3)**:427-431.

104. Volz EM, Koopman JS, Ward MJ, Brown AL and Frost SD. Simple epidemiological dynamics explain phylogenetic clustering of HIV from patients with recent infection. *PLoS Comput Biol* 2012; **8(6)**:e1002552.

105. Kouyos RD, von Wyl V, Yerly S, Boni J, Taffe P, Shah C, Burgisser P, Klimkait T, Weber R, Hirschel B, Cavassini M, Furrer H, Battegay M, Vernazza PL, *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J Infect Dis* 2010; **201(10)**:1488-1497.

106. Hue S, Brown AE, Ragonnet-Cronin M, Lycett SJ, Dunn DT, Fearnhill E, Dolling DI, Pozniak A, Pillay D, Delpech VC and Leigh Brown AJ. Phylogenetic analyses reveal HIV-1 infections between men misclassified as heterosexual transmissions. *AIDS* 2014; **28(13)**:1967-1975.

107. Hue S, Pillay D, Clewley JP and Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A* 2005; **102(12)**:4425-4429.

108. Lewis F, Hughes GJ, Rambaut A, Pozniak A and Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med* 2008; **5(3)**:e50.

109. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A and Leigh Brown AJ. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. *PLoS Pathog* 2009; **5(9)**:e1000590.

110. Anderson RM, May RM. *Infectious Diseases of Humans: Dynamics and Control*: Oxford Science Publications; 1991.

111. Blower SM, Hartel D, Dowlatabadi H, Anderson RM and May RM. Drugs, sex and HIV: a mathematical model for New York City. *Philos Trans R Soc Lond B Biol Sci* 1991; **331(1260)**:171-187.

112. Williams JR, Anderson RM. Mathematical models of the dransmission dynamics of HIV in England and Wales: Mixing between different risk groups. *J R Statist Soc* 1994; **157(1)**:69-87.

113. Handcock MS, Jones JH. Interval estimates for epidemic thresholds in two-sex network models. *Theor Popul Biol* 2006; **70(2)**:125-134.

114. Erdos P, Renyi A. On Random Graphs I. *Publ Math Debrecen* 1959; **6**:290-297.

115. Milgram S. The Small-World Problem. *Psychol Today* 1967; **1(1)**:61-67.

116. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature* 1998; **393(6684)**:440-442.

117. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science* 1999; **286(5439)**:509-512.

118. Anderson RM, Medley GF, May RM and Johnson AM. A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J Math Appl Med Biol* 1986; **3(4)**:229-263.

119. Albert R, Jeong H and Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000; **406(6794)**:378-382.

120. Dezso Z, Barabasi AL. Halting viruses in scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2002; **65(5 Pt 2)**:055103.

121. Moore C, Newman ME. Epidemics and percolation in small-world networks. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2000; **61(5 Pt B)**:5678-5682.

122. Boguna M, Pastor-Satorras R and Vespignani A. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys Rev Lett* 2003; **90(2)**:028701.

123. Hamilton DT, Handcock MS and Morris M. Degree distributions in sexual networks: a framework for evaluating evidence. *Sex Transm Dis* 2008; **35(1)**:30-40.

124. Handcock MS, Jones JH. Likelihood-based inference for stochastic models of sexual network formation. *Theor Popul Biol* 2004; **65(4)**:413-422.

125. Jones JH, Handcock MS. Social networks: Sexual contacts and epidemic thresholds. *Nature* 2003; **423(6940)**:605-606.

126. Jones JH, Handcock MS. An assessment of preferential attachment as a mechanism for human sexual network formation. *Proc Biol Sci* 2003; **270(1520)**:1123-1128.

127. Kelly WP, Ingram PJ and Stumpf MP. The degree distribution of networks: statistical model selection. *Methods Mol Biol* 2012; **804**:245-262.

128. Liljeros F, Edling CR, Amaral LA, Stanley HE and Aberg Y. The web of human sexual contacts. *Nature* 2001; **411(6840)**:907-908.

129. Schneeberger A, Mercer CH, Gregson SA, Ferguson NM, Nyamukapa CA, Anderson RM, Johnson AM and Garnett GP. Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in Britain and Zimbabwe. *Sex Transm Dis* 2004; **31(6)**:380-387.

130. Simon H. On a class of skew distribution functions. *Biometrika* 1955; **42(3/4)**:435-440.

131. Irwin JO. The place of mathematics in medical and biological statistics. *J R Statist Soc* 1963; **126**:1-45.

132. Keeling MJ, Eames KT. Networks and epidemic models. *J R Soc Interface* 2005; **2(4)**:295-307.

133. Eames K, Bansal S, Frost S and Riley S. Six challenges in measuring contact networks for use in modelling. *Epidemics* 2015; **10**:72-77.

134. Mercer CH, Tanton C, Prah P, Erens B, Sonnenberg P, Clifton S, Macdowall W, Lewis R, Field N, Datta J, Copas AJ, Phelps A, Wellings K and Johnson AM. Changes in sexual attitudes and lifestyles in Britain through the life course and over time: findings from the National Surveys of Sexual Attitudes and Lifestyles (Natsal). *Lancet* 2013; **382(9907)**:1781-1794.

135. Drumright LN, Frost SD. Rapid social network assessment for predicting HIV and STI risk among men attending bars and clubs in San Diego, California. *Sex Transm Infect* 2010; **86 Suppl 3**:iii17-iii23.

136. Frost SD. Using sexual affiliation networks to describe the sexual structure of a population. *Sex Transm Infect* 2007; **83 Suppl 1**:i37-i42.

137. Krivitsky PN, Morris M. Inference for Social Network Models from Egocentrically-Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US. *(submitted)* 2015:319-339.

138. Klovdahl AS. Social networks and the spread of infectious diseases: the AIDS example. *Soc Sci Med* 1985; **21(11)**:1203-1216.

139. Resik S, Lemey P, Ping LH, Kouri V, Joanes J, Perez J, Vandamme AM and Swanstrom R. Limitations to contact tracing and phylogenetic analysis in establishing HIV type 1 transmission networks in Cuba. *AIDS Res Hum Retroviruses* 2007; **23(3)**:347-356.

140. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E and Dunn DT. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J Infect Dis* 2011; **204(9)**:1463-1469.

141. Leventhal GE, Kouyos R, Stadler T, Wyl V, Yerly S, Boni J, Cellerai C, Klimkait T, Gunthard HF and Bonhoeffer S. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol* 2012; **8(3)**:e1002413.

142. Balfe P, Simmonds P, Ludlam CA, Bishop JO and Leigh Brown AJ. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. *J Virol* 1990; **64(12)**:6221-6233.

143. Holmes EC, Leigh Brown AJ and Simmonds P. Sequence data as evidence. *Nature* 1993; **364(6440)**:766.

144. Hue S, Clewley JP, Cane PA and Pillay D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 2004; **18(5)**:719-728.

145. Pillay D, Rambaut A, Geretti AM and Leigh Brown AJ. HIV phylogenetics. *BMJ* 2007; **335(7618)**:460-461.

146. Deng W, Nickle DC, Learn GH, Maust B and Mullins JI. ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics* 2007; **23(17)**:2334-2336.

147. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P and Drummond A. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012; **28(12)**:1647-1649.

148. Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Gunthard HF, Telenti A, Furrer H, Yerly S and Francioli P. Cohort profile: the Swiss HIV Cohort study. *Int J Epidemiol* 2010; **39(5)**:1179-1189.

149. Grant RM, Kuritzkes DR, Johnson VA, Mellors JW, Sullivan JL, Swanstrom R, D'Aquila RT, Van GM, Holodniy M, Lloyd JR, Jr., Reid C, Morgan GF and Winslow DL. Accuracy of the TRUGENE HIV-1 genotyping kit. *J Clin Microbiol* 2003; **41(4)**:1586-1593.

150. Shafer RW. Rationale and uses of a public HIV drug-resistance database. *J Infect Dis* 2006; **194 Suppl 1**:S51-S58.

151. Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AF, Hughes G, Fearnhill E, Gravenor MB, Leigh Brown AJ and Frost SD. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput Biol* 2009; **5(11)**:e1000581.

152. de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AM, van d, V, Boucher CA, Camacho R and Vandamme AM. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics* 2005; **21(19)**:3797-3800.

153. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 1999; **41**:95-98.

154. R Development Core Team. R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing; 2011.

155. Bennett DE, Camacho RJ, Otelea D, Kuritzkes DR, Fleury H, Kiuchi M, Heneine W, Kantor R, Jordan MR, Schapiro JM, Vandamme AM, Sandstrom P, Boucher CA, van d, V, *et al.* Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PLoS ONE* 2009; **4(3)**:e4724.

156. Pillay D. Current patterns in the epidemiology of primary HIV drug resistance in North America and Europe. *Antivir Ther* 2004; **9(5)**:695-702.

157. Shafer RW, Rhee SY, Pillay D, Miller V, Sandstrom P, Schapiro JM, Kuritzkes DR and Bennett D. HIV-1 protease and reverse transcriptase mutations for drug resistance surveillance. *AIDS* 2007; **21(2)**:215-223.

158. Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer R, Wensing AM and Richman DD. 2011 update of the drug resistance mutations in HIV-1. *Top Antivir Med* 2011; **19(4)**:156-164.

159. Edwards AWF, Cavalli-Sforza LL. Reconstruction of evolutionary trees. In: *Phenetic and Phylogenetic Classification*. Heywood VH, McNeill J (editors). London: Systematics Association; 1964. pp. 67-76.

160. Leitner T, Escanilla D, Franzen C, Uhlen M and Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proc Natl Acad Sci U S A* 1996; **93(20)**:10864-10869.

161. Holmes EC, Zhang LQ, Simmonds P, Rogers AS and Leigh Brown AJ. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *J Infect Dis* 1993; **167(6)**:1411-1414.

162. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 2006; **22(21)**:2688-2690.

163. Price MN, Dehal PS and Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* 2010; **5(3)**:e9490.

164. Zuckerkandl E, Pauling LB. Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in Biochemistry*. Kasha M, Pullman B (editors). New York: Academic Press; 1962. pp. 189-225.

165. Aulicino PC, Holmes EC, Rocco C, Mangano A and Sen L. Extremely rapid spread of human immunodeficiency virus type 1 BF recombinants in Argentina. *J Virol* 2007; **81(1)**:427-429.

166. Guimaraes ML, Vicente AC, Otsuki K, da Silva RF, Francisco M, da Silva FG, Serrano D, Morgado MG and Bello G. Close phylogenetic relationship between Angolan and Romanian HIV-1 subtype F1 isolates. *Retrovirology* 2009; **6**:39.

167. Bello G, Guimaraes ML, Passaes CP, Matos Almeida SE, Veloso VG and Morgado MG. Short communication: Evidences of recent decline in the expansion rate of the HIV type 1 subtype C and CRF31_BC epidemics in southern Brazil. *AIDS Res Hum Retroviruses* 2009; **25(11)**:1065-1069.

168. Tully DC, Wood C. Chronology and evolution of the HIV-1 subtype C epidemic in Ethiopia. *AIDS* 2010; **24(10)**:1577-1582.

169. Esbjornsson J, Mild M, Mansson F, Norrgren H and Medstrand P. HIV-1 molecular epidemiology in Guinea-Bissau, West Africa: origin, demography and migrations. *PLoS One* 2011; **6(2)**:e17025.

170. Posada D, Crandall KA. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (HIV-1). *Mol Biol Evol* 2001; **18(6)**:897-906.

171. Sanderson MJ. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* 1997; **14**:1218-1231.

172. Drummond AJ, Ho SY, Phillips MJ and Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 2006; **4(5)**:e88.

173. To TH, Jung M, Lycett SJ and Gascuel O. Fast dating using least-squares criteria and algorithms. *(submitted)* 2015.

174. Prosperi MC, Ciccozzi M, Fanti I, Saladini F, Pecorari M, Borghi V, Di GS, Bruzzone B, Capetti A, Vivarelli A, Rusconi S, Re MC, Gismondo MR, Sighinolfi L, *et al.* A novel methodology for large-scale phylogeny partition. *Nat Commun* 2011; **2**:321.

175. Lycett S, Hodcroft E, Leigh Brown AJ and Kao RR. Phylodynamics scenario simulation using the DiscreteSpatialPhyloSimulator. *(in preparation)* 2015.

176. Hodcroft E. Estimating the Heritability of Virulence in HIV. In: University of Edinburgh; 2015.

177. Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A and Suchard MA. piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics* 2014; **15**:133.

178. Butts CT. network: Classes for Relational Data. *J Stat Softw* 2008; **24(2)**:nihpa54860.

179. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal* 2006; **Complex Systems**:1695.

180. Efron BE. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* 1979; **7(1)**:1-26.

181. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* 1985; **39(4)**:783-791.

182. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 1951; **46(253)**:68-78.

183. Aliakbary S, Habibi J and Movaghar A. Feature Extraction from Degree Distribution for Comparison and Analysis of Complex Networks. *The Computer Journal* 2015.

184. Edwards CT, Holmes EC, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE and Drummond AJ. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol Biol* 2006; **6**:28.

185. Lemey P, Derdelinckx I, Rambaut A, Van LK, Dumont S, Vermeulen S, Van WE and Vandamme AM. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol* 2005; **79(18)**:11981-11989.

186. Brenner B, Wainberg MA and Roger M. Phylogenetic inferences on HIV-1 transmission: implications for the design of prevention and treatment interventions. *AIDS* 2013; **27(7)**:1045-1057.

187. Smith DM, May SJ, Tweeten S, Drumright L, Pacold ME, Kosakovsky Pond SL, Pesano RL, Lie YS, Richman DD, Frost SD, Woelk CH and Little SJ. A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS* 2009; **23(2)**:225-232.

188. Robertson DL, Anderson JP, Bradac JK, Carr JK, Foley B, Gao F. HIV-1 Nomenclature Poposal. In: *Human Retroviruses and AIDS*. Kuiken C, McCutchan FE, Marx P, Mellors JW, Mullins JI, Wolinsky S (editors). Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory; 1999. pp. 492-505.

189. Paradis E, Claude J and Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004; **20(2)**:289-290.

190. Chevenet F, Jung M, Peeters M, de OT and Gascuel O. Searching for virus phylotypes. *Bioinformatics* 2013; **29(5)**:561-570.

191. Parker J, Rambaut A and Pybus O. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect Genet Evol* 2008; **8**:239-246.

192. Mehta SR, Kosakovsky Pond SL, Young JA, Richman D, Little S and Smith DM. Associations Between Phylogenetic Clustering and HLA Profile Among HIV-Infected Individuals in San Diego, California. *J Infect Dis* 2012; **205(10)**:1529-1533.

193. EuroHIV: European Centre for the Epidemiological Monitoring of AIDS, UNAIDS. HIV/ AIDS Surveillance in Europe. In: 1999.

194. Csete J, Grob PJ. Switzerland, HIV and the power of pragmatism: lessons for drug policy development. *Int J Drug Policy* 2012; **23(1)**:82-86.

195. Nationales Programm HIV und ande-re sexuell übertragbare Infektionen2011-2017 (NPHS). HIV- und STI-Fallzahlen 2012: Berichterstattung, Analysen und Trends. In: 2013.

196. Alcantara LC, Cassol S, Libin P, Deforche K, Pybus OG, Van RM, Galvao-Castro B, Vandamme AM and de Oliveira T. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res* 2009; **37(Web Server issue)**:W634-W642.

197. Ragonnet-Cronin M, Hodcroft E, Hue S, Fearnhill E, Delpech V, Leigh Brown AJ and Lycett S. Automated analysis of phylogenetic clusters. *BMC Bioinformatics* 2013; **14(1)**:317.

198. Aghaizu A, Brown AE, Nardone A, Gill ON, Delpech V. HIV in the United Kingdom: 2013 Report. In. Public Health England, London: Health Protection Services; 2013.

199. Carnegie NB, Wang R, Novitsky V and De G, V. Linkage of viral sequences among HIV-infected village residents in Botswana: estimation of linkage rates in the presence of missing data. *PLoS Comput Biol* 2014; **10(1)**:e1003430.

200. Neogi U, Haggblom A, Santacatterina M, Bratt G, Gisslen M, Albert J and Sonnerborg A. Temporal trends in the Swedish HIV-1 epidemic: increase in non-B subtypes and recombinant forms over three decades. *PLoS ONE* 2014; **9(6)**:e99390.

201. Abecasis AB, Wensing AM, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DA, Albert J, Asjo B, Balotta C, Beshkov D, Camacho RJ, Clotet B, De GC, Griskevicius A, *et al.* HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* 2013; **10**:7.

202. Parry JV, Murphy G, Barlow KL, Lewis K, Rogers PA, Belda FJ, Nicoll A, McGarrigle C, Cliffe S, Mortimer PP and Clewley JP. National surveillance of HIV-1 subtypes for England and Wales: design, methods, and initial findings. *J Acquir Immune Defic Syndr* 2001; **26(4)**:381-388.

203. Chaix ML, Seng R, Frange P, Tran L, Avettand-Fenoel V, Ghosn J, Reynes J, Yazdanpanah Y, Raffi F, Goujard C, Rouzioux C and Meyer L. Increasing HIV-1 non-B subtype primary infections in patients in France and effect of HIV subtypes on virological and immunological responses to combined antiretroviral therapy. *Clin Infect Dis* 2013; **56(6)**:880-887.

204. Snoeck J, Van LK, Hermans P, Van WE, Derdelinckx I, Schrooten Y, Van de Vijver DA, De WS, Clumeck N and Vandamme AM. Rising prevalence of HIV-1 non-B subtypes in Belgium: 1983-2001. *J Acquir Immune Defic Syndr* 2004; **35(3)**:279-285.

205. Pyne MT, Hackett J, Jr., Holzmayer V and Hillyard DR. Large-scale analysis of the prevalence and geographic distribution of HIV-1 non-B variants in the United States. *J Clin Microbiol* 2013; **51(8)**:2662-2669.

206. UNAIDS. HIV Prevention Toolkit. In: 2008.

207. Tatt ID, Barlow KL, Clewley JP, Gill ON and Parry JV. Surveillance of HIV-1 subtypes among heterosexuals in England and Wales, 1997-2000. *J Acquir Immune Defic Syndr* 2004; **36(5)**:1092-1099.

208. Fox J, Castro H, Kaye S, McClure M, Weber JN and Fidler S. Epidemiology of non-B clade forms of HIV-1 in men who have sex with men in the UK. *AIDS* 2010; **24(15)**:2397-2401.

209. Gifford RJ, de Oliveira T, Rambaut A, Pybus OG, Dunn D, Vandamme AM, Kellam P and Pillay D. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. *J Virol* 2007; **81(23)**:13050-13056.

210. Barin F, Meyer L, Lancar R, Deveau C, Gharib M, Laporte A, Desenclos JC and Costagliola D. Development and validation of an immunoassay for identification of recent human immunodeficiency virus type 1 infections and its use on dried serum spots. *J Clin Microbiol* 2005; **43(9)**:4441-4447.

211. Ragonnet-Cronin M, Ofner-Agostini M, Merks H, Pilon R, Rekart M, Archibald CP, Sandstrom PA and Brooks JI. Longitudinal phylogenetic surveillance identifies distinct patterns of cluster dynamics. *J Acquir Immune Defic Syndr* 2010; **55(1)**:102-108.

212. Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 1997; **13(3)**:235-238.

213. Zeileis A, Kleiber C and Jackman S. Regression models for count data in R. *Journal of Statistical Software* 2008; **27(8)**:1-25.

214. Armitage P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* 1955; **11(3)**:375-386.

215. Cochran WG. Some methods for strengthening the common X^2 tests. *Biometrics* 1954; **10**:417-451.

216. Shao Y, Williamson C. The HIV-1 epidemic: low- to middle-income countries. *Cold Spring Harb Perspect Med* 2012; **2(3)**:a007187.

217. Stimson GV, Hunter GM, Donoghoe MC, Rhodes T, Parry JV and Chalmers CP. HIV-1 prevalence in community-wide samples of injecting drug users in London, 1990-1993. *AIDS* 1996; **10(6)**:657-666.

218. Xiridou M, van VM, Coutinho R and Prins M. Can migrants from high-endemic countries cause new HIV outbreaks among heterosexuals in low-endemic countries? *AIDS* 2010; **24(13)**:2081-2088.

219. Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, Nelson MI, Cummings DA, Bwanika JB, Mueller AC, Reynolds SJ, Munshaw S, Ray SC, Lutalo T, *et al.* The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. *PLoS Med* 2014; **11(3)**:e1001610.

220. Johnson AM, Wadsworth J, Wellings K, Bradshaw S and Field J. Sexual lifestyles and HIV risk. *Nature* 1992; **360(6403)**:410-412.

221. Anderson RM, May RM. Social heterogeneity and sexually transmitted diseases. In: *Infectious Diseases of Humans: Dynamics and Control*: Oxford Science Publications; 1991. p. 270.

222. Butts CT. Social network analysis with sna. *J Stat Softw* 2008; **24(6)**.

223. Dennis AM, Herbeck JT, Leigh Brown AJ, Kellam P, de OT, Pillay D, Fraser C and Cohen MS. Phylogenetic studies of transmission dynamics in generalized HIV epidemics: an essential tool where the burden is greatest? *J Acquir Immune Defic Syndr* 2014; **67(2)**:181-195.

224. Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, Van WE, Vandamme AM, Van LK and Lemey P. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLoS Comput Biol* 2014; **10(4)**:e1003505.

225. Dennis AM, Hue S, Hurt CB, Napravnik S, Sebastian J, Pillay D and Eron JJ. Phylogenetic insights into regional HIV transmission. *AIDS* 2012; **26(14)**:1813-1822.

226. Ndiaye HD, Toure-Kane C, Vidal N, Niama FR, Niang-Diallo PA, Dieye T, Gaye-Diallo A, Wade AS, Peeters M and Mboup S. Surprisingly high prevalence of subtype C and specific HIV-1 subtype/CRF distribution in men having sex with men in Senegal. *J Acquir Immune Defic Syndr* 2009; **52(2)**:249-252.

227. Tovanabutra S, Sanders EJ, Graham SM, Mwangome M, Peshu N, McClelland RS, Muhaari A, Crossler J, Price MA, Gilmour J, Michael NL and McCutchan FM. Evaluation of HIV type 1 strains in men having sex with men and in female sex workers in Mombasa, Kenya. *AIDS Res Hum Retroviruses* 2010; **26(2)**:123-131.

228. Volz EM, Frost SD. Inferring the source of transmission with phylogenetic data. *PLoS Comput Biol* 2013; **9(12)**:e1003397.

229. Punyacharoensin N, Edmunds WJ, De AD, Delpech V, Hart G, Elford J, Brown A, Gill N and White RG. Modelling the HIV epidemic among MSM in the United Kingdom: quantifying the contributions to HIV transmission to better inform prevention initiatives. *AIDS* 2015; **29(3)**:339-349.

230. Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, May S and Smith DM. Using HIV networks to inform real time prevention interventions. *PLoS ONE* 2014; **9(6)**:e98443.

231. Wertheim JO, Kosakovsky Pond SL, Little SJ and De G, V. Using HIV transmission networks to investigate community effects in HIV prevention trials. *PLoS ONE* 2011; **6(11)**:e27775.

232. Wang X, Wu Y, Mao L, Xia W, Zhang W, Dai L, Mehta SR, Wertheim JO, Dong X, Zhang T, Wu H and Smith DM. Targeting HIV Prevention Based on Molecular Epidemiology Among Deeply Sampled Subnetworks of Men Who Have Sex With Men. *Clin Infect Dis* 2015.

233. Brooks JI, Sandstrom PA. The power and pitfalls of HIV phylogenetics in public health. *Can J Public Health* 2013; **104(4)**:e348-e350.

234. O'Dea EB, Wilke CO. Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees. *Interdiscip Perspect Infect Dis* 2011; **2011**:238743.

235. kenah E, Britton T, Halloran ME and Longini IM. Algorithms linking phylogenetic and transmission trees for molecular infectious disease epidemiology. *arXiv* 2015.

236. Morelli MJ, Thebaud G, Chadoeuf J, King DP, Haydon DT and Soubeyrand S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol* 2012; **8(11)**:e1002768.

237. Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP and Haydon DT. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc Biol Sci* 2008; **275(1637)**:887-895.

238. Hall MD, Woolhouse M and Rambaut A. Using genomics data to reconstruct transmission trees during disease outbreaks. *(submitted)* 2015.

239. Rachinger A, Groeneveld PH, van AS, Lemey P and Schuitemaker H. Time-measured phylogenies of gag, pol and env sequence data reveal the direction and time interval of HIV-1 transmission. *AIDS* 2011; **25(8)**:1035-1039.

240. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C and Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol* 2014; **10(1)**:e1003457.

241. Pillay D, Herbeck J, Cohen MS, de Oliveira T, Fraser C, Ratmann O, Leigh Brown AJ and Kellam P. PANGEA-HIV: phylogenetics for generalised epidemics in Africa. *Lancet Infect Dis* 2015; **15(3)**:259-261.

242. Lemey P, Rambaut A and Pybus OG. HIV evolutionary dynamics within and among hosts. *AIDS Rev* 2006; **8(3)**:125-140.

243. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, *et al.* Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 2011; **364(8)**:730-739.