Estimating the Heritability of Virulence in HIV

Emma B. Hodcroft

This thesis is submitted for the degree of Doctor of Philosophy at the Institute of Evolutionary Biology, University of Edinburgh

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.

Signature of Candidate Date

Chapter 2

The R script for stripping codon sites associated with drug-resistance mutations was provided by Manon Ragonnet-Cronin. Dr. Jarrod Hadfield provided R code for using MCMCglmm to obtain inverse genetic matrices and pedigrees. JH also provided R code to analyse the change in viral load over time due to within-host and between-lineage selection, and ran simulations to assess the sensitivity of the method (see Chapter 2.7).

Chapter 5

Dr. Samantha Lycett provided the base code for the Discrete Spatial Phylo Simulator, as detailed in Chapter 5.2. Matthew Hall provided the VirusTreeSimulator program. Dr. Jarrod Hadfield provided equations to generate new viral loads dependant on a heritability value.

Abstract

The rate that HIV-infected individuals progress to AIDS and death varies greatly. Viral load taken during the asymptomatic phase of the disease is one of the best-known predictors of HIV progression rate and transmission risk, and is known to be influenced by both host and environmental factors. However, the role that the virus itself plays in determining the viral load is less clear. Previous studies have attempted to quantify the amount the viral genome influences viral load, or the heritability of viral load, using transmission pairs and phylogenetic signal in small sample sizes, but have produced highly disparate estimates.

Efficient and accurate methods to estimate heritability have been utilised by quantitative geneticists for years, but are rarely applied to non-pedigree data. Here, I present a novel application of a population-scale method based in quantitative genetics to estimate the heritability of viral load in HIV using a viral phylogeny. This new phylogenetic method allows the inclusion of more samples than ever previously used, and avoids confounding effects associated with transmission pair studies.

This new method was applied to the two largest HIV subtypes found in the UK, subtypes B and C, using sequences and clinical data from UK-wide HIV databases. For subtype B (n=8,483) and C (n=1,821), I estimated that 5.7% (CI 2.8–8.6%) and 29.7% (CI 14.8–44.7%) of the variance in viral load is determined by the viral genome, respectively. These estimates suggest that viral influence on viral load varies greatly between subtypes, with subtype C having much larger viral control over viral load than subtype B. I expanded the phylogenetic method to test whether the component of the viral load determined by the virus has changed over time. In subtype B, I found

evidence of a small but significant decrease in the viral component of viral load of -0.05 $\log_{10}$ copies/mL/yr.

I built a stochastic, individual-based model capable of simulating a realistic HIV epidemic, with heritable viral loads that influence transmission and disease progression, capable of generating data sets to assess the accuracy of phylogenetic methods. This was successfully used to generate epidemics approximating those in a small African village and a Western 'men who have sex with men' community under a variety of conditions. To test the accuracy of the new phylogenetic heritability estimation method, simulated datasets were generated with the heritability of viral load set at values of 30%, 50%, 70%, and 90%. Unfortunately, complications in the heritability equation used prevented full assessment of the new phylogenetic method on the simulated data. Future development of the model will enable simulation of realistic viral loads under varying heritability values, enabling simulation of data sets that can be used to test this and other heritability estimation methods.

This new phylogenetic method allows accurate estimation of heritability in large datasets, and has provided valuable insight into the viral influence on viral load in HIV.

Lay Summary

Human Immunodeficiency Virus, or HIV, is a virus that attacks and destroys the immune system, leaving you unable to defend yourself against infection. Though people infected with HIV can live many years without symptoms, the immune system is eventually so damaged that you become fatally ill, and are considered to have Acquired Immunodeficiency Syndrome, or AIDS.

However, there is a huge variation in how quickly people with HIV progress to AIDS. Some people get HIV in as little as two to three years, whereas others can remain AIDS free for over ten years. It's possible to estimate how quickly someone with HIV will progress to AIDS by measuring the amount of virus they have in their blood, known as the 'viral load.' People with lower viral loads have a slower disease progression and a longer time until they get AIDS. Those with a higher viral load progress more quickly, and get AIDS sooner.

The aim of my research is to identify what is responsible for the large variation in viral loads between individuals, which lead to such big differences in disease progression. This could be due to 'environmental' effects, such as diet and health, because of the effects of a particular person, such as their age, ethnicity, and immune system – or it could be due to the virus itself. The main question in my research is: do some strains of HIV make people sick more quickly than others?

This is an important question, because if the virus has a strong control over viral load, then it's possible that a more dangerous strain of HIV could appear, and spread in the population. In my PhD, I have investigated this question by developing a brand new method that allows me to use more data than any previous study.

Family trees allow us to see how everyone is related, and I have used a 'family tree' of many HIV sequences from individuals across the UK to see how all of the HIV viruses are related. Then, I can look at how similar or different the viral loads are in viruses that are more closely or distantly related, and estimate exactly what effect each type of virus has on the viral load.

Using this new method, I've estimated that one type of HIV controls only 6% of the viral load, and another controls about 30% of the viral load. This means that the majority of viral load is controlled by other things, like health, ethnicity, and immune system. And though it is possible that HIV could be evolving to become more dangerous, my analysis showed that viral load has actually gotten slightly lower in the last few years.

Finally, I have developed a new program that allows me to create realistic HIV epidemics on a computer. This program will let me, and other scientists, test my new method on data where we know the answer already, to ensure it works well.

Publications

The following papers have been published during my PhD:

Hodcroft, E., Hadfield, J. D., Fearnhill, E., Phillips, A., Dunn, D., O'Shea, S., Pillay, D., and Leigh Brown, A. J. (2014). The contribution of viral genotype to plasma viral set-point in HIV infection. PLoS pathogens, 10(5), e1004112.

This is the paper primarily arising from work performed in this thesis, and includes the methods and results detailed in Chapters 2 and 3.

Ragonnet-Cronin, M., **Hodcroft, E.**, Hu, S., Fearnhill, E., Delpech, V., Brown, A. J., and Lycett, S. (2013). Automated analysis of phylogenetic clusters. BMC bioinformatics, 14(1), 317.

My contribution to this paper (the ClusterMatcher program) was as a result of experience gained manipulating phylogenetic trees in Java on my PhD.

In total, I have made three Java programs publicly available as a result of work performed during my PhD:

1. TreeCollapseCL 4 (see Chapter 2.5.3) - *Available at www.emmahodcroft.com*
2. ClusterMatcher - *Available at hiv.bio.ed.ac.uk*
3. PareTree - *Available at www.emmahodcroft.com*

From the 17th February to the 28th March 2015, the webpages containing TreeCollapserCL and PareTree received visits from 31 different universities and research centres around the world.

Acknowledgements

First and foremost, I would like to thank the UK HIV Drug Resistance Database (UK HIV DRB) and the UK Clinical HIV Cohort (UK CHIC), for granting me access to one of the most comprehensive HIV data sets in the world. Without this data, none of my research would have been possible. In this same vein, I would also like to thank the UK HIV DRB steering committee, the multitude of centres who send in their samples, and the many clinician, technicians, and analysts that make the UK HIV DRB possible. In particular, I would like to thank Esther Fearnhill, Dr. David Dunn, Prof. Deenan Pillay, and Prof. Andrew Phillips for their invaluable insight into the UK HIV DRB, and their comments on and interest in my work.

Another sweeping thank-you goes to all of my colleagues in Ashworth Laboratories. I am so very lucky to have been surrounded by enthusiastic, interesting, friendly, and encouraging people in an environment where the most eminent professors share lunch with the most timid students. The 'environmental effect' of Ashworth on my experience as a student, and as a researcher, has been significant indeed, and will never be forgotten.

To fund PhD students is to invest in the future. I owe thanks to the Biotechnology and Biological Sciences Research Council and the Bill and Melinda Gates Foundation for investing in me, by funding my PhD research.

My love of biology is in no small way due to a long line of excellent teachers and professors. Mr. Harmon, Mr. Thornton, and Dr. Ben Pierce, thank you for nurturing the curiosity of a young girl so long ago. It was during my first class on evolution with Dr. John Horner that I realised that studying evolution was what I wanted to do. My

first forays into research, first conference, and first manuscript were all in his lab, and it was with his encouragement that I decided to pursue my MSc and PhD. Thank you Dr. Horner, for inspiring me to follow this wonderful path.

I have had the rare and wonderful privilege of doing my PhD among the most incredible peers. In my year as a research assistant I fell in with the PhD year above me, and am so glad to have been an 'honorary' member of their cohort. Sarah Mattey, Kay Boulton, Hannah Froy, Lucy Carter, Kevin Donnelly, Reuben Nowell, Mojca Zelnikar, Emily Moore, Elisa Schaum, and Julja Ernst, I am so pleased to have spent 4.5 years among you. My own 'official' PhD cohort is no less magnificent, and I wish to thank Manon Ragonnet, Jess Flood, Rebecca Callaway, Matthew Hall, Tom Godfrey, Lu Lu, Sam Lewis, Elisa Anastasi, Richard Allen, and Gytis Dudas for the many lunches, laughs, happy hours, and coffee breaks we have shared.

Doing a PhD is an adventure in success and failure unlike any other. To stay grounded in the highs, and mentally stable in the lows, there is no better resource than a close group of friends whom you can count on and confide in. To Kay Boulton, Sarah Mattey, Sam Lewis, Hannah Froy, Lucy Carter, Kevin Donnelly, Reuben Nowell, Elisa Schaum, Manon Ragonnet, Jess Flood, and Elisa Anastasi: thank you for all the dinners, the pub trips, the holidays, the practice talks, the funny emails, the Christmas parties, the vivas, the weddings, the birthday cards, and all the memories.

I would like to thank Kevin Donnelly and Bryony Jackson for always being up for an Saturday-night adventure, whether not starving together, archery in a dungeon, or smuggling criminals out of an asylum. I'll see you in Monaco.

There are a few individual people whom I also wish to thank more intimately. I have had the pleasure of working alongside many esteemed researchers, including Jess Hedge, Trevor Bedford, and Paul Wikramaratna. Gonzalo Yebra has shared not only an office with me, but advice, comments, and even a seemingly endless trek around down-town Seattle desperately trying to find dinner! I first met Melissa Ward almost six years ago, when I came to interview for an MSc at Edinburgh. Little did I realise then that she would become a colleague, friend, and advisor, to whom I could always go for help and reassurance about every step of the PhD process.

I have had the enormous privilege of working alongside Samantha Lycett, a truly

incredible scientist and programmer. I owe a debt of gratitude to Sam for her patience and help when I first arrived, and her continuing advice and inspiration. I also must thank my second supervisors: Andrew Rambaut, for his support and insightful comments, and Jarrod Hadfield, for his endless patience in helping me understand ASReml and MCMCglmm, and his continued advice on heritability simulation.

I was very lucky to have two excellent examiners in my viva, Loeske Kruuk and Angela McLean. They made the processes challenging, interesting, and enjoyable, and I am grateful for their excellent suggestions and thought-provoking questions.

A PhD is traditionally a uniquely lonesome quest into the unknown. I have had the rare privilege of sharing my journey with my PhD 'sister,' Manon Ragonnet. Meetings, holidays, conferences, meals, flights, trains, publications, hotel rooms, parties, hikes, code, and frustrations – we have shared them all. I am so glad we have two more years together to continue all our adventures, in research and in life!

I owe so very much to my supervisor, Prof. Andrew Leigh Brown. Under his guidance, I have grown from a near-silent, half-terrified research assistant into the independent, self-assured scientist I am today. He has nurtured my enthusiasm and love of programming, resulting in some of the most interesting aspects of my research, and leading to exciting future opportunities. His patience, understanding, and encouragement have been the bedrock of my research. Thank you, Andy.

Words are inadequate to express my gratitude to Darren Parker, my partner in crime, science, and life. Throughout the best times and worst, his tireless love and support have kept me going. Thank you, Darren, for keeping me focused, for never faltering in your faith in me, and for always listening to my many stories about Mendel, which are actually, very interesting, actually.

Finally, I would like to thank my family for their love, and for never doubting my ability, even when I did. In particular, I owe everything to my mother, Ellen Boyer, and my father, Ken Hodcroft. Without their support and endless encouragement, I would never have become the person, and scientist, I am today.

"Aids seems to have been around only since about 1978. By September 22^{nd} the Centres for Disease Control had counted 608 cases of it in America; there are about two new cases each day. Of the 608 reported, 249 people (41%) had died. Even this paints too cheerful a picture: of victims diagnosed before June, 1981, 70% had died by September this year. The reason is simple. No victim has been found whose immune system has recovered after an attack of Aids. Sooner or later sufferers fall prey to an infection that drugs cannot cure.

[...]

"*There has never been anything quite like this, and researchers are at a loss to explain it.* Most doctors studying the disease think it is caused by a new virus, or a new type of old virus, and that it is probably sexually transmitted.

[...]

"**Some good may come of all this ... Aids research will likely turn up something new about the way viruses work.** Aids may well disclose something about connections between immune system failures and cancer. Meanwhile, people who used not to worry about the consequences of a varied sex life now do."

– *The Economist* 'Next They'll Tax It' (October 1982)

Contents

List of Figures

List of Tables

"We live in a completely interdependent world, which simply means we cannot escape each other. How we respond to AIDS depends, in part, on whether we understand this interdependence. It is not someone else's problem. This is everybody's problem."

Bill Clinton

1 Introduction

1.1 Introduction to HIV

1.1.1 Discovery and Characteristics

AIDS (Acquired Immune Deficiency Syndrome) was first identified in the USA in 1981, when otherwise healthy young men began falling ill with rare opportunistic infections, usually only found in immunocompromised and elderly individuals (CDC, 1981; Greene, 2007). The cause of AIDS was identified with the discovery of a new retrovirus (Barré-Sinoussi et al., 1983; Gallo et al., 1984) named 'Human Immunodeficiency Virus' (HIV) (Coffin et al., 1986).

HIV is a lentivirus – a single-stranded, positive-sense, enveloped RNA virus. It primarily replicates by entering helper T-lymphocytes, a key immune system cell, using the CD4 receptor molecule (Klatzmann et al., 1984) on the cell surface. As a member of the family *Retroviridae*, HIV then reverse-transcribes its genetic material before inserting it into the host's own cellular DNA (integration). Once integrated, the viral DNA may lie dormant for years, leading to a 'latency' in the infection (see Section 1.2 on page 7 for more detail on the stages of HIV infection), or may be transcribed to mRNA immediately. Transcription can only occur when certain cellular transcription factors are present, many of which are produced when the T-lymphocyte is activated (Nabel and Baltimore, 1987; Siekevitz et al., 1987), and initially produces only short

viral regulatory proteins, which aid in the complete transcription and transportation of the mRNA. The HIV mRNA is transported to the cytoplasm, where further regulatory and structural proteins are translated and packaged along with the mRNA (the viral genome) into a viral particle (Sundquist and Kräusslich, 2012).

HIV's genome is about 10,000 base pairs, and encodes nine proteins: *gag, pol, env, tat, rev, nef, vpr, vif,* and *vpu* (Figure 1.1). HIV has a very high mutation rate of around 3.5×10^{-5} mutations per base per replication cycle (Mansky and Temin, 1995), due to the error-prone transcription of the reverse-transcriptase protein (Preston et al., 1988) and the lack of a 'proofreading' mechanism (Steinhauer et al., 1992). This high mutation rate enables the virus to continuously evolve to escape the host immune system and develop resistance to anti-retroviral treatment (ART). While the high mutation rate of HIV is one of the reasons why the virus has proven so difficult to treat, it also generates enough genetic diversity between viruses infecting different hosts, and viruses sampled at different time periods, that it is possible to reconstruct the spread of HIV between individuals and across populations using phylogenetic techniques, even over relatively short periods of time (Ou et al., 1992; Holmes et al., 1993). Many studies investigating transmissions between long-term partners have taken advantage of this to confirm that transmission was indeed from the long-term partner and not an external contact. The genetic distance between sequences from the transmitting individual and the newly-infected partner can be compared to the genetic diversity in the community at large to see if the sequences are similar, and phylogenetic analysis can be used to ascertain how close the sequences are placed to each other on a tree constructed with sequences from the study (Trask et al., 2002).

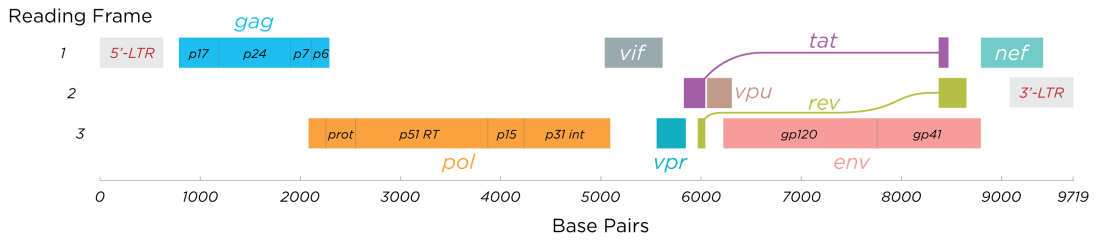

Figure 1.1: Structure of the HIV-1 genome *(CC BY-SA Thomas Splettstoesser www.scistyle.com)*

1.1.2 History of HIV

HIV is thought to have been transmitted to humans multiple times from different strains of Simian Immunodeficiency Virus (SIV), a lentivirus that infects a large number of African primate species (Klatt et al., 2012). HIV type 2 (HIV-2) is antigenically distinct from the first HIV viruses isolated (HIV type 1 or HIV-1; Figure 1.2), and is believed to have originated from west African Sooty Mangabey monkeys (Gao et al., 1992, 1994; Chen et al., 1997). HIV-1 is divided into four major groups: M, N, and O. and P (Figure 1.2). Group P is most recently discovered and least common of the HIV-1 groups, having been identified in only two individuals (Plantier et al., 2009; Vallari et al., 2011), and is thought to have originated from gorillas (Van Heuverswyn et al., 2006; Plantier et al., 2009), while HIV-1 groups, M, N, and O represent separate introductions from chimpanzees (Sharp et al., 2001). Within group M – the most common in the HIV epidemic, thought to account for >90% of HIV infections (Spira et al., 2003) – HIV has further diversified, and is classified into subtypes (see Figure 1.2). Recombination between these subtypes is not uncommon, and some recombinants have been formally identified and classified as 'circulating recombinant forms' (CRFs).

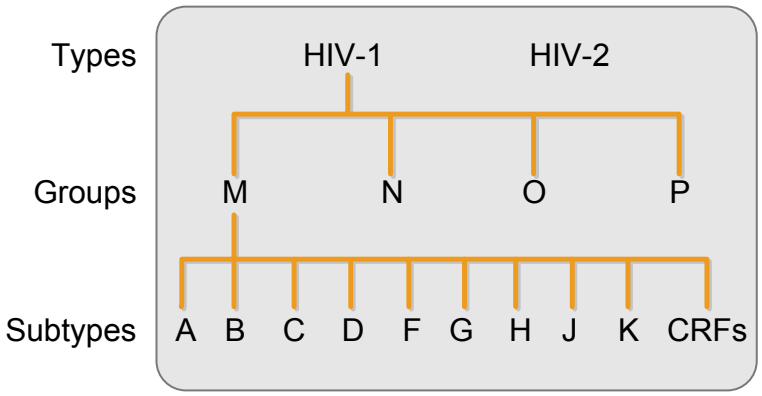

Figure 1.2: HIV is split into types HIV-1 and HIV-2, which are antigenically distinct. HIV-1 is divided into groups M, N, O, and P, each representing a different introduction to humans. Group M, the most common, is further divided into subtypes.

The ancestor for group M dates back to between 1910 and 1930 and is thought to have been located in the modern-day Democratic Republic of Congo (DRC) (Korber et al., 2000; Worobey et al., 2008; Faria et al., 2014). Sequences from Kinshasa in the late 1950's and early 1960's show that group M had already diversified into subtypes

that are now found across the globe (Worobey et al., 2008), and recent phylogenetic analysis suggests that a tripling in infection rate around this time may have been the start of the HIV pandemic (Faria et al., 2014).

As subtypes B and C are the most prevalent in HIV-infected individuals in the UK, they are the focus of my analysis. The histories of the two subtypes, and their arrival into the UK differs significantly, and so is outlined here.

Subtype B

Subtype B originated in Kinshasa before the mid-1940's, and is thought to have been spread to Haiti by infected Haitians who were employed in the DRC in the 1960's (Gilbert et al., 2007; Kuyu, 2008; Faria et al., 2014). Haiti has the oldest HIV epidemic outside of Africa, and the most genetically diverse subtype B clade, suggesting that the virus spread there for years before a single variant was transmitted to the U.S., and subsequently to the rest of the world, in the late 1960's or early 1970's (Gilbert et al., 2007).

Subtype B seems to have circulated within the U.S. for approximately twelve years before the first reported AIDS cases were recognised in 1981 (Gottlieb et al., 1981; Gilbert et al., 2007). Retrospective detection of HIV antibodies in blood samples taken from 1978-1980 in communities of men having sex with men (MSM) suggest a prevalence of 5% in San Francisco and 7% in New York (Jaffe et al., 1985; Stevens et al., 1986). During this circulating time the first introduction to the UK was made in the mid-1970's (Leigh Brown et al., 1997), and through the 1980's, subtype B was subsequently introduced to the UK from the US numerous times (Hué et al., 2005).

The first case of AIDS in the UK was identified in 1981 (du Bois et al., 1981), and as in the US, almost all early cases of HIV reported in the UK were subtype B and associated with homosexual behaviour (Arnold et al., 1995; Clewley et al., 1996; Parry et al., 2001). Subtype B also quickly gained a foothold in the injection-drug-user (IDU) community, where needle-sharing was common. As early as 1985, 38% of blood samples from hospitalized IDUs in Edinburgh tested positive for HIV antibodies (Peutherer et al., 1985), and retrospective testing of samples taken in 1982-83 suggests prevalence as high as 51% in Edinburgh IDUs (Robertson et al., 1986). In 1997, an

analysis of 621 HIV samples from across the UK showed 99.16% of samples from MSM and 75% of samples from all risk groups were subtype B (Parry et al., 2001), with non-B subtypes significantly associated with being born outside of the UK and Europe and heterosexual transmission. Despite the early dominance of subtype B across all risk groups, the proportion of subtype B infections in heterosexual individuals dropped from 40% in 1997 (Parry et al., 2001) to 25% by 2004 (Tatt et al., 2004). Subtype B is still the most predominant subtype present in UK MSMs, however, and is found in 88.4% of new diagnoses in this group (UK Collaborative Group on HIV Drug Resistance, 2014). Though non-B subtypes have recently overtaken subtype B as the majority of new diagnoses in the UK (UK Collaborative Group on HIV Drug Resistance, 2014), subtype B is still the most prevalent subtype in HIV-infected individuals in the UK (Figure 1.3).

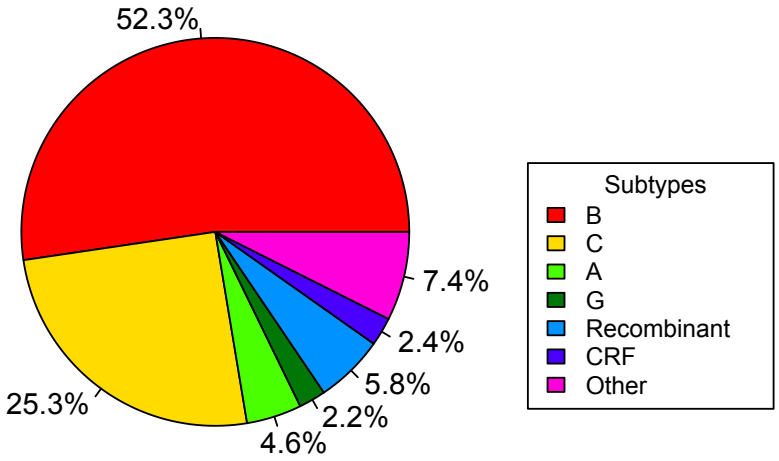

Figure 1.3: The distribution of subtypes found in the first sequence submitted in the UK HIV Drug Resistance Database, covering sequences submitted between 1997 and 2009. Subtype B is by far the most common subtype, followed by subtype C.

Subtype C and Non-B Subtypes in the UK

Subtype C seems to have diverged slightly earlier than subtype B, in the mining regions of the DRC (Faria et al., 2014). Early sequences from across southern and eastern Africa cluster with sequences from Lubumbashi, the capital of a southern region of the DRC, leading Faria et al. (2014) to conclude that migrant workers aided the early transport of subtype C across Africa. Subtype C is now thought to be responsible for around 50% of HIV infections worldwide (Hemelaar et al., 2006). Though such detailed analysis

has not yet been done for the other HIV subtypes, they too seem to have spread from the DRC across Africa, though less successfully than subtype C. Some subtypes are highly associated with certain regions; 80% of HIV infections in Eastern Europe and central Asia are subtype A, while 85% of infections in South and south-East Asia are CRF01_AE (Hemelaar et al., 2006). As with subtype B, this is likely largely due to a founder effect, where the first subtype introduced is the one now most common.

An analysis of *gag* sequences from 211 individuals taken in 1993-95 in Scotland, Ireland, and Northern England showed that all sampled sequences were of subtype B (Leigh Brown et al., 1997), but within 10 years of the first reported AIDs case in 1981 (du Bois et al., 1981) multiple non-B subtypes were being reported (Arnold et al., 1995). In conjunction with the rise of non-B subtypes, the number of HIV infections acquired by heterosexual contact rose steadily through the 1990's, eventually overtaking MSM as the risk group with the highest number of new HIV diagnoses (The UK Collaborative Group for HIV and STI Surveillance, 2004) (Figure 1.4 on the facing page), though in 2011 the MSM risk group again comprised the majority of new diagnoses (Aghaizu et al., 2013). Though the majority of samples taken from heterosexually-acquired infections were of subtype B through 1996-7 (Parry et al., 2001), only 25% of heterosexually-acquired infections were subtype B by 2000 (Tatt et al., 2004). This increase in heterosexually-acquired infection and non-B subtype infection in the late 1990's coincides with a sharp rise in the number of infections acquired in South-East Africa (Health Protection Agency et al., 2003; Sinka et al., 2003), quite possibly a result of the increase in the number of African-born immigrants to the UK during the same time period (Owen, 2009).

By 2000, 88% of new diagnoses in heterosexuals were thought to have been acquired abroad (Unlinked Anonymous Surveys Steering Group, 2002). Individuals infected with non-B subtypes are more likely to have been born outside of the UK and Europe compared to those infected with HIV subtype B (Parry et al., 2001). Most non-B subtype infections identified in Britain were probably acquired in Africa through heterosexual contact, or from a partner who was infected outside of Europe (Health Protection Agency et al., 2003; Tatt et al., 2004).

Unlike the UK subtype B epidemic, where most transmissions occur within the UK,

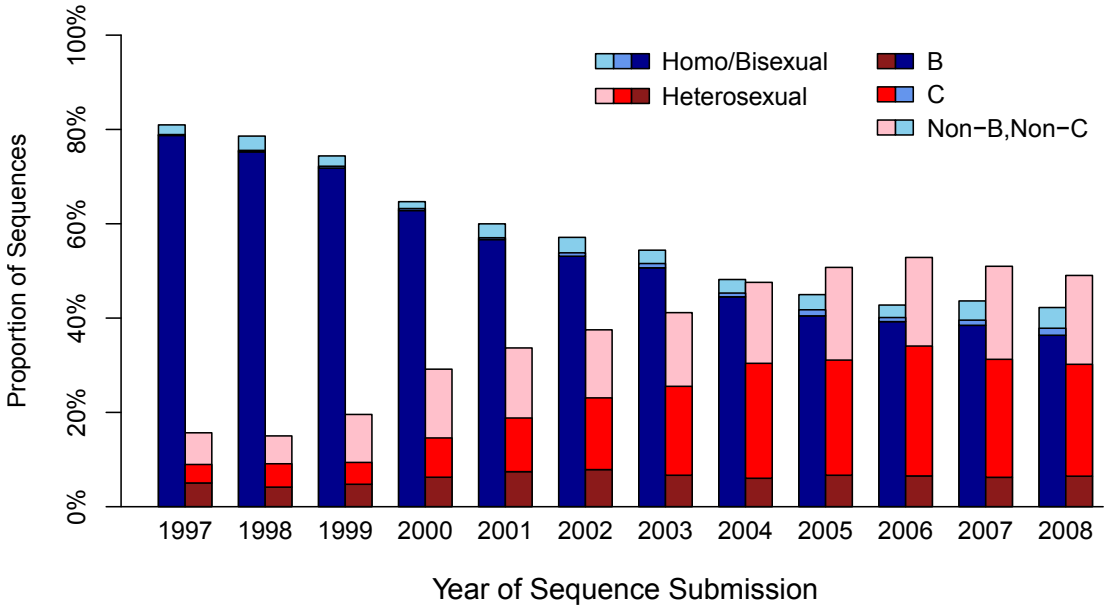

Figure 1.4: The proportion of first sequences submitted to the UK HIV Drug Resistance Database from patients in homo/bi-sexual (blues) and heterosexual (reds) risk groups, broken down by subtype. Though the vast majority of first sequences submitted where initially from patients infected through homo/bi-sexual contact, infections acquired through heterosexual contact have steadily risen.

until around 2010 non-B subtypes primarily had ongoing transmission from countries where these subtypes are more prevalent and seemed contained among heterosexual risk groups in the UK (Ragonnet-Cronin et al., 2013). Within the last few years, the number of heterosexual infections acquired abroad has fallen below the number acquired within the UK, marking a change in the UK heterosexual epidemic (Yin et al., 2014). Subtype C is now the most common non-B subtype in the UK. In 2008, the most recent year for which the 2010 UK HIV Drug Resistance Database (see Section 2.2.2 on page 30 for full details about the Database) release has full data, 30% of all sequences and 55% of non-B sequences submitted by new patients to the HIV DRB were subtype C (Figure 1.5 on the following page).

1.2 Prognostic Markers and Plasma Viral Load

The $CD4^+$ cells primarily targeted by HIV are killed both by cytotoxic T-lymphocytes recognising HIV epitopes presented on the cell surface, and when reproduction is unsuccessful but triggers cell death in the process (Doitsh et al., 2010). Decline in the

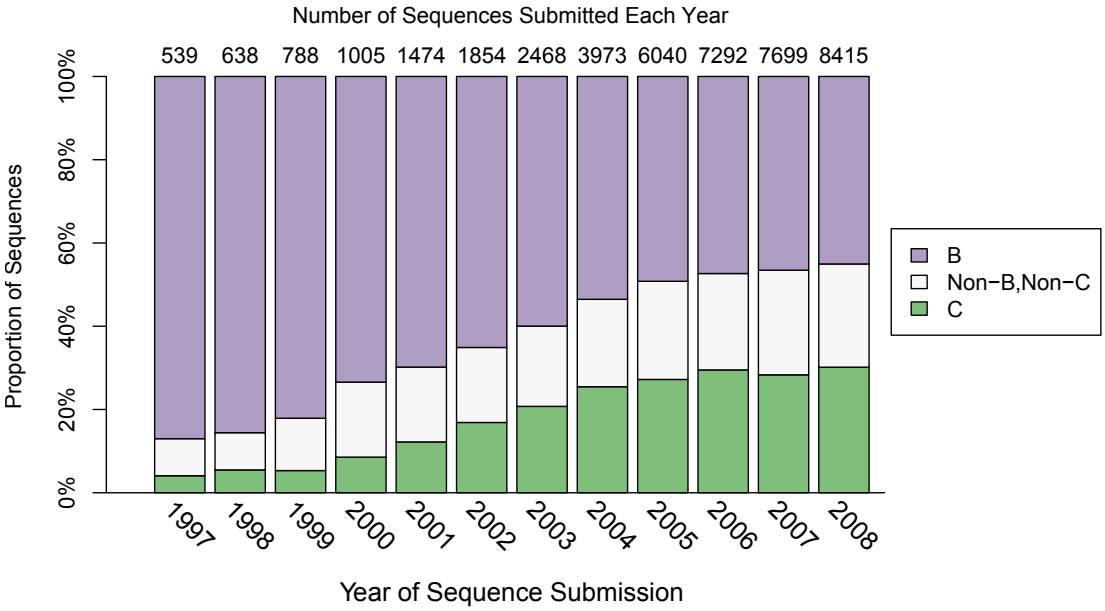

Figure 1.5: The proportion of subtype B (purple), C (green), and non-B/non-C (white) subtypes in the first sequence submitted to the UK HIV Drug Resistance Database, by year of submission. Subtype B sequences comprised the majority of first sequences submitted through the early 2000's, but by 2008 non-B subtypes now have a narrow majority of the sequences submitted that year.

number of $CD4^+$ cells in HIV-infected patients was first noticed in 1981 (Gottlieb et al., 1981), and marks the virus' slow destruction of the immune system that leaves the body vulnerable to opportunistic infections.

As research into the treatment and outcome of HIV infection progressed, the importance of prognostic markers became apparent. Plasma viral load, or the amount of HIV virus in the blood, has long been considered one of the most important clinical measures in HIV-positive patients (Mellors et al., 1995). Plasma viral load varies greatly through the course of HIV infection. In the weeks immediately after infection, known as the 'acute stage,' plasma viral load spikes and $CD4^+$ cell count dips as HIV begins replicating throughout the body, destroying immune cells (Figure 1.6 on the next page). During the acute stage, the patient may experience 'flu-like' symptoms and enlarged lymph nodes as the body begins to mount an immune response. As the immune system begins to develop antibodies to HIV, $CD4^+$ cell count rises, viral replication is limited, and viral load falls. This marks the beginning of the phase known as the 'chronic stage' or 'clinical latency' (Figure 1.6), during which patients are usually asymptomatic. The length of this phase is extremely variable, lasting from a few years to over twenty years

(Lackner et al., 2012). Though $CD4^+$ cells are slowly depleted during this stage, viral load has a very gentle rise (Sabin et al., 2000), and the viral load at the beginning of this period, initially considered as being one year after infection, is known as 'set-point' viral load. At the end of the chronic phase, the $CD4^+$ cell count has dropped to a point where the immune system can no longer fight off opportunistic infections or keep HIV replication in check (Figure 1.6). Viral load soars as $CD4^+$ count plummets, marking the onset of AIDS, and eventually, death.

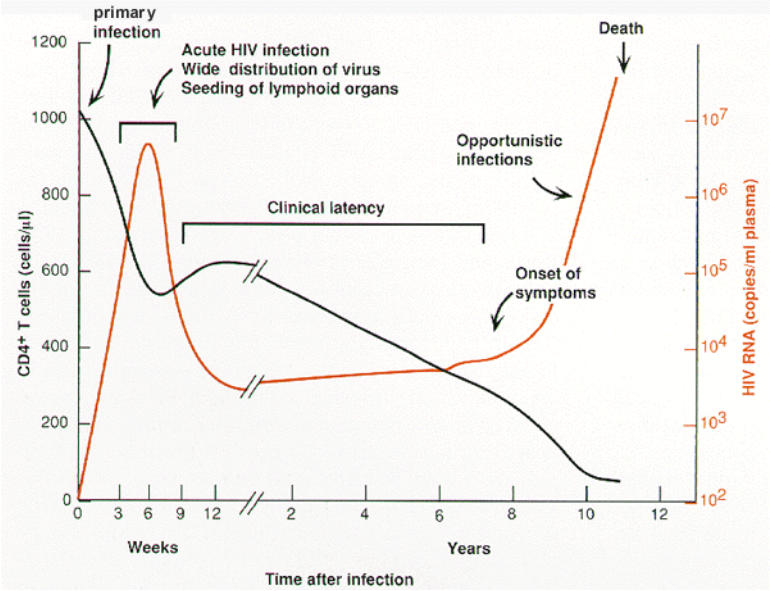

Figure 1.6: The change in viral load and $CD4^+$ count over the course of HIV infection. Figure highlights the acute and chronic or clinical latency stages, and the onset of symptoms that corresponds with the onset of AIDS (Fauci and Desrosiers, 1997).

Just as progression time from infection to AIDS or death varies enormously from a few years to decades, the 'set-point' viral load of an individual differs greatly between patients, and is the best known predictor of disease progression (Mellors et al., 1996; Fraser et al., 2007; Langford et al., 2007) and transmission risk (Quinn et al., 2000; Fideli et al., 2001) for an individual patient. Patients with a high viral load have a shorter chronic stage and progress to AIDS and death faster than those with a low viral load (Mellors et al., 1996) (Figure 1.7 on the following page). High viral load is also associated with a greater risk of transmitting HIV (Quinn et al., 2000; Fideli et al., 2001).

Because of the strong positive association between viral load and disease progression

and transmission risk, viral load is often used as a proxy for investigating 'virulence.' Virulence is usually understood to be the ability of a pathogen to cause disease and successfully transmit itself. As both of these things are linked to viral load in HIV, virulence has often been used to describe measures of one of these things ($CD4^+$ count, time to AIDS, and time to death as measures of disease progression; number of transmissions or time to transmission as measures of transmission risk), with the implication that it will affect the other. As ways of investigating HIV transmission and disease progression have changed as more clinical measures have become available, the measure being equated with 'virulence' in papers has also changed. When describing previous studies, I use 'virulence' to mean both disease progression and transmission risk, as both of these are closely linked to viral load. (For more discussion on the use of 'virulence,' particularly with regard to my own analyses, see Section 2.7 on page 56.) A better understanding of what influences viral load and thus virulence is vital to gaining insight into the dynamics of HIV.

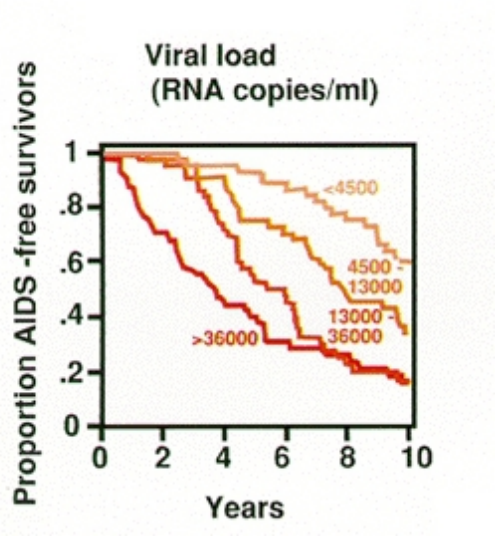

Figure 1.7: The relationship between viral load and length of AIDS-free survival. Those with higher viral loads (red) have a lower proportion of AIDS-free survivors over time than those with lower viral loads (orange). (Modified from Fauci and Desrosiers (1997), originally from Mellors et al. (1995))

1.2.1 Influences on Viral Load

Host genetic effects on viral load and HIV virulence have been of particular interest to researchers. Deletion of CCR5, a surface molecule that allows HIV to enter $CD4^+$ T-

cells (Deng et al., 1996), is found in Northern European populations (Martinson et al., 1997; Lockett et al., 1999) and prevents HIV from replicating, conferring a strong protective effect against HIV infection when homozygous (Huang et al., 1996; Smith et al., 1997). Research on human leukocyte antigen (HLA) types has also confirmed that variants with a range of effect sizes can be either protective or risk-increasing (Steel et al., 1988; Kaslow et al., 1990; O'Brien and Nelson, 2004; Tang et al., 2004; Fellay et al., 2009; Salgado et al., 2010).

Though many host genetic factors have been identified (reviewed in Telenti and Johnson (2012)), the viral genetic effect on virulence is still much less clear. However, assessing the viral influence on disease progression could have important implications for studying, preventing, and treating HIV. Identification of genetic variations associated with a higher viral load could be used to identify patients with more virulent virus strains, who could then be monitored more closely or targeted for treatment to prevent disease spread in populations where treating all HIV-positive individuals is not feasible. Importantly, if a proportion of the viral load variation is identified as being due to the virus' genetics, HIV set-point viral load could have the ability to evolve under natural selection.

Changes in Viral Load Over Time

Evolutionary theory predicts that pathogens evolve to modulate their density within hosts in order to maximize transmission rate. In the classic studies of myxomatosis (Fenner and Chapple, 1965), viral genotypes with reduced replication rate that permitted longer host survival were selected for when host density, and thus transmission probability, declined as the epidemic progressed. HIV is poorly transmissible relative to other human viruses such as measles or influenza, raising the possibility that in the 100 years HIV is known to have infected humans (Korber et al., 2000; Worobey et al., 2008; Faria et al., 2014), it might have adapted to different levels of transmission probability associated with different infected populations (Fraser et al., 2007; Wawer et al., 2005).

The hypothesis that HIV could be evolving and perhaps becoming more deadly has been a driver for decades of research into HIV virulence. As early as the mid-1980's, it became clear that some HIV isolates, deemed 'high/fast' lines, had a much higher

replicative capacity in cancer cell lines than others (Åsjö et al., 1986; Fenyo et al., 1988; Fiore et al., 1990). A few years later, Hutchinson et al. (1991) and Weiss et al. (1992) reported a decline in $CD4^+$ cell count at diagnosis, and speculation began as to whether the viral properties that produced 'high/fast' lines could be spreading and responsible for the drop (Weiss et al., 1992; Gorham et al., 1993; Holmberg et al., 1995). As larger cohort research became increasingly feasible and the collection of disease progression measures such as $CD4^+$ cell count and viral load became more commonplace, a number of studies looking at long-term trends in HIV virulence were published, drawing mixed conclusions on whether there was evidence of HIV becoming more virulent (Veugelers et al., 1994; O'Brien et al., 1995; Galai et al., 1996; Keet et al., 1996; Carré et al., 1997; Sinicco et al., 1997; Vanhems et al., 1999; Concerted Action on SeroConversion to AIDS and Death in Europe, 2000; CASCADE Collaboration, 2003; Dorrucci et al., 2005; Müller et al., 2006; Crum-Cianflone et al., 2009; Müller et al., 2009a). Lack of standardization of when measurements were taken, what measures were used, and whether patients were on anti-retroviral therapy (ART), as well as differences in the subtypes, risk groups, and demographics of the patients involved mean that these studies are difficult to directly compare. Despite this, two meta-analyses of the data have been performed, both concluding that a decrease in $CD4^+$ count and an increase in viral load can be observed over the last few decades, implying an increase in HIV virulence that both papers suggest could be caused by viral factors (Dorrucci et al., 2007; Herbeck et al., 2012).

In an epidemiological analysis, Fraser et al. (2007) studied the relationship between transmission risk, time to progress out of the asymptomatic phase, and viral load, confirming that a higher viral load lead to a higher chance of transmission during a risk activity but a quicker disease progression, shortening the time when transmissions are most likely. Thus these two virulence factors are negatively correlated and changes in the viral load are expected to force a 'trade-off' between one and the other (Fraser et al., 2007). Fraser et al. (2007) hypothesized that selection toward both a high per-event transmission risk and a long potential transmission period could lead to viral load evolving towards an 'optimal' value that maximises the overall transmission potential. Data from Amsterdam, Zambia, and Uganda was used in a parametric model to describe

the relationship between transmission risk, asymptomatic phase length, and viral load (Fraser et al., 2007). When the viral load values that maximized both transmission risk and asymptomatic phase length for early- and late-stage HIV epidemics were calculated from the model, these were found to match the mean viral loads from cohorts in early- and late-stage epidemics, suggesting that viral load might indeed be at an 'optimal value,' though this may or may not be due to viral adaptation (Fraser et al., 2007).

Shirreff et al. (2011) developed a mathematical model for Fraser et al. (2007) hypothesis to track the estimated dynamics of viral load over time, using parameter values from Fraser et al. (2007)'s analysis that describe the effect of viral load on transmission and infection duration and disease transmission parameters estimated from transmission partner studies (Wawer et al., 2005; Hollingsworth et al., 2008). Regardless of whether the model was started with a high or low initial viral load, the viral load converged to an intermediate value close to that predicted by Fraser et al. (2007), with this convergence taking place in the same length of time that HIV is estimated to have been in the human population (Shirreff et al., 2011).

However, Fraser et al. (2007)'s theory is dependent on the presence and influence of viral factors that affect virulence. If viral factors are present and found to play a large role in determining set-point viral load and thus transmission potential, selection could indeed be playing a strong role in pushing viral load to an 'optimal' value, but if viral factors only have a weak influence on viral load, selection will be much less efficient.

Evidence that virulence is influenced by major viral genetic differences across the HIV phylogeny seems strong. Early studies identified virulence differences linked to HIV type and group, with HIV-1 group M being shown as more virulent than HIV-2 and HIV-1 group O (Pepin et al., 1991; Marlink et al., 1994; Whittle et al., 1994; Arien et al., 2005). Comparing differences in disease progression and viral load between the 7 subtypes of HIV-1 group M has proven more difficult, as often only one subtype is prevalent in a demographic, ethnic, or risk group, making comparative studies problematic.

Kanki et al. (1999) identified a population of female sex workers in Senegal where subtype A was prevalent, but subtype C, D, and G were also present. The time from HIV-infection until AIDS or AIDS-related death was tracked in 53 such women, and

those infected with subtype A viruses were found to have a longer AIDS-free survival period than those with non-A subtypes.

Uganda has a high prevalence of both subtypes A and D coexisting in the same populations, which has been taken advantage of by two studies. Kaleebu et al. (2001) studied 1,045 HIV-positive individuals who were subtyped as A (51%) or D (49%) from Entebbe, Uganda, measuring time to death. Subtype D was associated with a significantly faster progression to death, with a relative risk of 1.29 compared to subtype A (Kaleebu et al., 2001). A later study by Kiwanuka et al. (2008) found similar results when investigating disease progression in 350 HIV-positive individuals from Rakai, Uganda, infected with subtypes D, A, and C, as well as recombinant subtypes and multiple subtypes. The study found that those infected with non-A subtypes had a significantly higher risk of progression to AIDS and death than those with subtype A (Kiwanuka et al., 2008). One study has suggested the difference in progression between subtypes could be due to differences in co-receptor use, finding subtype D viruses were more likely to use CXCR4 receptors compared to subtype A (Kaleebu et al., 2007). These studies seem to imply that virulence could be associated with the large genetic differences found between subtypes, as well as HIV type and group, but did not investigate the impact of within-subtype variant, which is less clear.

1.3 Previous Estimates of Viral Genetic Effect on Viral Load

With the practice of taking viral load measures becoming widespread, analyses could begin to quantify the viral genetic effect on viral load instead of simply identifying a correlation. The Zambia-UAB HIV Research Project enrolled 1,022 heterosexual 'serodiscordant' cohabiting couples, where one person was HIV-negative and the other HIV-positive, infected primarily with subtype C but also with A, G, D, and J (Fideli et al., 2001). The HIV-negative partner seroconverted in 162 couples, and transmission from their cohabiting partner was confirmed in 129 pairs by genetic distance and phylogenetic analysis (Fideli et al., 2001; Trask et al., 2002). 115 of these couples had available viral load measures and were studied in order to identify correlation between

the transmitter's viral load and the seroconverter's viral load and estimate the effect of viral factors (Trask et al., 2002; Tang et al., 2004). Tang et al. (2004) reported a Pearson's correlation coefficient of $r = 0.21$ (p=0.030) after adjustment for sex, age, and HLA class, concluding that the viral genome makes a small but significant contribution to viral load (Table 1.1).

Another study looking into the correlation of transmission pair's viral loads was conducted in San Francisco, California, where subtype B is almost exclusively prevalent. HIV-positive and HIV-negative enrolees to the University of California San Francisco (UCSF) Options Project refer sexual or intravenous drug user (IDU) partners, and potential partner transmission is confirmed with phylogenetic analysis (Hecht et al., 2010) (http://labs.ucsf.edu/options/). Viral load samples are taken after a recent infection is confirmed, and compared to the most recent viral load available for the transmitting partner (Hecht et al., 2010). Comparing the viral loads from 23 transmitters and 24 seroconverters (one transmitter had infected two partners), Hecht et al. (2010) reported a Pearsons correlation coefficient of $r = 0.55$ (p=0.006) and suggested this was evidence of strong viral genetic influence on viral load (Table 1.1).

van der Kuyl et al. (2010) identified transmission pairs in the Amsterdam Cohort using phylogenetic analysis, finding 75 pairs with viral load information (van der Kuyl, 2012). 56 of these pairs had information about the HIV infection stage, with 60% being men who have sex with men (MSM) and 40% being heterosexual (van der Kuyl et al., 2010; van der Kuyl, 2012). The resulting Pearsons correlation coefficient was reported as $r = 0.25$, similar to Tang et al. (2004)'s results (Table 1.1).

The Rakai Community Cohort Study followed over 12,000 individuals in the rural Rakai area of south-western Uganda, taking sera samples every 10-12 months from 1994-2003, with 16.5% of the cohort testing HIV-positive (Hollingsworth et al., 2010). 200 suspected transmission pairs were identified where at least one partner seroconverted during the study period, and 29 heterosexual couples were identified as having strong support for transmission by phylogenetic analysis (Hollingsworth et al., 2010). The amount of variance in viral load explained by the model, or coefficient of determination (R^2), was reported for these couples as $R^2 = 27\%$ (Hollingsworth et al., 2010) (Table 1.1).

Table 1.1: Estimated Viral Genetic Effect on Viral Load in Previous Studies

Paper	**Analysis**	**Country**	***N***	**Risk Group**	**Subtype**	**Estimate**
Tang *et al.* 2004	Transmission pair	Zambia	115 pairs	Hetero	95% C	r=0.21[a]
Hecht *et al.* 2010	Transmission pair	USA	22 pairs, 1 triplet	MSM & IDU	B	r=0.55[a]
Hollingsworth *et al.* 2010	Transmission pair	Uganda	29 pairs*	Hetero	62% D, 17% A, 21% A/D recomb.*	R^2 = 27%*[b]
van der Kuyl *et al.* 2010	Transmission pair	Netherlands	56 pairs	60% MSM, 40% Hetero	77% B (pers comm)	r = 0.25[a]
Alizon *et al.* 2010	Phylogenetic signal	Switzerland	134 individuals 404 individuals	MSM	B	K = 0.59[c], λ = 0.51[d] K = 0.09[c], λ = 0.13[d]
Yue *et al.* 2013†	Transmission pair	Zambia	195 pairs (1 VL) 143 pairs (mean VL)	Hetero	>95%C	R^2 = 0.020*[b] R^2 = 0.013*[b]
Lingappa *et al.* 2013†	Transmission pair	East & South Africa	141 pairs	Hetero	43%A, 37% C, 13% D, 6% other	R^2 = 0.06*[b]

'MSM' stands for 'men who have sex with men'; 'IDU' stands for 'intravenous drug user'; 'Hetero' stands for 'heterosexual'

† These papers were published after the start of my PhD research and analysis

* Pairs with strong genetic support for transmission

[a] Pearson's correlation coefficient

[b] Coefficient of determination (Fraser and Hollingsworth, 2010)

[c] Blomberg's measure of phylogenetic signal (Blomberg et al., 2003; Alizon et al., 2010)

[d] Pagel's measure of trait similarity deviation from the expected (Pagel, 1999; Alizon et al., 2010)

One previous study has moved away from the transmission-pair model by instead looking for a signal of inherited viral effect in a phylogeny reconstructed from many HIV sequences in a population (Alizon et al., 2010). The Swiss HIV Cohort Study is estimated to include approximately 45% of the HIV infections declared to health agencies in the country, and includes over 15,000 patients, mostly male (71.2%), across MSM (34.7%), heterosexual (38.8%), and IDU (29.5%) risk groups (The Swiss HIV Cohort Study 2010). From cohort participants in all risk groups, Alizon et al. (2010) included 661 participants who had three consecutive viral load measurements within 1 log (the 'liberal' case) and a subset of 230 of these patients whose viral load measures were all within 1 log (the 'strict' case). Phylogenetic signal measures the amount that the connections in a phylogeny explain the similarity in trait values seen in different individuals, often calculated using methods based on Felsenstein's independent contrasts (Felsenstein, 1985) which measures the difference in traits weighted by the distance between tips on the phylogeny (Alizon et al., 2010) (for a full explanation of independent contrasts, see Section 2.1 on page 23). Alizon et al. (2010) used two estimators of phylogenetic signal: K, described by Blomberg et al. (2003) and based on the mean squared error of the contrasts, and λ, introduced by Pagel (Pagel, 1999), which measures the ability of the evolutionary pathways estimated from the phylogeny to produce the trait measures observed. K and λ were estimated for all risk groups and only MSM in both the 'strict' and 'liberal' datasets, with the 'strict' MSM group of 134 individuals producing a significant result of $K = 0.59$ and $\lambda = 0.51$, which is stated as support for the presence of viral factors with a strong influence on viral load (Alizon et al., 2010). However, the estimate of the viral genetic contribution in the 'liberal' MSM group (n=404) was also significant at $K = 0.09$ and $\lambda = 0.13$ (Table 1.1).

After I had begun work on my PhD, two further papers were published investigating the viral genetic effect on viral load in African cohorts.

Yue et al. (2013) identified 195 phylogenetically-linked heterosexual transmission partners from the Zambia-Emory HIV Research Project in Zambia, for whom at least one viral load from the seroconverter was available $\geq$9 weeks after the estimated date of infection (EDI), and for 143 seroconverting partners multiple viral loads were available over a period of up to 12 months after EDI. More than 95% of the transmission pairs

were infected with subtype C, and all individuals were HLA-I genotyped, to control for the influence of HLA on viral load. The entire dataset (n=195) was used to investigate the correlation between the transmitter's and the seroconverter's first $\geq$9 week viral load, generating a significant correlation of $R^2 = 0.020$ (p=0.046). For the subset where multiple viral loads were available, the geometric mean of all viral loads between 3 and 12 months was used to repeat the analysis, but the resulting correlation of $R^2 = 0.013$ was not significant (p=0.18) (Table 1.1).

Most recently, 3,408 stable heterosexual serodiscordant couples were recruited to the Partners in Prevention HSV/HIV Transmission Study from 14 sites in East and Southern Africa, where viral loads were collected four times within a year for the transmitting partner, and five times within twelve months after infection for the seroconverting partner (Celum et al., 2010; Lingappa et al., 2013). Lingappa et al. (2013) observed 151 seroconversions, of which 141 seroconverting partners had at least one set-point viral load measure from $\geq$4 months after seroconversion, with 101 of those pairs being genetically linked. 43% of the seroconverters were subtype A, 37% were subtype C, 13% subtype D, and 6% were another subtype or the subtype was not identified (Lingappa et al., 2013). A multivariate linear mixed-effect model was used to look at the relationship between transmitting and serconverting partner's viral loads, incorporating multiple viral loads where available, and found that the proportion of variation in seroconverting partner's viral loads explained by their partner's viral load was $R^2 = 0.06$ (Table 1.1).

1.4 Defining 'Set-Point' Viral Load

Alizon et al. (2010)'s use of both 'liberal' and 'strict' measures of viral load highlights the difficulty and lack of standardization in assessing what measurements should be taken as 'set-point' viral load in practice. As previously stated, set-point viral load is usually understood to be the viral load at the beginning of the asymptomatic clinical latency phase, after the viral load spike during the acute phase that follows initial infection (Daar et al., 1991) (Figure 1.6 on page 9). During this phase, the amount of virus in the blood stabilizes, and it is this measurement that is associated with

disease progression (Mellors et al., 1996; Fraser et al., 2007; Langford et al., 2007). However, the exact date of HIV infection in a patient is often unknown, and at best is usually estimated as halfway between the last HIV-negative and first HIV-positive tests, making it hard to determine whether a sample was taken during the beginning of the clinical latency stage, or indeed, during the clinical latency stage at all, and should be taken as the set-point viral load.

The difficulty of knowing a patient's disease stage and the tendency of the viral load to fluctuate marginally (Raboud et al., 1996) or have a slight upward slope during the clinical latency stage (Lyles et al., 1999, 2000; Sabin et al., 2000) means that many different methods of determining set-point viral load have been used. Set-point viral load is sometimes simply the first sample taken after a positive HIV test, or the median value if more than one viral load sample is available (Mellors et al., 2007). Some transmission pair studies have aimed to use the viral load measures closest to the suspected time of transmission, or the recipient's viral load closest to transmission and a matching disease-stage viral load from the transmitter (Tang et al., 2004; Hecht et al., 2010; van der Kuyl et al., 2010). Others have defined the set-point viral load as the first measurement taken a set period of time after the estimated infection, which varies from study to study (García et al., 1997; Vidal et al., 1998; Richardson et al., 2003; Yue et al., 2013), or used the mean value of viral load measures taken some time after initial infection but before the onset of AIDS, starting ART, or a few months before death (Fraser et al., 2007; Hollingsworth et al., 2010; Yue et al., 2013). Lingappa et al. (2013) chose to use a more complex model which allowed multiple viral loads taken at set intervals to be included for each seroconverter. When multiple viral load samples are available, studies sometimes choose to be more restrictive in the data they include, using only patients where a certain number of consecutive measures or all measures fall within a 0.5 or 1 log band around the mean patient viral load measure (Fellay et al., 2007, 2009; Alizon et al., 2010).

Mei et al. (2008) point out that many studies have to choose between having more viral loads from fewer patients, or more patients with fewer viral loads, due to limited resources. Studies analysing cohort data often must make a similar choice between including many viral loads that might not actually be 'set-point,' or being stricter

about which viral loads to include, but losing a large amount of available data. In a study using the more liberal definition of set-point as the mean of viral loads taken during a set period after infection and before death, a cohort of 200 possible couples was reduced to 112 (Hollingsworth et al., 2010). However, in a study with a stricter definition of set-point that only utilized patients who had at least 3 consecutive samples within a 1-log band, only 661 patients were included from a cohort that contains over 15,000 (Alizon et al., 2010; The Swiss HIV Cohort Study, 2010). The decision of how to define set-point must be carefully made, as being too liberal could include viral loads from acute-stage or AIDS-onset patients, or patients on ART, while being too strict runs the risk of over-fitting the analysis to include only 'perfect' patients that may not represent the population as a whole, as well as decreasing the power of the study because of the reduction in sample size.

1.5 Limitations of the Transmission-Pair Model

Six of the previous studies estimating the viral genetic contribution to viral load studied seroconverting transmission pairs in their analysis. Transmission pair analyses provide a direct way to observe viral load in both the transmitting and seroconverting partner, often over a period of months or even years. Viral sequences from each partner can confirm whether the transmission was from the long-term partner, or from extra-pair contact, so that only confirmed direct transmissions are included. In many HIV databases and HIV studies, no partner data is available, meaning that transmission-pair studies are one of the few datasets where access to information on both partners is available. Measuring the genetic contribution to viral load in transmission pairs is commonly done by using regression, correlation, or linear mixed models to measure the similarity in viral load between the two partners after controlling for other effects (Tang et al., 2004; Hecht et al., 2010; Hollingsworth et al., 2010; van der Kuyl et al., 2010; Yue et al., 2013; Lingappa et al., 2013). Despite the advantages of transmission-pair studies, there are also some limitations that could affect estimates of viral genetic influence on viral load.

Sample sizes in transmission pair analyses are usually limited, due either to the relative rarity of transmissions between partners in a cohort under observation (Tang et al., 2004; Hollingsworth et al., 2010; Lingappa et al., 2013; Yue et al., 2013), or the lack of available partner data for recently seroconverted individuals (Hecht et al., 2010). From a study cohort of over 12,000 individuals, Hollingsworth et al. (2008) identified only 29 couples with genetically-linked infection. Similarly, of the 3,408 long-term serodiscordant couples recruited and followed for up to 24 months in the cohort used by Lingappa et al. (2013), only 101 transmissions could be genetically linked. The sero-discordant couples where transmission does occur may not accurately reflect the epidemic as a whole, as viral load is linked to transmission risk. Tang et al. (2004) found that partners who transmitted HIV and thus got included in the analysis had higher viral loads than the average for the rest of the study population.

Cohabiting or long-term sexual partners are also likely to share confounded environmental factors such as diet and exposure to other pathogens, which could affect health and thus viral load. There is also a potential confounding effect due to shared human leukocyte antigen (HLA) alleles, which allow the immune system to recognise pathogens. A study of mother-to-child HIV transmissions showed that cytotoxic T lymphocyte (CTL) escape variants acquired by HIV in the mother can be transmitted to the child, increasing the chance of a successful infection and causing faster disease progression because the child shares half of the mother's HLA profile, which the virus already has mutations to escape (Goulder et al., 2001). Studies have shown that HIV-negative partners in heterosexual couples with discordant HIV status are more likely to become HIV-positive if they share HLA alleles with their partner (Lockett et al., 2001; Dorak et al., 2004). Tang et al. (2004) found evidence that shared HLA alleles can not only influence transmission risk, but also viral load, citing an increase in the viral load of seroconverters whose partners had shared HLA alleles, though the increase was not statistically significant in the 115 transmission pairs studied. In a much smaller cohort of 24 transmission pairs, Hecht et al. (2010) failed to find evidence that HLA had a significant effect on the correlation between transmitter and seroconverter viral load. Yue et al. (2013) investigated this effect again on a larger study of 195 couples and found a highly significant increase in seroconverters viral load when the partners

shared HLA alleles. As none of the other three transmission-pair studies had data on the HLA profile of individuals in their studies, estimates of the viral influence on viral load in these studies could be confounded by the effect of shared HLA alleles.

Finally, transmission pair studies are often done in cohorts that were initially formed for other reasons. The seroconverting transmission pairs in Hollingsworth et al. (2010)'s study were identified retrospectively from the larger study cohort, meaning that much of the data used was not collected with a transmission-pair study in mind, which the authors admit could lead to unidentified biases. Lingappa et al. (2013) selected transmission pairs from a study originally founded to investigate whether suppressing genital herpes simplex virus with acyclovir in HIV infected partners could reduce transmission risk. Though acyclovir was not found to reduce HIV transmission, it did reduce viral load in the HIV infected partner by 0.25 $\log_{10}$ copies/mL Celum et al. (2010). Acyclovir use was included as a factor in the subsequent transmission-pair study Lingappa et al. (2013), but its significant effect on viral load in the transmitting partner could introduce bias that is hard to control for.

Given the potential issues in estimating the viral genetic contribution to viral load in small and restricted sample sizes and in long-term transmission pair studies, I wished to utilize a new method that would allow the inclusion of very large datasets, where direct partner information is not available, while controlling for environmental and other effects.

"With numbers, with facts of all sorts, Lacon never faltered. They were the gold he worked with, wrested from the grey bureaucratic earth."

John le Carré - 'Tinker Tailor Soldier Spy' (1974)

"Kynes – direct, savagely intent Kynes – knew that highly organized research is guaranteed to produce nothing new."

Frank Herbert - 'Dune' (1965)

2 Methods

In this chapter, I introduce a new, quantitative genetics-based method for estimating the heritability of traits using phylogenies. First, I give a brief overview of heritability estimation and its challenges, and some background on the previous methods used to estimate heritability. I then introduce and describe the theory behind the new method used in Chapters 3 and 4 to estimate the heritability of viral load in HIV. Next, I give an overview of dataset used for these analyses, the UK HIV Drug Resistance Database, and outline the data cleaning procedure, including the process of choosing the 'set-point' viral load. I detail the methods used to build phylogenies from the HIV sequences, and describe how collapsing unsupported splits in the phylogeny was used to investigate the effect of uncertainty in the tree on my estimates. Finally, I outline the details of the heritability estimation pipeline, and the method used to investigate change in set-point viral load over time.

2.1 Background of Heritability Estimation

In quantitative genetics the proportion of the total trait variation (V_P) caused by genetic factors (V_A) is described as the heritability (h^2) of the trait. Though numerous models have been proposed to estimate variance components and heritability by a variety of methods, many of the initial methods were superseded, as they can be prone to bias and inflexibility (Hadfield and Nakagawa, 2010).

All methods of estimating heritability are based on the premise that a heritable trait should have a higher covariance in individuals that are more closely related than in those that are more distantly related. A simplified example of a phylogeny and how the traits on it might covary is given in Figure 2.1. If the heritability of the trait is high, as in a), the covariance will be high in traits from closely related individuals, and will be lower and more randomly distributed in distantly related individuals. If the heritability of the trait is low, as in b), traits from more closely related individuals will have a higher covariance than traits from distantly related individuals, but the covariance will be lower and more widely distributed than in a). Methods to estimate heritability attempt to determine whether the covariance of traits in closely related individuals is higher than expected by chance and, if so, quantify how strongly the genetic relationships influence the trait.

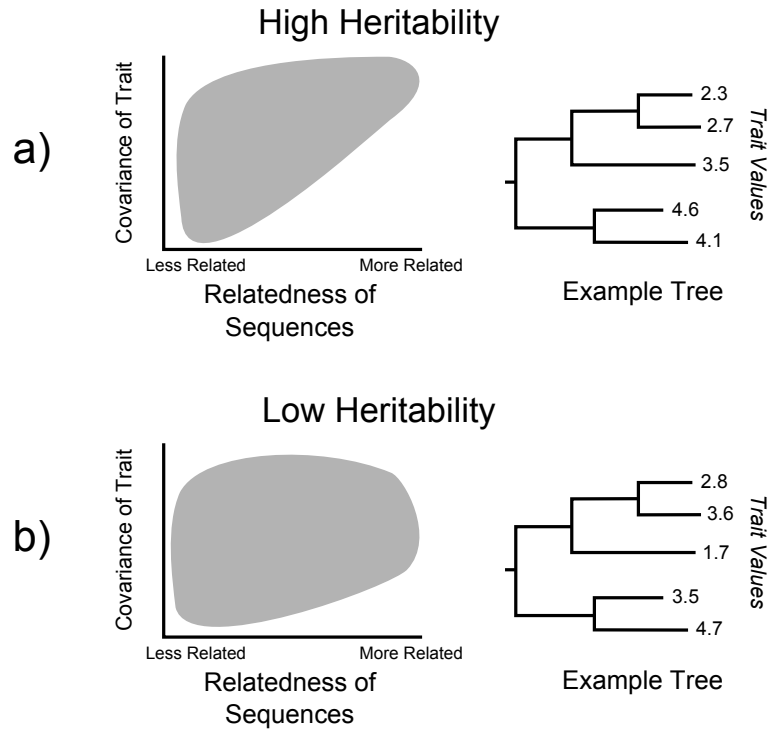

Figure 2.1: An example of trait covariance in phylogenies with high and low heritability. The trait in a) has a high heritability, causing the trait values of more closely related individuals to have a high covariance. The covariance in trait values between more distantly related individuals is lower and more randomly distributed. In b), heritability is low. Though the covariance of traits between closely related individuals is still higher than in traits between more distantly related individuals, it is more widely distributed than in a), as the influence of the genetic relationships on the trait is weaker.

Early attempts to correlate traits in phylogenies had been done mostly by tech-

niques based on linear regression, which assumes the trait values of species are independently drawn from a distribution (Felsenstein, 1985). Felsenstein (1985) pointed out that species are not independent, but related according to their phylogeny, and so their relationships must be taken into account. Solutions to this problem had been suggested, but were based upon inaccurate assumptions about the Linnean classification system or parsimony approaches which try to use the fewest number of state changes to construct the observed relationships, but may not accurately reflect the actual evolutionary path taken (Felsenstein, 1985). Felsenstein (1985)'s method proposed a way of using the phylogeny to weight the observed trait differences in pairs of tips by the distances between the tips on the tree to create 'independent contrasts' which can be tested for correlation, avoiding the problems of previous methods. Though Felsenstein (1985)'s method has proven popular and been used as the basis for many other models, it assumes that the variance in the evolutionary change of the trait down the tree is constant, causing it to perform very poorly if trait changes at speciation events earlier in the phylogeny are larger than trait changes at speciation later in the phylogeny, or vice versa (Felsenstein, 1985; Pagel, 1999; Housworth et al., 2004). This assumption also restricts analysis to only using only species' mean trait values, as within-species trait variation could include rapid, reversible trait evolution that does not reflect the variance of gradually accumulated changes in the earlier phylogeny (Housworth et al., 2004). To include within-species variation in an updated version of the model, multiple measurements from each species were incorporated instead of mean trait values (Housworth et al., 2004; Felsenstein, 2008). However, the model still makes the assumption that Brownian motion, or random changes, in the phylogeny's history can completely explain the trait values, and lacks a residual term to explain deviations from the expected trait value, such as those caused by non-heritable effects and fast genetic changes within a species (Housworth et al., 2004; Hadfield and Nakagawa, 2010).

'Mixed models' were first introduced by Henderson in the early 1950's (Henderson, 1950, 1973) to include both fixed and random effects into a model explaining the observed data values. Fixed effects are those correlated with the independent variable that need to be controlled for, such as sex, and random effects are those that are associated 'randomly' with the data points, where their contribution to the variance is being

investigated, such as maternal effects. Before mixed models, models that assumed all effects were random effects could be used to estimate variance components, but if some of the effects should instead have been treated as fixed effects, the estimated variances were biased (Henderson, 1953). A solution to this was to try and estimate the fixed effects and correct the data for these effects before proceeding with the analysis, but this was a lengthy multi-step process that could still result in biased variance estimates (Henderson, 1953; Henderson et al., 1959; Thompson et al., 2005). Henderson's mixed model can be described as

$$y = X\beta + Zu + \epsilon$$

Where y represents the observed data values, β represents the fixed effects, u represents the random effects, ϵ represents the error term, and X and Z represent matricies of the regressors that relate y to β and u. Pedigrees illustrate the parental lineage and offspring of a group of animals or people, and provide information that can be used to calculated the expected genetic relatedness of individuals (Figure 2.2 on page 29). In order to calculate the amount of variance in the trait due to the genetic relationships of the individuals, one of the random effects included in the mixed model could be the pedigree, and this specific application of the mixed model became known as the 'animal model' (Henderson, 1973, 1976; Thompson et al., 2005).

In 1991, Lynch showed that animal models could also be applied to phylogenetic data, since an absolute time scale, available to pedigrees as discrete generations, is not needed as long as units proportional to time are used (Lynch, 1991; Hadfield and Nakagawa, 2010). Lynch (1991)'s new model was based on previous animal breeding methods of estimating variance components (Henderson et al., 1959; Henderson, 1973), but also implemented expectation-maximization (EM) algorithms (Dempster et al., 1977) to maximize the likelihood values for the variance component and heritability estimates by iteration. Lynch denoted heritability estimates found using this method as h^2. Unlike Felsenstein (1985)'s earlier model, Lynch realised that rapidly evolving traits could have variation in their current state that is not predicted by their phylogenetic past of gradually accumulating changes, and so included a residual term to explain mean trait values that differ from their expected value (Housworth et al., 2004; Hadfield and

Nakagawa, 2010). However, at the time the paper was published the computational intensity of finding the maximum-likelihood (ML) of these values and the fact that the proposed EM algorithms were often slow to converge made the method largely impractical, and it failed to gain much popularity (Thompson et al., 2005; Hadfield and Nakagawa, 2010).

Ordinary least squares (OLS) is a form of linear regression that attempts to minimize the vertical distance between the observed response and the response predicted by the regression. Generalized least squares (GLS) is a special case of OLS that can be used when the data are not independent (Pagel, 1999), and is equivalent to applying OLS to linearly transformed data, which is independent. Pagel (1999)'s GLS-based method generates a variance-covariance matrix to predict the covariance of each pair of tip values based on their phylogenetic distance, and multiplies these covariances by a value, designated λ, to generate a distribution of predicted trait values. The value of λ that best replicates the observed tip values measures how well the phylogeny alone produces the pattern of trait measures observed, or the amount of phylogenetic contribution to the observed variation in trait values (Pagel, 1999; Alizon et al., 2010).

Lynch (1991)'s h^2 and Pagel (1999)'s λ are equivalent, but both models fail to take into account the uncertainty in the fixed effect estimates, which usually have a high sampling error due to being associated with the ancestral state, causing the variance component estimates to be downwardly biased (Hadfield and Nakagawa, 2010). Restricted maximum-likelihood (REML) methods emerged as the preferred choice for variance component and heritability estimation in animal breeding and quantitative genetics due to their ability to give unbiased variance parameter estimates (Patterson and Thompson, 1971; Thompson et al., 2005; Hadfield and Nakagawa, 2010). In ML estimates of mixed models, the fixed effects are estimated first and used to estimate the variance components without recognising the resulting loss of degrees of freedom, which introduces bias due to the uncertainty in the fixed effect estimate (Falconer and Mackay, 1996). In REML, the fixed effects are estimated in a similar fashion, but then 'error contrasts' are created from the observed data by transforming contrasts to have an expected fixed effect, or mean, of zero, which can then be used to estimate the variance without bias (Lynch and Walsh, 1998).

Methods to improve the efficiency of matrix calculations and the use of ML and REML were slowly realised, allowing more complex models and much larger datasets to be utilized (Thompson et al., 2005). By 1996 the software package ASReml introduced an efficient implementation of REML-based variance estimation specifically designed for use with animal models (Gilmour et al., 2009; Hadfield and Nakagawa, 2010).

By measuring the relationships between individuals on the pedigree as the probability their alleles are identical by decent (IBD) and linking this to the observed differences in trait measures, the amount of trait variation explained by the pedigree, or genetic relationships, can be estimated. In animal breeding these IBD relationship measures are calculated from the pedigree and inserted into a genetic relatedness matrix, usually referred to as $\mathbf{A}$. Hadfield and Nakagawa (2010) showed that mathematically an inbred pedigree is equivalent to a phylogeny, where the branch lengths of the phylogeny correspond to inbreeding coefficients. For a phylogeny, the genetic relatedness of two taxa is already present in the topology and branch lengths of the reconstructed phylogeny, and is equivalent to the total length from the taxa's most recent common ancestor (MRCA) to the root (Hadfield and Nakagawa, 2010). In order to calculate variance components the inverse of $\mathbf{A}$, $\mathbf{A}^{-1}$, is usually needed, but can be computationally resource intensive to calculate (Henderson, 1976; Hadfield and Nakagawa, 2010). Henderson (1976) was one of the first to show that for pedigrees this problem can be made easier by including 'phantom parents' for all individuals with unknown parentage so that the population could be traced back to unrelated ancestors. Hadfield and Nakagawa (2010) extended this technique to phylogenies by expanding $\mathbf{A}$ to include all the internal ancestral nodes in the tree, allowing the inverse matrix to be calculated by Henderson (1976)'s method. This proof allows a reconstructed phylogeny to be analysed using well-established methods from animal breeding such as REML, with branch lengths and ancestral splits measuring relatedness instead of discrete generations (Hadfield and Nakagawa, 2010) (Figure 2.3 on page 30).

Here, I take advantage of this development and use ASReml to estimate the heritability of viral load, allowing HIV sequences with matched viral load data from thousands of individuals to be included in the analysis without exceeding processing capacity.

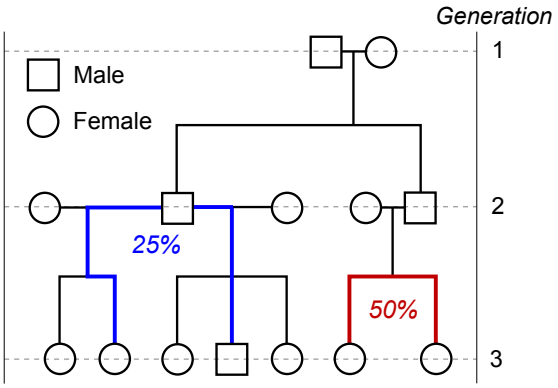

Figure 2.2: A simplified example of how genetic relationships can be determined from a pedigree. By tracing the relationship of two individuals and their parents and grandparents, the expected relationship between two individuals can be identified. For example, siblings that share both parents ('full-sibs') are expected to share half of their genes (shown in red). Siblings that share just one parent ('half-sibs') are expected to share one quarter of their genes (shown in blue). When inbreeding is involved, calculating relationships becomes more complex, but can still be done by tracing the amount of shared ancestry between the parents of an inbred individual.

2.2 Data

2.2.1 Drug Resistance in HIV

ARTs for HIV are often divided into six classes, depending on which step of HIV replication they inhibit: cell entry, reverse transcription, integration, transcription, virus assembly/production, and virion maturation (protease processing) (Arts and Hazuda, 2012). A large number of drugs aim to stop reverse transcription, when the viral RNA is transcribed into DNA to be inserted into the host genome, and protease processing, when the final proteins are cleaved to produce a mature virion that can go on to infect other cells, by targeting the *RT* (reverse transcriptase) and *protease* genes respectively (Arts and Hazuda, 2012). *RT* and *protease* are located next to each other within the *pol* region of the HIV genome, and as major targets of ART, are locations where drug resistance mutations (DRMs) arise frequently. If DRMs arise or are present in a patient, the efficacy of the drug is compromised, and viral replication persists, even if at a low level. Failure to detect resistance or counter it with an effective drug regimen can lead to a vicious cycle of increasing drug resistance that eventually cannot be controlled by available treatment (Clavel and Hance, 2004). DRMs are also associated with poorer long-term outcome and increased mortality (Kozal et al., 2007; Hogg et al., 2006).

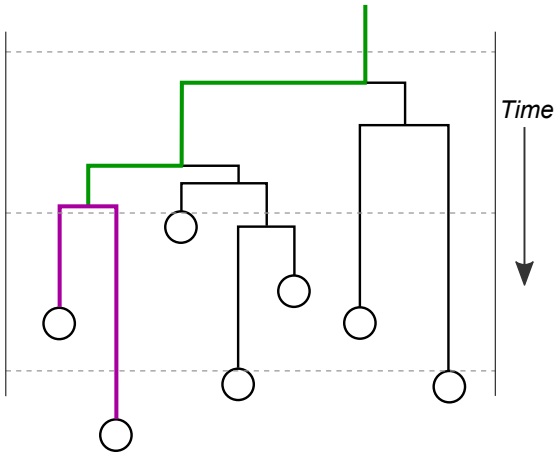

Figure 2.3: Unlike in a pedigree (see Figure 2.2), phylogenies do not have discrete generations or 'parents' to trace ancestry through. However, in a similar way to how relatedness can be calculated in an inbred pedigree, the relatedness of two individuals in a phylogeny can be determined by looking at the amount of time they share ancestry (time before they diverged, green), and the amount of time they evolved separately (time after they diverged, purple). Hadfield and Nakagawa (2010) demonstrated how these two measures can be used to calculate the genetic relatedness of two individuals in a phylogeny.

DRMs can be acquired by HIV if viral replication is not totally suppressed by ART, allowing the treatment to select for resistant strains replicating with the patient. DRMs can also be transmitted from one individual to another drug naive individual (transmitted drug resistance), which can lead to immediate treatment failure and further DRMs if not detected before ART is started. The amount of transmitted drug resistance has changed over time (Little, 2001; Grant et al., 2002), and poses a potential threat to the efficacy of available treatment if it were to become wide-spread, leading to recommendations that all patients be tested for DRMs by genotyping *pol* prior to starting ART and at any sign of treatment failure (Miller et al., 2001; Thompson et al., 2012; Williams et al., 2014). To aid in recognising and recording DRMs, in 2007 the World Health Organization pushed for the development of a comprehensive consensus list of DRMs, which is kept updated as new ART drugs become available and new DRMs arise (Shafer, 2006; Bennett et al., 2009).

2.2.2 The UK HIV Drug Resistance Database

As an effort to monitor drug resistance in HIV-positive patients within the UK, the UK HIV Drug Resistance Database (UK HIV RDB) was established in 2001, which collects

pol sequences from HIV-positive patients attending clinics across the UK before starting and during ART in order to detect DRMs. Viral load measurements are also taken before starting ART as a proxy for disease progression, and during ART to ensure that the therapy is successfully suppressing viral replication. On effective ART, a patient's viral load should be very low or undetectable, so rising viral load measures while on ART often indicates the development of drug resistance. Finally, some limited clinical information is collected.

The UK HIV RDB 2010 release, used for my research, contained 55,556 HIV sequences from 43,002 patients, with at least one viral load before starting ART being available for 13,309 patients. In 2006, 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 in 2006 (Leigh Brown et al., 2011). Fully anonymised clinical data corresponding to many of the sequences was made available by the UK Clinical HIV Cohort (UK CHIC) (The UK Collaborative HIV Cohort Steering Committee, 2004), and linked to the corresponding sequences and viral loads. The data used were the most current available, with sequences and clinical data collected up to mid-2009.

As well as recording viral load counts and the dates of viral load tests, the UK HIV DRB has information on the date that patients started ART treatment. As ART greatly reduces the viral load, all viral loads taken after the reported start of ART were excluded from the analysis.

Though the collection and management of the UK HIV DRB sequences and UK CHIC data is rigorous, patient-reported information is always potentially incorrect. Particularly with the stigma associated with HIV, it is not unreasonable to believe that some patients may provide erroneous information, particularly with regard to the suspected exposure route. In the analysis performed here, most information is clinical values, not patient-provided data. Sex, ethnicity, date of birth, and country of origin are provided by the patient; it seems unlikely that many patients would feel the need to falsify this information.

A potentially larger problem is how the data from the UK HIV DRB and UK CHIC are linked together when gathered from the participating clinics across the UK. Though all efforts are taken to accurately track and identify patients and their samples,

patients moving between clinics may not be correctly recognised. For example, patients moving between two clinics that provide data to the UK HIV DRB could be erroneously recorded as two independent patients. The fast evolutionary rate of HIV means that no two patients have identical sequences (indeed, even sequences from the same patient over time are often not identical), so if identical sequences were found in the dataset, they were excluded as possibly being from the same patient, recorded under two or more different identifiers. Patients moving from a clinic that does not provide data to the UK HIV DRB or UK CHIC to a clinic that does provide data could have information about the history of their treatment missed out. For example, the patient may have been diagnosed and put on treatment at the first clinic, and fail to inform the second clinic, who will erroneously record a later date of diagnosis and treatment start. Any viral loads taken before the recorded date that ART began would seem to be pre-ART values, but might actually be influenced by unreported ART prescribed at a different clinic.

2.3 Choosing 'Set-Point' Viral Loads

As explained in the introduction (Section 1.4 on page 18), deciding what viral load measurements to take as 'set-point' viral load is rarely straightforward due to the lack of information about when the patient was infected. Inclusion of non-set-point values could bias the heritability estimates obtained, but being too restrictive in the inclusion of viral load values could lead to a loss of power due to a smaller sample size, and produce heritability estimates that only apply to a specific subset of the HIV infected population, and are not generalizable to the UK epidemic.

To maintain both the representativeness of the HIV epidemic in the UK and as large a sample size as possible to improve power, a liberal definition of set-point viral load was chosen. If multiple pre-ART viral load measures were available for a patient, the first viral load was generally taken as the 'set-point' viral load.

However, to try and minimise the number of viral load values included that were actually taken during the acute phase, during AIDS, or on unreported ART, data cleaning rules were introduced.

2.3.1 Multiple Viral Loads

As first viral load was generally taken as the set-point viral load, patients with exceptionally high or low first viral loads and multiple viral loads available were examined to determine whether a viral load other than the first should be taken as the 'set-point.' Example figures illustrating some of the rules can be found in Appendix A on page 59.

High Viral Loads

Patients with first viral loads ≥1,000,000 copies/mL could either be in the acute phase or have progressed to AIDS. Some methods of measuring viral load have an upper measurement limit of 500,000 or 750,000 copies/mL. Viral loads recorded at exactly these values are possibly much higher than this, and so were treated as if they were ≥1,000,000 copies/mL. Patients with such high first viral load values were assessed differently according to the number of viral loads available.

Though efforts were taken not to entirely exclude patients with multiple viral loads, and instead use a viral load other than the first viral load if acute phase, AIDS, or unreported ART was suspected, if a patient seems to go from acute phrase or AIDS to being on ART, none of the viral loads may be usable. A very quick drop from very high viral load to relatively low viral load in a short space of time could indicate that a patient in acute phrase or AIDS started unreported ART. If a patient's viral load was ≥1,000,000 copies/mL (or exactly at the measurement limits of 500,000 or 750,000 copies/mL), and dropped below 1,000 copies/mL within 3 months (93 days), the patient was excluded entirely, to prevent viral loads taken during AIDS, acute phrase, or while on unreported-ART from being used as set-point.

Rules for exclusion of potential non-set-point viral loads when multiple viral loads are available, and the first viral load is high:

1. If more than three viral loads were available when the first viral load was ≥1,000,000 copies/mL, the viral loads of the patient were inspected by hand. An example of the R interface coded to accomplish this is shown in Figure A.1 on page 60. After plotting all the pre-ART viral loads, I looked for evidence that the patient

had moved beyond the acute phase and into the chronic phase by looking for an 'inflection point,' within one year of the first viral load, where the viral load values stopped decreasing and starting increasing (Figure 2.4 on the next page).

   (a) If an 'inflection point' is present, this viral load was used as the set-point.

   (b) If no 'inflection point' was obvious, then the lowest viral load within one year of the first viral load was used as the set-point.

2. If exactly three viral loads were available when the first viral load was ≥1,000,000 copies/mL, the middle value was inspected to see if it appeared to be an 'inflection point' (lower than the other two values), and if so, it was taken as the set-point viral load (Figure A.2 on page 61). If the middle value was not the lowest value, the last viral load was taken as the set-point.

3. If only two viral loads were available when the first viral load was ≥1,000,000 copies/mL, the lowest value was taken as the set-point viral load. If the two viral loads were equal, the first value was taken as the set-point.

Low Viral Loads

Though the only viral loads considered were those before the reported start of ART, patients with first viral loads ≤50 copies/mL could be on unreported ART, as detailed above (Section 2.2.2 on page 30). Some methods of measuring viral load have a lower measurement limit of 400 copies/mL. Viral loads recorded at exactly this value are possibly lower than this, and so were treated as if they were ≤50 copies/mL. Some patients do have the ability to 'control' their viral load at very low or near undetectable levels, and with the data currently available there is currently no way to distinguish these patients from those on continuous ART. In cases where all viral loads remain low, the first viral load was used as the set-point. However, sudden rises from a low viral load to high viral loads in short periods of time are not consistent with patients who can control their viral load, and may indicate a patient stopping unreported ART, which causes the viral load to 'rebound' to the set-point value. To detect this, I looked for a change of at least 2 logs within a year of the first viral load. Patients with low

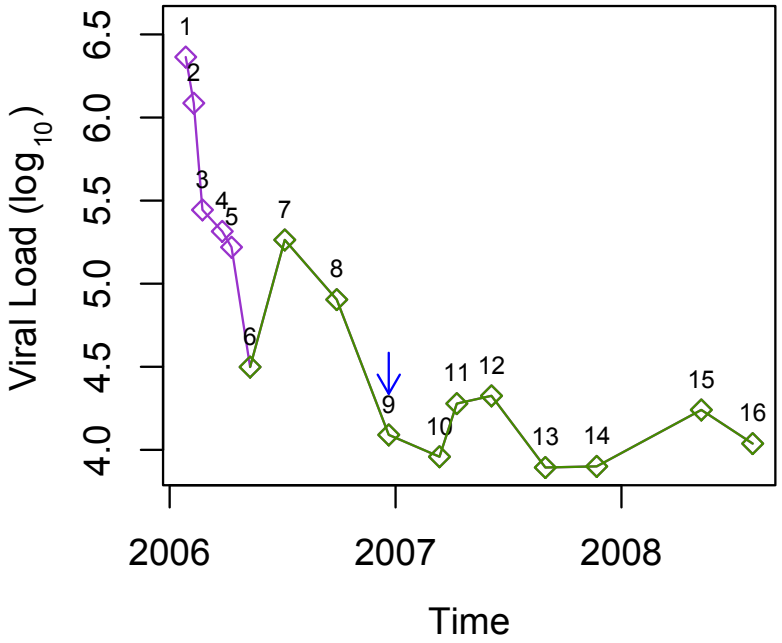

Figure 2.4: An example plot of a patient with a very high first viral load and more than three viral loads available. The blue arrow indicates the last viral load taken within a year of the first viral load. Each viral load measure is numbered. Here, the 'inflection point' where the viral load stops decreasing and starts increasing is viral load measure number six. Selection of value six as the viral load is visually confirmed by changing the colour of the plot from measure six onwards. An example of the R interface coded to manually select these viral loads is shown in Figure A.1.

first viral load values were again assessed differently according to the number of viral loads available.

Rules for exclusion of potential non-set-point viral loads when multiple viral loads are available, and the first viral load is low:

1. If more than three viral loads were available for patients with a first viral load of ≤50 copies/mL, they are examined for evidence of a sharp rise in viral load that may indicate discontinuation of ART.

   (a) First, the difference between the $\log_{10}$ 50 copies/mL value and the maximum viral load values is calculated, then any leading 50 copies/mL measures are removed, and the difference between minimum and maximum viral load is measured again. If the difference between these two values is ≥2 logs, the leading 50 copies/mL values are excluded (Figure 2.5 on page 37). This is because the leading 50 copies/mL values are significantly different from the

remaining values, indicating that the leading measures may have been taken on unreported ART, and after ART was stopped, the viral load rebounded to the actual 'set-point' values.

(b) After this step, the viral loads were examined to see if there was any increase of more than 2 logs between the first viral load and any other viral load within one year of the first viral load. If an increase of 2 logs or more is found in the first year, the value after the increase is used as the set-point (Figure A.3 on page 62). If not, the first viral load is used.

2. If two to three viral loads are available for patients with a first viral load of $\leq$50 copies/mL, viral loads were examined to see if any increase of more than 2 logs between the first viral load and any other viral load within one year was present. If so, the value after the increase was taken as set-point, and if not, the first viral load is used.

3. Patients with viral loads <400 copies/mL could also be on unreported ART, and were examined to see if there were signs of viral load rebound after stopping ART. If more than one viral load was available, the viral loads were checked for evidence of a $\geq$2 log increase within one year of the first viral load. If present, this value was used as set-point. If there was no large rise in viral load, or if only one viral load was available, the first viral load was taken as the set-point value.

2.3.2 Only One Viral Load

For a subset of patients in the UK HIV DRB, only one viral load value is available before the reported start of ART. When only one viral load measure is available, additional information about the potential disease stage or ART status cannot be gathered from overall viral load trends, making these cases more likely to be included as set-point viral loads when they are not. In these cases, $CD4^+$ count can offer some insight into disease stage and ART status (see Section 1.2 on page 7), and so was used to assess whether to include these single viral load values.

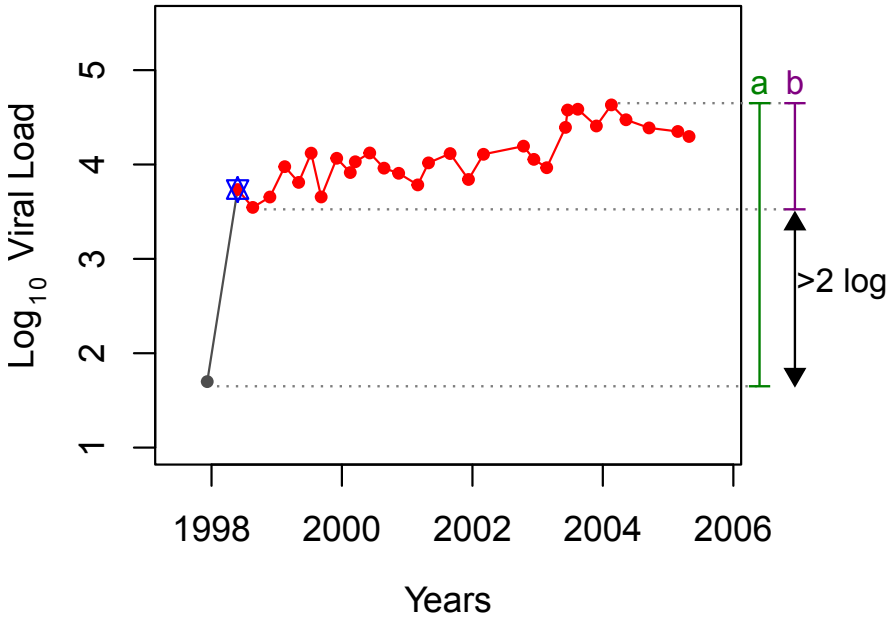

Figure 2.5: An example plot of a patient with a very low first viral load and more than three viral loads available. The leading measure of 50 copies/mL is in grey. It has been excluded, as there is a ≥2 log difference between the difference in the minimum and maximum viral loads including the leading 50 copies/mL value (a), and excluding the 50 copies/mL value (b). The point covered by the blue star is the value taken as set-point, and given the remaining viral loads, is likely a much better estimate of the set-point value than the first viral load.

Rules for exclusion of potential non-set-point viral loads when only one viral load is available:

1. The combination of a very high viral load and a reasonably high $CD4^+$ cell count is likely to occur during the acute phase of HIV infection, when the virus is replicating unchecked, but the $CD4^+$ cell count has not yet been severely affected by the HIV infection. In patients where only one viral load was available and this situation was present, the patient was excluded to prevent an acute-stage viral load being counted as set-point viral load.

   (a) If only one viral load measure is available, and is ≥1,000,000 copies/mL, a $CD4^+$ cell count within one month of the viral load was examined. If the $CD4^+$ cell count was greater than 500 cells/μL, or if no $CD4^+$ was available within one month, the patient was excluded.

   (b) As explained in the methods 2.3, viral loads recorded at exactly the measurement limits of 500,000 or 750,000 copies/mL were treated as if they were

≥1,000,000 copies/mL. If only one viral load measure is available and is equal to 500,000 or 750,000 copies/mL, a $CD4^+$ cell count within one month of the viral load was examined. If the $CD4^+$ cell count was greater than 500 cells/μL, or if no $CD4^+$ was available within one month, the patient was excluded.

2. The combination of a very low viral load and reasonably high $CD4^+$ cell count is likely to occur when the patient is on ART, suppressing the virus and allowing the immune system to recover. Though all viral loads used in the analysis were taken before ART was reportedly started, it is possible that some patients were on unreported ART. In cases where only one viral load was available and this situation was present, the patient was excluded to prevent a viral load taken on ART being counted as a set-point viral load.

   (a) If only one viral load measure is available, and is ≤50 copies/mL, a $CD4^+$ cell count within one month of the viral load was examined. If the $CD4^+$ cell count was greater than 200 cells/μL, or if no $CD4^+$ was available within one month, the patient was excluded.

   (b) As explained in the methods, 2.3, viral loads recorded at exactly the measurement limit of 400 copies/mL were treated as if they were ≤50 copies/mL. If only one viral load measure is available and is equal to 400 copies/mL, a $CD4^+$ cell count within one month of the viral load was examined. If the $CD4^+$ cell count was greater than 200 cells/μL, or if no $CD4^+$ was available within one month, the patient was excluded.

3. One patient had only one viral load value, which was reported as 1 copy/mL. As the lower measurement limit on most methods of measuring viral load is 50 copies/mL, and no other viral load or $CD4^+$ data was available to verify this measurement, the patient was excluded as a potential database error.

Before analysis, all viral loads were $\log_{10}$ transformed to make their distribution approximately normal.

2.4 Sequences

Though the *pol* sequences collected by the UK HIV DRB are gathered primarily for the purpose of detecting DRMs, *pol* sequences are also valuable in analysing transmission and epidemic dynamics using phylogenetic tools. Initially, most genetic analysis of HIV was carried out on *env* and *gag* (Balfe et al., 1990; Simmonds et al., 1990; Ou et al., 1992; Holmes et al., 1993; Albert et al., 1994; Leigh Brown et al., 1997; Yirrell et al., 1997), with *pol* criticised as being too genetically conserved to accurately reconstruct transmission dynamics (Stürmer et al., 2004). However, Hué et al. (2004) confirmed that *pol* had sufficient variability to allow phylogenetic analysis and reconstruction of transmission clusters, and the *pol* region is now widely used to study transmission and epidemic dynamics (Pao et al., 2005; Brenner et al., 2007; Gifford et al., 2007; Leigh Brown et al., 2011; Ragonnet-Cronin et al., 2013; Wertheim et al., 2014).

All sequences collected and stored by the UK HIV DRB are aligned using the Stanford HIVdb Program (Shafer, 2006), with manual checks for high levels of ambiguity and poor quality. Only the first sequence available for each patient was analysed.

Phylogenetic methods use differences and similarities between sequences to reconstruct the evolutionary history and relationship between the sequences. DRMs occur at specific locations and involve particular base changes, and so are often identical between otherwise independent sequences. These identical sites could potentially make sequences appear to be more closely related than they actually are, affecting the structure of the phylogeny. The actual effect of drug-resistance mutations in this regard is debated (Bansode et al., 2011), but as 11.3% of subtype B sequences submitted to the UK HIV DRB from ART-naive patients between 2002 and 2009 were found to have at least one DRM (UK Collaborative Group on HIV Drug Resistance, 2012), all sequences were stripped of codons in positions associated with drug-resistance mutations (Rhee et al., 2003; Shafer, 2006) before phylogenetic analysis to eliminate any possible bias in the phylogeny caused by identical DRMs. Similarly, having large regions where sequence is missing could affect how phylogenetic methods place sequences in the phylogeny. Missing regions are usually ignored by phylogenetic methods, but the remaining part of the sequence will end up completely determining the relationship to the other

sequences, which means that decisions are being made on a limited amount of data. Because of limitations in sequencing technology, the variety of sequencing method used, and errors in sequencing, some sequences were missing the entire *RT* or *protease* gene, and others had up to 200bp missing in *RT*. All sequences with these large fragments missing were discarded from the analysis (for numbers discarded in subtypes B and C, see Table 3.1 on page 68 and Table 4.1 on page 90, respectively). 603 subtype B sequences and 229 subtype C sequences were acquired through the TruGene genotyping kit (Grant et al., 2003; Kuritzkes et al., 2003), which leaves a gap of 120 missing basepairs. As this gap is relatively small and discarding these samples would lead to losing a number of sequences, these sequences were included in the analysis.

As already mentioned, identical sequence pairs and triplets were also removed from the analysis, as they are likely to include multiple samples from the same individual submitted under different patient identifiers (see Section 2.2.2 on page 30).

2.5 Phylogenetic Analysis

Phylogenetic trees organize sequences based on the similarities and differences between base pairs or codons, with more similar sequences being placed closer together on the tree. Methods of constructing phylogenetic trees have grown more complex over time and as computational power increased, and now range from relatively simple stepwise comparisons to exhaustive maximum-likelihood (ML) -based searches and complex Bayesian inference.

Phylogenetic analysis has been used extensively to study the evolution, history, and dynamics of HIV (Balfe et al., 1990; Holmes et al., 1993; Albert et al., 1994; Leigh Brown et al., 1997; Yirrell et al., 1997; Pao et al., 2005; Brenner et al., 2007; Gifford et al., 2007; Leigh Brown et al., 2011; Ragonnet-Cronin et al., 2013; Wertheim et al., 2014), with a variety of methods for constructing phylogenies being implemented.

Though the *pol* sequences from the UK HIV DRB are relatively short (about 1,500bp), the number of sequences available for analysis in my two datasets is 1,821 (see Section 4.2 on page 90) and 8,483 (see Section 3.2 on page 67), which limits the phylogenetic methods available due to the computational resources required to anal-

yse that number of sequences. Within the last few years, two phylogenetic programs focusing on rapid analysis of large datasets have been released: FastTree (Price et al., 2009, 2010) and RAxML (Stamatakis, 2006; Stamatakis et al., 2007, 2008).

2.5.1 Maximum-Likelihood-Based Methods: FastTree and RaxML

ML-based methods work by evaluating all reasonable tree topologies by calculating the likelihood of the tree structure given the probability of finding each nucleotide in the ancestral nodes, and choosing the topology with the highest likelihood as the final tree (Edwards and Cavalli-Sforza, 1964). The methods for executing ML searches on sequence data are complex, and much work has been dedicated to finding more efficient algorithms and optimization steps to generate accurate phylogenies more quickly.

RAxML (Randomized Accelerated Maximum Likelihood) and FastTree are both ML based phylogenetic programs that are designed to handle large alignments of over one thousand sequences. RAxML implements numerous mechanisms to make topology searches and parameter optimization more efficient than previous programs (Stamatakis, 2006). FastTree, which is much faster than RAxML, uses similar methods but does fewer optimization steps. Despite the differences in optimization, the phylogenies produced by the two programs have been shown to be comparable (Liu et al., 2011). The two programs also differ in how they assign support values to branch splits in the phylogeny. FastTree only completes a ML phylogeny search, and estimates the support value of a split in the phylogeny using the Shimodaira-Hasegawa test to compare possible alternate topologies (Price et al., 2010). RAxML is able to do an ML search, then bootstrap the alignment and create bootstrapped trees, and write the frequency of each split in the bootstrapped trees onto the ML tree to create bootstrap support values.

The resulting difference in the support values generated by these two programs is visible when the values are plotted for both RAxML and FastTree (Figure 2.6 on the following page). In phylogenies constructed from the same subtype B dataset (n=8,483), FastTree is much more likely to assign high support values (>70%) to splits, while RAxML is most likely to assign either a very high (>90%) or very low (<10%) support value to splits. As using bootstrapped phylogenies to generate support values is

well-recognised as a way of ascertaining confidence in a phylogeny (Van de Peer, 2003), support values generated by FastTree using the Shimodaira-Hasegawa were ignored.

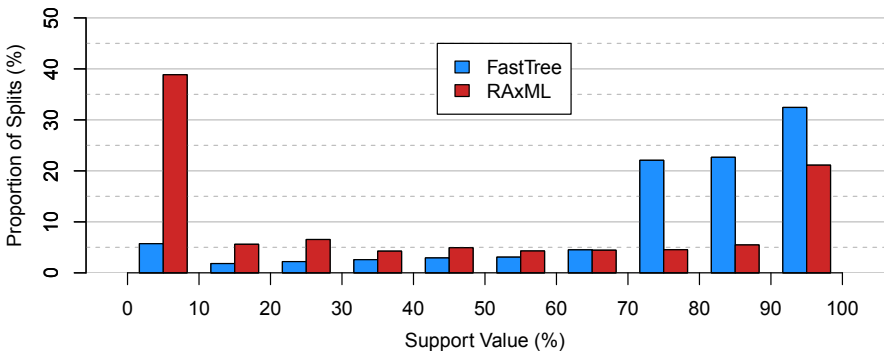

Figure 2.6: An example of the differences in the support values assigned to splits by FastTree's calculated support values (blue) and RAxML's conventionally generated bootstrap values (red). The full subtype B dataset (n=8,483) was run in both programs, with 100 bootstraps in RAxML, and the values assigned to the tips in the phylogeny were plotted. FastTree is much more likely than RAxML to assign high support to splits in the tree.

On the larger subtype B dataset (n=8,483), a full ML search and split-support calculation takes just 30 minutes in FastTree. Due to a time limit of 48 hours on the Edinburgh Compute and Data Facility computer cluster (ECDF), RAxML runs must be split into three separate stages. Generating 100 bootstrapped trees takes approximately 30 hours using the multi-threaded 'MPI' version of RAxML on 16 nodes on ECDF. The comprehensive ML search takes considerably longer – after 48 hours the run is stopped by ECDF, and must be restarted using the most recently found ML tree as a starting point to continue the ML search. In total, generating the ML tree using the multi-threaded 'PTHREADS' version of RAxML takes approximately 86 hours of processing time using 12 nodes with 2GB of memory each on ECDF. Generating the support values for the splits in the ML tree from the bootstrapped trees takes only a few minutes after the other two analyses have completed. However, none of the times quoted above include time spent queueing for available nodes on ECDF. When requesting as many as 12 or 16 nodes, it could sometimes take 4 or 5 days before the run begins during busy periods. Choosing the number of nodes to request for a run is a trade-off between the time saved due to a quicker run using more nodes and the increased time spent queueing if many nodes are requested. Through trial and error, 12-16 nodes seemed to be an appropriate number in both regards, though no formal

testing was carried out, and queue time varies greatly through the year. The subtype C dataset (n=1,821) is small enough to take significantly less time to run; using the same number and type of nodes specified above, 100 bootstrapped trees can be generated in 4 hours, and the ML search is completed in 8.

Analysis with both programs on subtype B produced analogous heritability estimates (Section 3.3.1 on page 72), but FastTree runs on subtype C failed to produce estimates significant at the Bonferroni-corrected level (Section 4.3 on page 94), though runs with RAxML did. This is possibly due to the much smaller sample size of the subtype C dataset, though could also be due to the very different histories of the two subtypes making the phylogenetic reconstruction more challenging with subtype C (Section 4.4 on page 102). Thus, though the time to produce a phylogeny with FastTree was significantly faster than RAxML, RAxML was chosen as the primary phylogenetic reconstruction method as it seemed to produce more accurate trees on smaller and potentially more complex sequence data. FastTree runs were also conducted on the final datasets in order to compare the estimates obtained from two different ML-based programs.

In hope of potentially identifying the differences that resulted in the successful detection of a heritability signal using one program but not the other, the overall tree structure produced by RAxML and FastTree on subtype C data was compared. Using visual comparison of manually coloured tips and the splits near the roots, and a specially-written Java program based on TreeCollapseCL 4 that compares the topologies of two trees in four different ways, I compared the differences between the RAxML and FastTree trees to the normal variation found between two FastTree or RAxML runs on the same data. Unfortunately, the complex nature of phylogenetic trees and the uncertainty in knowing what differences may be most important to detecting a heritability signal meant that no definitive answer was found. Interestingly, the size and number of 'clusters' (grouped sequences on a phylogeny that fall within specified genetic distance and bootstrap cut-off values) found in phylogenies constructed with FastTree and RAxML on a variety of data sizes does not seem to differ greatly (Ragonnet-Cronin, 2014). This implies that both methods are equally proficient at reconstructing structure at the tips of the trees, and that the differences that lead to

a failure to detect a heritability signal are somewhere deeper in the phylogeny. That both methods perform well in reconstructing the tips of trees may not be surprising, however, as the tips, being the most recent part of the tree, contain the most data. Nodes with at least one 'child' (a node branching from this node) that is a tip are much more likely than internal nodes (nodes where neither child is a tip) to be given a very high (>90%) support value by both RAxML and FastTree (Figure 2.7).

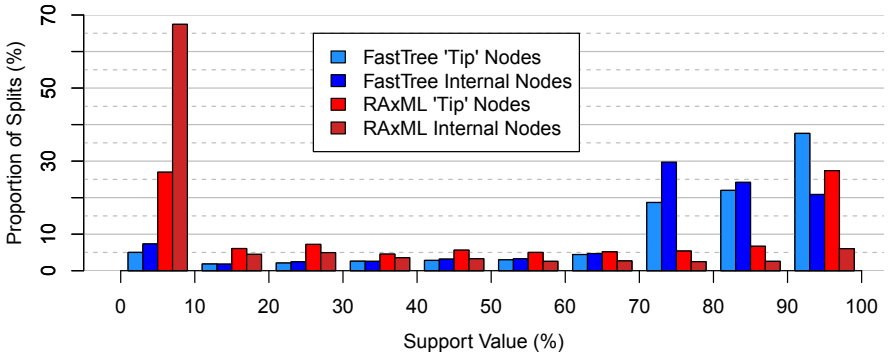

Figure 2.7: An example of the differences in the support values assigned to splits by FastTree's calculated support values (blues) and RAxML's conventionally generated bootstrap values (reds). The full subtype B dataset (n=8,483) was run in both programs, with 100 bootstraps in RAxML. 'Tip' nodes are those with at least one child is a tip (sequence), and internal nodes are those where neither child is a tip. Tip nodes are more likely to be assigned a very high support value (>90%) by both RAxML and FastTree.

2.5.2 Bayesian MCMC Methods: BEAST

BEAST (Bayesian Evolutionary Analysis by Sampling Trees) (Drummond and Rambaut, 2007; Drummond et al., 2012) is a phylogenetic program that uses Bayesian MCMC methods rather than maximum likelihood. Bayesian MCMC methods to estimate phylogenies take in 'prior' information provided by the user (e.g. a rough substitution rate, dates of the samples, sequences), then randomly 'walk' through parameter space, rejecting some 'steps' if they are not consistent with the provided information, and sample the values at set intervals of the walk. This generates a 'posterior' distribution of each parameter from the samples taken during the random walk. The most likely values of the parameters, including the topology of the tree, evolutionary rate, and effective population size, should lie in the densest part of the resulting distribution, assuming the priors were reasonable and enough sampling took place.

One of the main functions of BEAST is to provide rooted, time-scaled phylogenies (Drummond and Rambaut, 2007). This can be very useful in phylogenetic analysis, as it provides a way to examine changes in traits, population size, or transmission dynamics over time. In order to investigate change in viral load over time (full method in Section 2.7), BEAST was used to reconstruct time-resolved phylogenies. Because of the computational intensity of BEAST analysis, dataset sizes must be limited in order for the run to complete within a practical amount of time. The full details of the sub-sampled dataset run in BEAST, and the parameters of the BEAST runs can be found in Section 3.2.2 on page 70 (subtype B) and Section 4.2.2 on page 92 (subtype C).

2.5.3 Rooting and Tree Uncertainty

Rooting

Most of the phylogenies for the analyses performed on the datasets were produced using ML-based FastTree and RAxML, both of which produce 'un-rooted' trees. Phylogenies trace all the sequences back to one point of common ancestry, but in un-rooted trees, the point where all the branches converge to one ancestor is not assigned to a meaningful position in the tree. Often, it is simply assigned to the mid-point of the path between the most distant tips ('mid-point rooting,' see Figure 2.8 (a)), which produces a 'balanced-looking' tree. An alternative approach is to include an 'outgroup' – sequences that are known from other information to be distantly related to the other samples – and place the root somewhere along the branch to the outgroup ('outgroup rooting,' see Figure 2.8 (b)). As shown in Figure 2.8, the two types of rooting can produce quite different phylogenies, and these differences can be important. In (a), one could decide that sequences A and B are more closely related to C, and that the split between E and D occurred before the split between A and B. With the information that C is an outgroup (b), it's clear that A and B are more closely related to C and D, and that the split between A and B occurred before the split between C and D.

When measuring the relatedness of sequences on a phylogeny, it is imperative that the tree is arranged in a meaningful way, so that the splits and branch lengths most

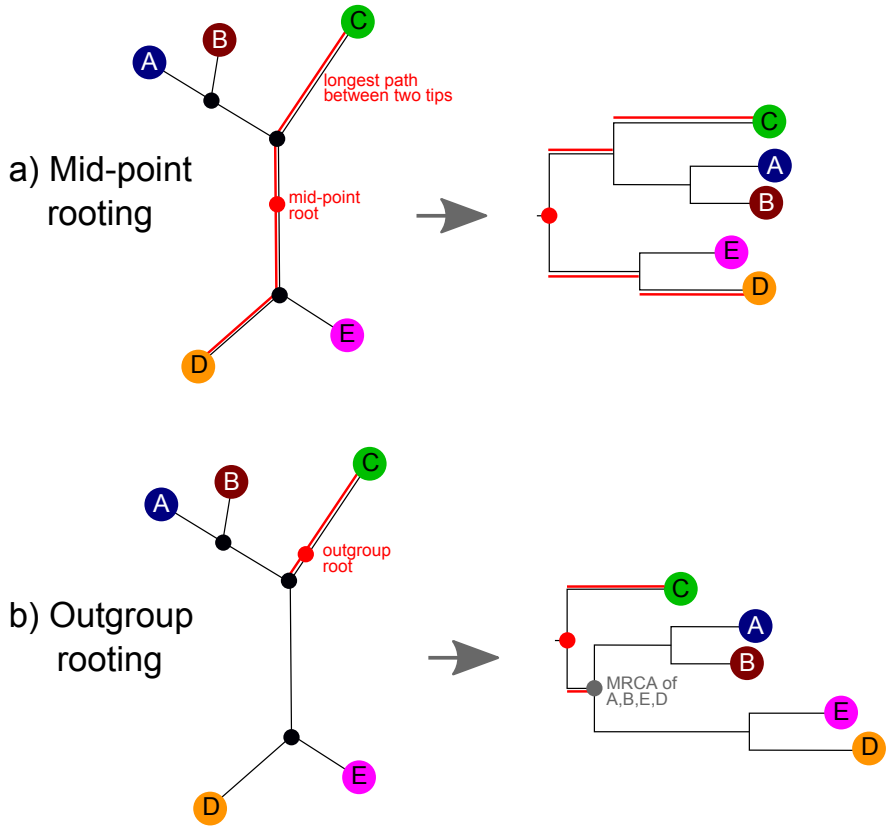

Figure 2.8: Diagram showing two different kinds of rooting for phylogenetic trees. In mid-point rooting (a), the mid-point of the path between the two most distant nodes in the tree is used as the root. In outgroup rooting (b), a point on the branch to the outgroup is used as the root. With outgroup rooting, the non-outgroup sequences' (A, B, E, D) most recent common ancestor (MRCA) is the node just above the root that leads to the non-outgroup sequences.

accurately reflect the true relationships. To accomplish this with the UK HIV DRB sequence, I included 38 subtype reference *pol* sequences (subtypes A-K) from the Los Alamos HIV Database (www.hiv.lanl.gov) to function as an outgroup. Subtype reference sequences include up to four selected genomes from each HIV subtype that are calculated by Los Alamos to be broadly representative of that subtype. As each of the non-recombinant HIV subtypes forms a distinct monophyletic clade (a cluster that includes all ancestors and descendants of one group), a group of subtypes other than the subtype being studied can be used reliably as an outgroup.

One aspect of outgroup rooting is still quite arbitrary – where along the branch to the outgroup to place the root. As shown in Figure 2.9 on the next page in (a) and (b), the point where the root is placed on the outgroup branch influences the distance each tip is from the root. Often, the root is placed halfway along the path to the

outgroup, as shown in (b) and (c). In this case, distance between the root and the MRCA of the sampled sequences is determined by the length of the outgroup branch, which depends heavily on what outgroup is used. If a more closely related outgroup is used, the distance from root to MRCA will be smaller (b). If a relatively distant outgroup is used, the distance from root to MRCA will be larger (c).

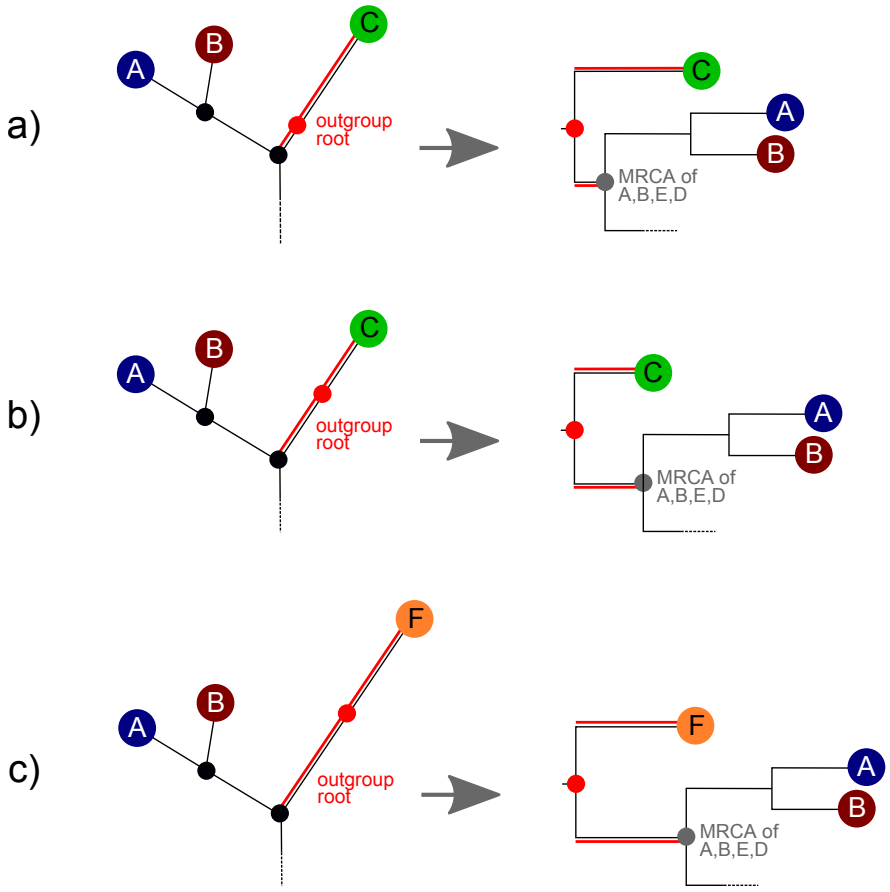

Figure 2.9: The impact of different kind of outgroup rooting on phylogenies. The point on the branch to the outgroup where the root is placed influences the distance from each tip to the root, as illustrated by the difference in using a random root placement (a) and placing the root halfway along the branch to the root (b). If the root is placed at the midpoint of the outgroup branch, the outgroup used can still influence the distance from the tips to the root. In (b) a more closely related outgroup has been used, so the tips are closer to the root, whereas in (c) a more distant outgroup has been used, so the tips are further from the root.

When using analyses that are dependant on the average length of the phylogeny (the mean distance from each tip to the root), it is important to keep in mind that the outgroup used will determine the length of the branch between the MRCA of the sampled sequences and the root, and the length is therefore somewhat arbitrary. As

an alternative, the distance from each non-outgroup tip to the MRCA (the node just before the root) can be calculated and used to find the average length of the phylogeny, excluding the influence of the outgroup.

Tree Uncertainty

As is typical for phylogenies based on population samples of HIV *pol* sequences, there is relatively little well-supported internal structure found within the phylogenies I created. The fast evolutionary rate of HIV and the relatively limited sampling often produces 'star-like' trees, with long internal branches whose relationships are difficult to resolve, and possibly affected by 'long-branch attraction,' where samples with long branches can cluster more closely than their true relationship (Felsenstein, 1978; Gribaldo and Philippe, 2002). Because estimating heritability depends on having reliable information about the relationships between sampled individuals, there was concern that spurious splits with low bootstrap support values could be providing incorrect information about the relationships between sequences, biasing heritability estimates.

In an effort to see if the estimates obtained could possibly be heavily influenced by poorly-supported splits, all splits with bootstrap support values less than 90% were collapsed using a new piece of software, TreeCollapseCL 4 (available at hiv.bio.ed.ac.uk). In most phylogenies, all splits are 'bifurcating' meaning that each node has only two children. The bootstrap support value for a node expresses what percentage of the time its two child nodes (and their child nodes, if appropriate) group together. If two child nodes group together less than 90% of the time, the node that separates them from the rest of the tree may be unreliable, and so is collapsed. While keeping the overall branch length from root to tip constant, the poorly supported node is removed, and its children are linked to the parent of the poorly-supported node (Figure 2.10 on the facing page) The parent node now has multiple children, and is called a 'polytomy.'

After collapsing poorly supported splits, the heritability estimation can be re-run. If the resulting estimate differs significantly from the original estimate, it is possible that spurious splits are influencing the heritability estimate. If there is no significant change in the estimate, poorly supported and possibly incorrect nodes in the tree are unlikely to be influencing the heritability estimate.

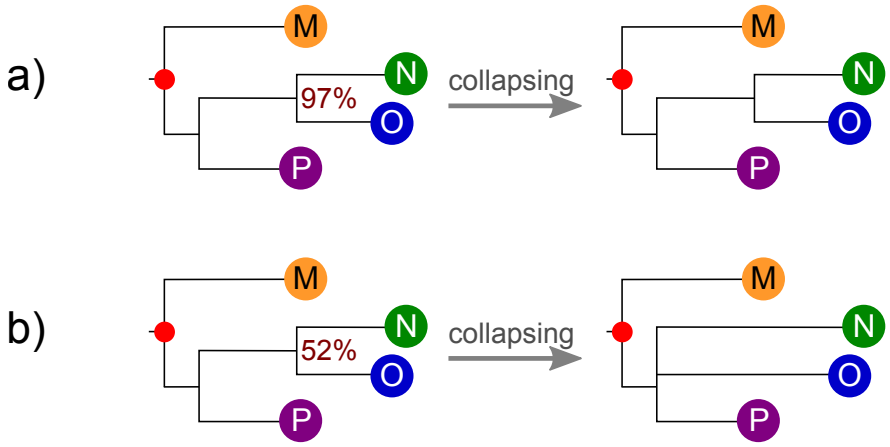

Figure 2.10: An illustration of collapsing poorly-supported nodes. In (a), N and O cluster 97% of the time. After collapsing nodes with support <90%, they remain unchanged. In (b), N and O cluster together only 52% of the time. After collapsing nodes with support <90%, the node with 52% support has been removed, and N and O now link directly back to the removed node's parent. The parental node of N, O, and P is now a 'polytomy' as it has more than two children. Note that after collapsing, the total distance from each tip back to the root (red) is preserved.

The bootstrapped trees generated to create bootstrap support values can also be used to further investigate the effect of uncertainty in the tree. As bootstrapping randomly samples nucleotide columns from the original alignment, we expect the resulting bootstrapped trees to have a distribution around the true relationship between the sequences. If most or all of the bootstrapped trees give heritability estimates similar to the one produced by the original tree, this provides evidence that the method is robust to a degree of uncertainty in the phylogeny, and thus is not affected by spurious splits.

Finally, to ensure the validity of the estimates obtained, each analysis was performed in duplicate or more, with two trees independently created with RAxML, and ten independent phylogenies generated using FastTree. The significance thresholds for these runs were Bonferroni corrected. Due to the computational intensity of BEAST runs, only one phylogeny was produced.

2.6 Heritability Estimation Pipeline

A 'pipeline' of code was written in R to automate the process of correctly processing the phylogenies, creating the files needed to run in ASReml, running the analysis and checking the run, and processing the output. This code can run many phylogenies at

once: automatically running both with and without poorly-supported nodes collapsed, ensuring unique file names, Bonferroni-correcting significance estimates, and generating human-readable output and summaries in one file.

After phylogenies were constructed using RAxML, a new piece of software, TreeCollapseCL 4, was developed to aid in preparing the phylogenies for further analysis. Using TreeCollapseCL 4, each phylogeny was rooted using the outgroup sequences, and the average length of the tree was calculated from the tips to the MRCA of the UK HIV DRB sequences (the second node from the root). Branch length to the root was not calculated because the distance from the root to the MRCA of the UK HIV DRB sequences can be severely affected by the choice of outgroup used (see Section 2.5.3 on page 45 and Figure 2.9 on page 47).

The viral sequences and all of the internal nodes of the phylogeny were incorporated into a genetic relatedness matrix from which the inverse was calculated using the R (R Development Core Team, 2011) package MCMCglmm (Hadfield, 2010). The phylogenetic covariance of two individuals on a phylogeny is assumed to be proportional to the distance between their MRCA and the root (Felsenstein, 1985). Thus the covariance of an individual with itself is its distance from the root in units of substitutions per site per year. In phylogenetic comparative methods that use 'ultrametric' trees, where the distance from root to tip is the same for every node, the distance between the tips and the root is often rescaled to one unit. Although the units are arbitrary, the variance explained by the phylogeny is directly interpretable as the variance explained in the sample of individuals used in the analysis. However, when trees are not ultrametric, as in the case of the UK HIV DRB sequences, the root-to-tip distances vary. In order to scale the heritability estimate to be directly interpretable as the variance explained over the length of the phylogeny created from the samples, I standardized the heritability estimate obtained from ASReml by multiplying it by the average distance from MRCA to tip for each tree, as calculated by TreeCollapseCL 4.

2.6.1 Running Phylogenies in ASReml

MCMCglmm was used to export the inverse genetic relatedness matrix in file formats readable by ASReml. The '.ped' (pedigree) input file in ASReml usually contains one

line for each individual, with its 'sire' and 'dam' (male and female parent). For my phylogenies, this becomes a file with one line for each node (including internal nodes), with each nodes' parental node listed as the 'sire' and no entry for the 'dam,' becoming a list representation of the topology of the tree. The '.giv' (general inverse variance) input file describes a sparse matrix of the genetic relationships between individuals and for each individual with itself. Using phylogenies this remains essentially the same, except that the relationships are calculated using the topology and branch length of the phylogeny, rather than pedigree information.

After the generation of the .giv and .ped files, an '.as' file is created for each phylogeny. The .as file is the main ASReml command file, which introduces the variables to be used in the model and their format (e.g. date, character, number), gives the location of the corresponding '.giv,' '.ped,' and demographic data files, and describes the model to be run. An example '.as' file is given in Appendix B (Code B.1).

The ASReml Model: Fixed and Random Effects

Preliminary runs were carried out on the subtype B dataset in ASReml in order to identify fixed effects to include in the final model. Age at the sample date taken for set-point viral load, sex, ethnicity, time from HIV diagnosis to the set-point viral load sample date, subtype, exposure risk group, and year of HIV diagnosis (as a continuous variable) were included in the preliminary models. This was all of the data made available by the UK HIV DRB and UK CHIC, with the exception of first and last $CD4^+$ count, which were not included as $CD4^+$ count is linked to viral load, and the variance in $CD4^+$ count would likely explain much of the variance in viral load. All of the terms except subtype and exposure risk group were found to be highly significant ($p<0.001$) and therefore were included in the final model. When preliminary runs were carried out on the subtype C dataset, the same terms were found to be significant, with the exception of ethnicity. As there is currently no evidence for why ethnicity might influence viral load in subtype B but not subtype C, it is possible the lack of significance is due to the much smaller sample size in subtype C, rather than a real effect. As including ethnicity might help absorb variance that might otherwise be incorrectly partitioned to other effects, ethnicity was also included in the final subtype

C ASReml model.

Country of origin and year of HIV diagnosis (as a categorical variable) were included as random effects along with the phylogeny (in the form of the inverse genetic relationship matrix). Year of HIV diagnosis was included as a continuous fixed effect to model the linear change in set-point viral load and as a categorical random effect to account for any random deviations around this trend from year to year. As country of origin has many discrete levels (>70), it was included as a random effect in order to estimate the variance of their effects.

Post-Processing: Getting h^2

The R code pipeline also automatically creates a '.pin' file for each run. The '.pin' file contains post-analysis processing that uses the variance components generated by the main analysis to calculate genetic and phenotypic variance and heritability. An example '.pin' file is given in Appendix B (Code B.2).

The '.pin' file reads in all the variance components from the ASReml run; in this case, the variance in viral load explained by categorical year of HIV diagnosis, country of origin, the phylogeny, and the residual variance. To get the total phenotypic variance (V_P), these values are summed. As discussed earlier (Section 2.6 on page 49), in order to ensure that the estimate of the heritability is interpretable as being the variance explained by the phylogeny in the sample of individuals used, the genetic variance (V_A) estimate is standardized by the average root-to-tip distance of the phylogeny. To do this, the average root-to-tip distance calculated by TreeCollapseCL 4 is included in the '.pin' file when it is created by the R pipeline, and multiplied by the V_A estimate. A new V_P can then be calculated using the standardized V_A, and the ratio of the standardized V_A to the new V_P is used to calculate the standardized heritability estimate. After calculation, all of these values are printed to an output file with their standard errors.

Convergence and Significance

Each ASReml run was initially allowed to run for 100 iterations. Though very few runs failed to converge during this time, the pipeline R code checks each run to ensure convergence, and if convergence has not been reached, automatically restarts the run,

allowing unlimited iterations. The few runs that did not converge within 100 iterations did so very quickly when restarted.

The significance of the effect of the phylogeny in explaining the variance in viral load was assessed by first running a model without the phylogeny, as a 'null' model, and then including the phylogeny and running again. A log-likelihood ratio test with one degree of freedom was then used to test whether the model with the phylogeny was significantly better at explaining the variation in viral load than the null model.

To further check that the heritability estimates given were not spurious, the R pipeline creates a parallel run for each phylogeny where the tips of the phylogeny are randomly permuted, then the analysis is performed as normal, with significance tested against a model without a phylogeny. None of these runs gave a significant result, and the heritability estimates for these runs was always zero, or very close to zero (<0.1%).

2.7 Change in Set-Point Viral Load Over Time

In order to look at the change in set-point viral load due to selection, I examined the total effect of within-host and between-lineage selection (with assistance from Jarrod Hadfield). Within-host selection occurs when variation in set-point viral load determines the relative frequency of the genotype within a host (Figure 2.11 (b)). The within-host change can be estimated in longitudinal data by fitting sequence sample date as a covariate in the model, which was done using the MCMCglmm package. However, any directional change due to environmental factors not controlled for in the model that influence viral load, such as the background level of ART in the population, could give rise to identical patterns, and currently there is no way to distinguish between the two.

Between-lineage selection happens when variation in set-point viral load determines the probability of transmission (speciation) and host death (extinction) (Figure 2.11 (a)). In this context, it is known that evolutionary change can be estimated by taking the difference in the means of predicted breeding values (the equivalent of phylogenetic effects) over time (Walsh and Lynch, 2012). Markov chain Monte Carlo methods can be used to average over the uncertainty in the heritability estimates and the predicted

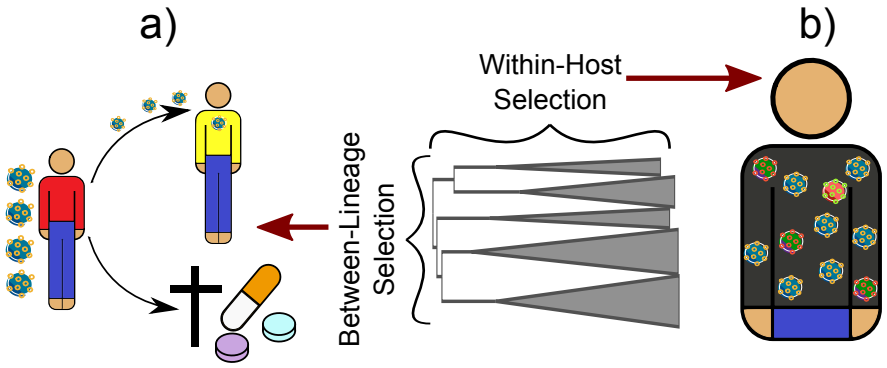

Figure 2.11: Illustration of between-lineage and within-host selection. In between-lineage selection (a), the set-point viral load defines the probability of transmission and/or time to death or treatment. In within-host selection (b), the set-point viral load determines the relative frequency of a viral genotype within the host.

breeding values in order to derive the posterior distribution of evolutionary change (Sorensen et al., 1994). Posterior predictive tests can then determine whether evolutionary change has occurred and whether the change is greater than would be expected by chance (drift) (Hadfield et al., 2010). Here, I implemented an equivalent model on the phylogeny.

It is known that selection estimates rely on any missing data being missing at random and not dependent on the value of the data, such as when the trait value determines survival probability and thus sampling (Rubin, 1976; Im et al., 1989). Because the data will not be missing at random if speciation and/or extinction is dependent on viral load, which is the case in HIV, then the method may be expected to give biased estimates of evolutionary change. In addition, phenotyping of a pedigree is often comprehensive in comparison to the phenotyping of species in a phylogeny, where all ancestral taxa usually have missing data. Although more appropriate methods exist for this type of problem (FitzJohn, 2010), the size and complexity of the data set here prohibit their use. To gauge the magnitude of the problem, Jarrod Hadfield simulated data under a model of speciation and extinction with trait-dependant rates using the make.quasse function in the Diversitree R package (FitzJohn, 2012).

JH simulated 500 100-tip trees with the probability of speciation or extinction either being a constant or depending on the trait through a linear model on the logit scale. The two parameters of the linear model were an intercept and a slope on trait value. In the first set of simulations, extinction probability was set to zero, and speciation

Table 2.1: Means and standard errors of observed and estimated evolutionary change for 500 data sets simulated under models of trait-mediated speciation and extinction. Estimate I is the estimate made when only the phenotypic data of extant taxa are observed, and estimate II is the estimate made when the phenotypic data of all ancestral nodes of extant taxa are also observed.

	Speciation	Extinction
Actual Change	0.359+-0.016	-0.352+-0.011
Average Estimate I	0.295+-0.011	-0.053+-0.006
Average Estimate II	0.377+-0.014	-0.104+-0.009

modelled with an intercept of zero and slope of 0.75. In the second set of simulations, speciation probability was set to one, and extinction modelled with an intercept of -2 and slope of 5. The rate of drift (proportional to the phylogenetic variance) was set to 0.1, and independent random normal deviation with a variance of 0.05 added to each character. The models of speciation and extinction depend on the trait value before adding the random normal deviates (i.e. they depend on the phylogenetic effects only). The magnitudes of evolutionary change in the two sets of simulations were considerably larger than the rate of drift and roughly equal to each other (although opposing in sign).

The method was capable of detecting evolutionary change under a trait-mediated speciation model, but the magnitude of the change was underestimated, with an average change across simulations of 0.359 but an average estimate of 0.295 (i.e. 82% the true value). When between-lineage evolutionary change was mediated by differential extinction the average change across simulations was -0.352 but the average estimate was -0.053 (i.e. 15% the true value). The results, together with those of analyses that included the missing phenotypes of taxa ancestral to the extant taxa (i.e. the missing data of the speciation only model) are presented in Table 2.1.

The simulations suggest that although power to detect evolutionary change via differential speciation was relatively good and the downward bias in the estimate of the magnitude of evolutionary change not too severe, any evolutionary change caused by differential extinction is hard to detect and its magnitude considerably underestimated. However, in HIV differential extinction and differential transmissions are not easy to distinguish as viral load has a positive effect on transmission probability and a negative effect on infection duration and therefore potential for transmission (Fraser et al., 2007).

What Does Changing Virulence in HIV Mean?

It is worth noting here some disagreement in how the term 'virulence' should be used with regard to HIV. Previous studies investigating changes in viral load over time have looked at the overall clinical measure of viral load, as influenced by many host, environmental, and demographic factors, which are never completely controlled for (Dorrucci et al., 2007; Herbeck et al., 2012; Pantazis et al., 2014), and concluded that the detected rise in viral load indicates an increase in virulence for HIV, as higher viral loads will lead to faster disease progression and a higher transmission risk. Similarly, in my own analyses to estimate the heritability of viral load, I fitted year of HIV diagnosis as a fixed effect in the model (see Section 2.6.1), to estimate the change in viral load due to year of diagnosis (see Tables 3.4 and 4.4). As in previous studies, this term estimates the overall change in viral load over time (though less exactly than a study designed to investigate this change specifically), which many be due to many causes, including demographic and environmental effects as well as changes due to the virus itself, and may not be the same as the change in viral load due to viral genetics alone.

In the analysis I have described in this section, I have separated out the change in viral load over time due to the viral genetics from change due to other factors, and found the viral influence on viral load has lead to a small decline in viral load over time (see Section 3.3.3 on page 75), which I have interpreted as evidence that the virulence of HIV has not increased, as the viral component influencing viral load has not lead to increasing viral load. These two conclusions about whether viral load has changed over time are not irreconcilable, as the viral component of viral load could cause a decrease in viral load, while other, non-viral factors could cause a larger increase in viral load, leading to an overall increase in viral load (see Section 3.4.4 on page 80 for further discussion).

However, this does cause disagreement in whether an overall change in viral load should be interpreted as a change in the virulence of HIV, or whether only a change in viral load due to the viral genetics should be interpreted as a change in HIV virulence. In my research here, I determine that the decline in viral load due to the viral genetics indicates that HIV, as a virus, has not become more virulent. I believe this use of

'virulence' is more accurate, as while it may be correct to say that any change that influences an organism's ability to cause disease and transmit itself is a change in virulence, in the context of HIV 'virulence' is often strongly linked to the idea of genetically distinct strains of HIV that differ in their ability to transmit themselves and cause AIDS (Palm et al., 2013; Payne et al., 2014; Kouri et al., 2015). In other words, changes in 'virulence' are often implicitly attributed to the virus itself, when this is not what has been explicitly tested, and when in fact the change could be due to many other factors that are difficult to control for.

This slight difference in using the term 'virulence' when non-viral factors are the cause of a change in viral load can be examined by considering that though HIV-infected men usually have higher viral loads than women (Farzadegan et al., 1998; Sterling et al., 1999; Gandhi et al., 2002), it would seem strange to say that HIV is more virulent in men. Though obviously still controversial, in my own results in investigating the change in viral load over time (Section 3.4.4 on page 80), I have interpreted 'change in virulence' to mean change due to the virus itself, a distinction not made by most previous papers using the term 'virulence.'

A

Example Graphs of Rules Used to Choose Set-Point Viral Loads

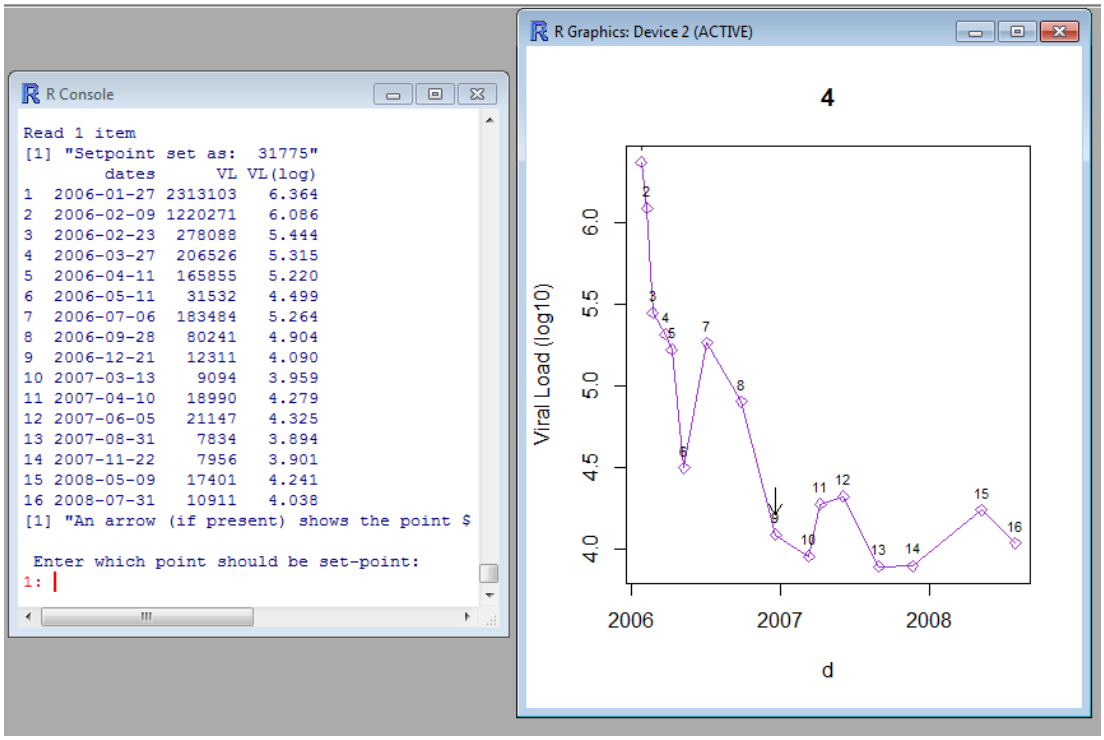

Figure A.1: This screen-shot shows an example of how set-point viral load was inspected manually. R code was written so that for all patients with a very high initial viral load, and more than three viral loads available, all viral loads were plotted. Last viral load measure taken within a year of the first is marked by a blue arrow. All viral load measures are numbered, and the date, viral load, and $\log_{10}$ viral load are printed to the screen. The code then asks which viral load should be taken as the set-point. The number corresponding to the viral load is typed in, and the code confirms this choice by colouring the graph differently past this point. The code records the chosen set-point value, and displays the next patient's viral load graph.

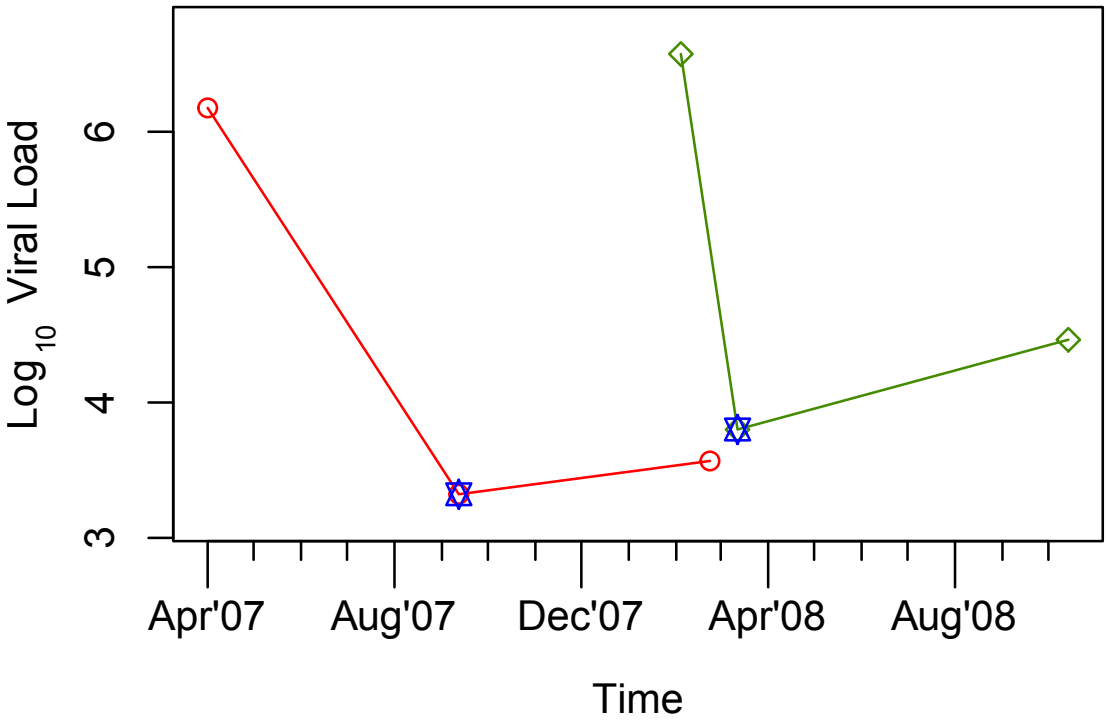

Figure A.2: An example plot of two patients with very high initial viral loads, and three viral load measures available. In both of these cases, the middle viral load is the lowest. This is taken to be an 'inflection point' at the end of the acute phase, and is taken as the set-point viral load, marked with a blue star.

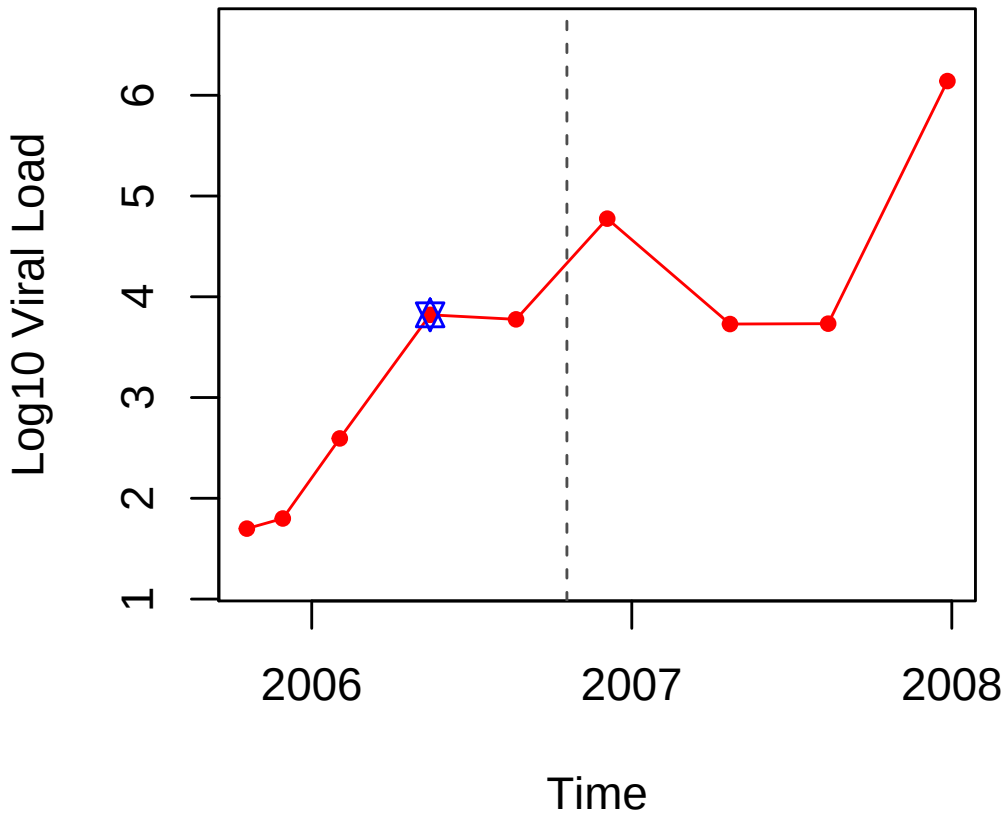

Figure A.3: An example plot of a patient with a very low first viral load and more than three viral loads available. The dashed grey line indicates a year after the first viral load. Within one year after the first viral load, there was a 2 log increase between the first viral load and the fourth viral load, possibly indicating a patient going off ART. The value after this increase is used as the set-point viral load, marked with a blue star.

B

Example ASReml Files

Code B.1: An example of the .as file used to specify the model to be run in ASReml

```
ASREML HIV Trial
 id !I
 Subtype !A
 age
 sexid !A
 hivposYr
 hivposYr2 !A
 hivposAgeDay
 exposureid !A
 ethnicityid !A
 countryid !A
 rndID !P
 setpt

hivc-1.ped !ALPHA
hivc-1.giv
hivc-1.csv !SKIP 1 !MVREMOVE  !MAXIT 100

setpt ~ mu hivposAgeDay hivposYr sexid age ethnicityid !r giv(rndID,1)
    countryid hivposYr2
```

Lines 2-13 describe all the variables that are included in the '.csv' file that contains the demographic and clinical information for each sample. They are listed in the same order as the '.csv' file, and are qualified with the data type. Values without modifiers are read in as continuous numerical variables. '!I' indicates the value should be read in as an integer that does not run from 1...n. '!A' indicates alphanumerical values, and the resulting data is treated as a factor. For example, the names of countries are essentially meaningless, but ASReml should recognise that two individuals with the same country value are from the same 'group.' '!DATE' indicates the value should be read in as a date, and '!P' indicates the value that should be used to link the information in the '.csv' data file to the '.giv' and '.ped' file (here, the sequence ID). Not all information read in must be used in the model. It also worth noting that 'hivposYr' (year of diagnosis) is given twice in the '.csv' file so that it can be read in twice: once as a continuous variable (hivposYr) and once as a factor (hivposYr2). (See Chapter 2.6.1 on page 51 for why this is done.)

Lines 15-17 give the names of the accompanying data files, the '.ped', '.giv', and '.csv' files. Qualifying the '.ped' file with '!ALPHA' indicates that some identities may not be numerical (in this case, the subtype reference files have alphanumeric names). The '.csv' file is qualified with '!SKIP 1' which means to skip the first line (the column names), '!MVREMOVE' indicating removal of any individuals where data used in the model is missing, and '!MAXIT 100' sets the maximum number of iterations to 100. Some ASReml runs may not converge in 100 iterations, and the R code in the pipeline checks for this. If any run has not converged, it is restarted, and allowed unlimited iterations. Very few runs failed to converge in 100 iterations, and those that did converged very quickly when allowed unlimited iterations.

Finally, line 19 gives the mixed model to be run, modelling set-point (setpt) as a function of the fixed and random effect included. 'mu' is the intercept, and hivposAgeDay, hivposYr, sexid, age, and ethnicityid are the fixed effects. '!r' indicates the variables following should be random effects: 'giv(rndID,1),' which is the genetic relatedness matrix (the information from the phylogeny), countryid, and hivposYr2. See Chapter 2.6.1 on page 51 for a full description of the fixed and random effects.

Code B.2: An example of the .pin file used to calculate components of variance in ASReml

```
1  #Post-processing file to estimate heritability
2  #Variance components in .asr (results) file are :
3  #1) hivposYr2
4  #2) countryid
5  #3) Genetic (giv(rndID,1)) (Va)
6  #4) Residual (Vr)
7
8  F  Vp   1+2+3+4 #Gives total phenotypic variance (becomes component 5)
9  F  Va   3    #Rewrites Va with SE (becomes component 6)
10 F  Vr   4    #Rewrites Vr with SE (becomes component 7)
11 F  Vc   6*0.149725774035225  #Va corrected by avg branch length (becomes
       component 8)
12 F  Vp   8+7+1+2   #Gives total new Vp with corrected Va (becomes component 9)
13
14 H  h2   3 5  #Ratio of (original) Va to (original) Vp (i.e. uncorrected
       heritability) with SE
15 H  r2   4 5  #Ratio of Vr to (original) Vp
16 H  hc   8 9  #Ratio of (corrected) Va (Vc) to (corrected) Vp (i.e. corrected
       heritability) with SE
```

'.pin' files are used to calculate variance components from the results of the main ASReml analysis. Comments are denoted with '#' symbols, and what follows them is not processed. Here, comments are used to explain the '.pin' file. The variance in the trait explained by the random effects (here, year of HIV diagnosis, country of origin, and the phylogeny) of the model and the residual variance are given numbers. As shown in lines 3-6, the variance due to year of HIV diagnosis is number 1, the variance due to country of origin is number 2, and so on.

Starting a line with 'F' allows linear combination of variance components (addition, subtraction, multiplication), and creates a new variance component with a number to store the results. For example, on line 8, the four variance components from the main analysis are summed, and the result becomes component 5. The letters after 'F' indicate a 'label' for the new result. This is used in the output file, so that the user can more easily identify the numbers resulting from the post-processing. The result of each 'F' line is printed to the output file, with the standard error. Because the initial four variance components are not printed with standard error, 'F' lines can also be used to force ASReml to simply print the estimate with a standard error (as in lines 9 and 10).

Starting a line with 'H' forms the ratio of two components, as when calculating heritability. Again, the letters following the 'H' denote a 'label' which is used in the

output file. The first component number is used as the numerator of the ratio, and the second as the denominator, and the resulting number is printed to the output with its standard error. The results of calculations performed on 'H' lines does not create a new numbered variance component.

"But all evolutionary biologists know that variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions."

Stephen Jay Gould - 'The Median Isn't the Message' (1982)

3 Virulence in Subtype B

3.1 Introduction: Subtype B in the UK

It was in subtype B where differences in disease progression over time were first reported, leading to speculation that HIV could be becoming more virulent (Hutchinson et al., 1991; Weiss et al., 1992; Gorham et al., 1993; Holmberg et al., 1995). Of the seven previous papers that have estimated the heritability of set-point viral load, three have had a majority of subtype B-infected individuals (Hecht et al., 2010; van der Kuyl et al., 2010; Alizon et al., 2010), more than any other subtype.

As the oldest and most prevalent subtype in the UK, subtype B sequences make up the majority of the UK HIV DRB, with 22,507 first sequences available (52.3% of first sequences). One of the concerns about previous studies' estimates is the possibility that datasets were subject to narrow selection criteria, and one of the strengths of my phylogenetic method is the ability to take in large and complex datasets. This, along with the ability to directly compare my results with three previous studies on the heritability of viral load made subtype B the ideal first dataset to use in my analysis.

3.2 Methods: Subtype B Data from the HIV DRB

8,700 initial subtype B sequences from the UK HIV RDB had viral load measures before starting ART available from UK CHIC (see Methods Section 2.2.2 on page 30).

Table 3.1: Number of Subtype B Sequences Discarded During Data Cleaning

Number of Subtype B sequences with at least 1 viral load measure before ART			8,700
1.	(a) Potential acute-stage viral load	-12	
	(b) Potential acute-stage viral load (measurement limit)	-21	
2.	(a) Viral load potentially taken on ART	-26	
	(b) Viral load potentially taken on ART (measurement limit)	-3	
3.	Suspected database error	-1	
	Acute stage or AIDS patient starting unreported ART	-8	
	Missing RT or protease	-20	
	Missing 200bp in RT	-7	
	Identical sequences removed	-119	
Final number of records and sequences			8,483

As described in the methods, the sequences were aligned, rules were applied to exclude viral loads taken during the acute phase, AIDS, or on unreported ART, and set-point viral load was chosen (see Introduction Section 1.4 on page 18 for issues around choosing 'set-point' viral load values, and Methods Section 2.3 on page 32 for details of how set-point viral load was chosen).

As outlined in Methods Section 2.3.2, when only one viral load was available and was exceptionally high or low, $CD4^+$ count was used to exclude patients suspected of being in acute stage, having progressed to AIDS, or being on unreported ART when the viral load was taken. A few patients were also excluded as they had very high viral loads that dropped to very low viral loads in a short space of time, implying a transition from the acute stage or AIDS to unreported ART. Samples were also excluded if large portions of the sequences were missing, or if the sequences were identical to other sequences (see Section 2.4 on page 39). A total of 217 sequences and data were excluded from the subtype B analysis due to these rules; the number excluded due to each rule can be found in Table 3.1 (the numbers correspond to the rule list in Methods Section 2.3.2 on page 36).

After applying the exclusion rules, 8,483 subtype B sequences with matched viral loads remained. 80% of the patients included in the dataset had the viral load used as set-point taken within three years of HIV diagnosis. More information about the

Table 3.2: Median, quartiles, and range of HIV diagnosis date, set-point viral load test date, and the number of days between HIV diagnosis and set-point viral load testing.

	1st Quartile	Median	3rd Quartile	Range
HIV Diagnosis	5-Oct-1998	17-Jun-2003	27-Apr-2006	1-Jan-1980 to 6-May-2009
Date of Set-point Viral Load Test	7-Oct-1999	22-Dec-2003	14-Aug-2006	29-Jul-1986 to 13-May-2009
Days between HIV Diagnosis & Set-Point Viral Load Test	3	20	245	-3,333 to 8,181

dates of HIV diagnosis and set-point viral load tests is available in Table 3.2, and a histogram showing dates of HIV diagnosis and set-point viral load testing is shown in Figure 3.1. Before analysis, set-point viral load values were $\log_{10}$ transformed to make the distribution approximately normal, and sequences were stripped of codons in positions associated with drug-resistance mutations (Rhee et al., 2003; Liu and Shafer, 2006) (see Methods Section 2.4). The analysis was repeated on sequences not stripped of resistance-associated codons, but no significant differences in heritability estimates were observed (Table 3.5).

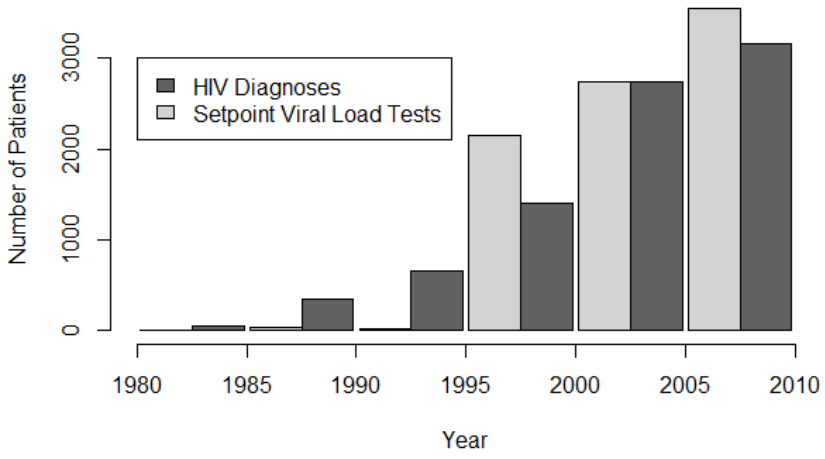

Figure 3.1: Histogram of the dates of HIV diagnosis (dark grey) and set-point viral load tests (light-grey) in the subtype B dataset.

As detailed in the methods (see Methods Section 2.5.3), to provide an unbiased root for the tree, 38 subtype reference *pol* sequences (subtypes A-K) from the Los Alamos HIV Database (www.hiv.lanl.gov) were used as an outgroup, and a boot-

strapped maximum-likelihood phylogeny was created using RAxML (Stamatakis, 2006; Stamatakis et al., 2007). Ten phylogenies were also created with the program FastTree (Price et al., 2009, 2010) in order to compare the results obtained using a different maximum-likelihood based method.

TreeCollapseCL 4 was used to root each phylogeny and to calculate the average length of the tree from tips to the most recent common ancestor (MRCA) of the UK sequences in the dataset (the second node from the root), and this was used to standardize the results obtained from ASReml (see Methods Section 2.6). For the main subtype B analysis, the average distance from root to tip was 0.14 substitutions per site per year as calculated by TreeCollapseCL 4. For all analyses, average root-to-tip distance was calculated and used for each phylogeny.

3.2.1 Choosing Fixed and Random Effects

Preliminary runs were carried out on the dataset in ASReml (Gilmour et al., 2009) in order to identify the fixed effects to include in the final model (see Methods Chapter 2.6.1 on page 51). Age at the sample date taken for set-point viral load, sex, ethnicity, time from HIV diagnosis to the set-point viral load sample date, subtype, exposure risk group, and year of HIV diagnosis (as a continuous effect) were included in the preliminary models. All of the terms except subtype and exposure risk group were found to be highly significant ($p<0.001$) and therefore were included in the final model. Country of origin and year of HIV diagnosis (as a categorical effect) were also included as random effects along with the phylogeny. Year of HIV diagnosis was included as a continuous fixed effect to model the linear change in set-point viral load and as a categorical random effect to account for any random deviations around this trend from year to year. As country of origin has many discrete levels, it was included as a random effect in order to estimate the variance of their effects.

3.2.2 Analysis Strategy

The significance of the effect of the phylogeny in explaining the variance was assessed by first running the model without the phylogeny as a 'null' model and then including the phylogeny. A log-likelihood ratio test with one degree of freedom was then used to

test whether the model with the phylogeny was significantly better at explaining the variation in viral load than the null model.

As ASReml assumes all pedigree information provided is correct, the original analysis was repeated using TreeCollapseCL 4 to collapse splits with bootstrap-support values less than 90% down to polytomies (see Methods Section 2.5.3). To further evaluate how uncertainty in the tree could affect the heritability estimates, one hundred bootstrapped trees were generated in RAxML and analysed. Because of the close phylogenetic relationships between subtypes B and D in the *pol* region, bootstrapped subtype D sequences can sometimes cluster within the B clade, making it necessary to remove the subtype D Los Alamos outgroup sequences from the phylogeny in order to root all 100 trees by the same outgroup.

Each analysis was performed in duplicate, with the sequences being run through RAxML and the analysis pipeline twice, except when using FastTree, where ten phylogenies were created and run through the pipeline, and in BEAST, where only one time-scaled phylogeny was created. The significance threshold used was adjusted using a Bonferroni correction for the number of replicates.

In addition, to further investigate the phylogenetic effects on viral load, a time-scaled phylogeny was produced using BEAST, a Bayesian phylogenetic program (Drummond et al., 2012) (see Methods Section 2.5.2). Because the complexity of the analysis performed by BEAST limits the number of samples which can be run in a reasonable time-frame, a sub-sample of the main dataset was used. After collapsing nodes with bootstrap support values less than 90%, 965 sequences remained in un-collapsed clusters of fifteen or more sequences. A random subsample of 652 of these sequences was taken for analysis in BEAST to improve the chance of convergence. BEAST was run with a relaxed log-normal clock and a constant population size for 100,000,000 steps, sampling every 10,000 steps. The SRD06 nucleotide substitution model was used, which consists of the HKY substitution model with four categories of site heterogeneity (γ+4) and partitions the first and second codon positions from the third to parametrize separately. The run was performed in duplicate, and after 10% burn-in was removed the two resulting files were combined using an in-house script. A summary tree was then generated using the BEAST program TreeAnnotator, and run in

Table 3.3: Demographics of Patients whose Samples were Analysed

		Subtype B (*n*=8,483)
Age at Set-point (years) (mean, range):		35.4 (15 - 83)
Log_{10} Set-point Viral Load (copies/mL) (mean, SD):		4.493 ±0.86
Sex	Female:	464 *(5.5%)*
	Male:	8019 *(94.5%)*
Risk Group	Homo/Bisexual:	7278 *(85.8%)*
	Heterosexual:	711 *(8.4%)*
	IDU:	239 *(2.8%)*
	Other/Unknown:	255 *(3.0%)*
Ethnicity	White:	6990 *(82.4%)*
	Black:	597 *(7.0%)*
	Asian:	221 *(2.6%)*
	Other/Unknown:	675 *(8.0%)*

the ASReml pipeline to obtain heritability estimates. Using the time-scaled phylogeny from BEAST, the amount of change in viral load due to selection was estimated using Markov Chain Monte Carlo methods to calculate the total contribution of between-lineage and within-host selection, though change due to within-host selection cannot be distinguished from environmental factors (see Methods Section 2.7).

3.3 Results: Initial Findings

3.3.1 Estimates of the Effect on Viral Load

After removing all cases where there was uncertainty over disease or treatment status or large sections of the sequence were missing, 8,483 subtype B sequences and associated viral load measurements remained. 8,253 (97.3%) records had a known risk group, 8,324 (98.1%) had a known ethnicity, and 8,114 (95.7%) had information on both risk group and ethnicity. The demographics of the dataset show that of those with known ethnicity and risk group 76% (6,198 individuals) were white MSM, reflecting the historical preponderance of this subtype among MSM (Table 3.3).

Preliminary runs in ASReml were used to determine the fixed and random effects for the model. Sex, ethnicity, country of origin, age when the set-point viral load was taken, year of HIV diagnosis, and time from HIV diagnosis to the date when set-point viral load was taken were all included in the final model (all effect estimates given in

Table 3.4: Mean Fixed Effect Estimates of $\log_{10}$ Set-point Viral Load Influence

Effect	**Units**	**Estimate**	**Standard Error**
Intercept		4.508	0.117
Age at viral load test*	per year	8.48E-3	1.13E-3
Year of HIV Diagnosis*	per year	-4.00E-3	3.46E-3
Time from HIV Diagnosis to VL test*	per day	-6.18E-5	1.49E-5
Sex	Male	0	0
	Female	-0.200	0.044
Ethnicity	White	0	0
	Black-African	-0.260	0.075
	Black-Caribbean	-0.165	0.053
	Black-other	-0.127	0.073
	Asian/Oriental	-0.038	0.076
	Indian/Pakistani/Bangladeshi	-0.136	0.093
	Other/Mixed	-0.142	0.039
	Other	-0.267	0.161
	Not known	-0.171	0.079

*Values were means-adjusted before analysis

Table 3.4). Set-point viral load was found to increase with age, but decrease with a more recent year of diagnosis and with a longer time period between HIV diagnosis and viral load testing. HIV-positive females and non-white individuals were found to have decreased set-point viral load measures compared to males and white individuals. The total variance in set-point viral load (V_P) was 0.736 (SE 0.0124). The variance explained by the random effects year of HIV diagnosis and country of origin was estimated at 3.11×10^{-3} and 6.55×10^{-4}, respectively. The variance explained by the viral genome (V_A) was 0.0418 (SE 0.0113), and the residual variance was estimated at 0.690 (SE 0.0125).

Bootstrapped phylogenetic trees were reconstructed in duplicate on the 8,483 sequences using RAxML, and both trees were analysed independently with ASReml. Using the comparison of the resulting log-likelihood values from running the model with and without the tree to estimate significance, both replicates produced highly significant ($p<0.0001$) heritability estimates of 5.8% (CI 2.9–8.7%) and 5.6% (CI 2.6–8.5%; Table 3.5). The analysis was also repeated on ten replicates in FastTree, all of which

Table 3.5: Estimate of Viral Genetic Influence on Set-Point Viral Load in HIV Subtype B in the UK

Dataset	**Method**	***N***	**Replicate**	**Viral h^2**	**Standard Error**	**Sig.[1]**
Full dataset *(Resistance -codons-stripped)*	RAxML	8,483	1	5.8%	0.0148	***
			2	5.6%	0.0151	***
Full dataset *(Resistance -codons-included)*	RAxML	8,483	1	5.1%	0.0132	***
			2	7.8%	0.0179	***
With bootstraps <90% collapsed	RAxML	8,483	1	5.1%	0.0138	***
			2	6.0%	0.0146	***
Full dataset	FastTree	8,483	1	6.7%	0.0161	***
			2	6.7%	0.0172	***
			3	6.6%	0.0141	***
			4	5.7%	0.0164	***
			5	5.2%	0.0161	***
			6	5.7%	0.0144	***
			7	5.6%	0.0150	***
			8	6.5%	0.0138	***
			9	6.0%	0.0158	***
			10	6.9%	0.0168	***
BEAST Sub-Sample	BEAST	652	1	5.1%	0.0308	***
Sequences with only 1 VL removed	RAxML	6,757	1	7.8%	0.0177	***
			2	6.6%	0.0165	***

[1]'***' indicates significance at the Bonferroni-corrected p value. Significance is determined by a Log-Likelihood Ratio Test against the same model with no genetic relationships, as described in Chapter 2.6.1.

produced significant estimates with a mean heritability value of 6.2% (CI 3.0–9.2%; Table 3.5).

3.3.2 Phylogenetic Uncertainty

As is typical for phylogenies based on population samples of HIV *pol* sequences, there is relatively little well-supported internal structure (see Method section 2.5.3). In order to avoid possible bias in the heritability estimates, the analysis was repeated after splits with bootstrap-support values less than 90% were collapsed, which removes 78% of internal nodes. Nevertheless, the heritability estimates remained significant in each case, with estimates of 5.1% (CI 2.4–7.8%) and 6.0% (CI 3.1–8.8%; Table 3.5). However, when the entire tree was collapsed (excepting the split to the outgroup) leaving only

branch-length information, the estimate was not significant, highlighting that detecting the heritability signal relies on at least some tree structure.

One hundred bootstrapped phylogenies were analysed to further examine the effect of uncertainty in the tree. Only four of the resulting heritability estimates failed to reach significance after Bonferroni correction (though their p-values were still <0.002), resulting in a mean heritability estimate of 5.5% (CI 2.6–8.5%; see Appendix C on page 85).

3.3.3 Phylogenetic Effect on Viral Load and Change over Time

In order to investigate how the viral genetic effect on set-point viral load varies across the phylogeny and through time, a time-resolved phylogeny was constructed using BEAST. For reasons of computational tractability (see Methods Section 2.5.2), this phylogeny had to be generated on a 652 sequence sub-sample of the dataset, but still produced a significant heritability estimate of 5.1% (CI -0.9–11.2%; $p<0.005$). Unsurprisingly, the loss in power due to the major reduction in dataset size caused the confidence intervals to expand, with the lower CI crossing zero. BEAST estimated the root of the subtype B sub-sample sequences to be 23.9 years before the most recent sample (95% highest posterior density (HPD): 21.8-26.0 years), placing the root at around 1985 (1983-1987). ASReml was then used to estimate the phylogenetic effect of each node on viral load, and these estimates were mapped onto the time-resolved phylogeny, allowing the distribution of the effects across the tree and over time to be visualised (Figure 3.2 on page 83). This showed some viral lineages to be clearly associated with substantial positive genetic effects on viral load, relative to the mean, and others to be associated with similarly large negative effects.

To investigate more formally the change in set-point viral load over time, I conducted an analysis in the R package MCMCglmm (R Development Core Team, 2011; Hadfield, 2010) in order to estimate the change in viral load due to selection on the virus and environmental effects using information from the temporal variation in sample dates (see Methods Section 2.7). The analysis showed there was a small change in viral load due to between-lineage selection of 0.002 $\log_{10}$ copies/mL/year, though this was not significantly different from what could be explained by drift. The analysis

also revealed that within-host selection on the virus and environmental effects would have contributed a small but significant negative change in viral load of -0.05 $\log_{10}$ copies/mL/year (Figure 3.3 on page 84). When the effects of the two types of selection are combined, the much larger magnitude of the estimated change due to within-host selection gives an overall change in viral load of -0.05 $\log_{10}$ copies/mL/year (Figure 3.4 on page 84).

3.4 Discussion: The Genetic Basis of Set-Point Viral Load in Subtype B

3.4.1 Heritability Estimates and Phylogenetic Uncertainty

The analysis showed that viral genotype has a small but significant effect on set-point viral load in the UK subtype B population, with an estimated mean heritability of 5.7% (CI 2.8–8.6%). When the analysis was repeated after subsampling and using two different phylogenetic methods, the heritability remained significant and did not differ greatly from the original estimate. As the star-like structure of HIV phylogenies can cause poor resolution of the internal nodes, resulting in low split support values, the impact of this effect was tested by collapsing weakly-supported nodes and analysing one hundred bootstrapped phylogenies. This showed that the heritability estimates and their significance were not due to spurious or poorly-supported splits.

Analysing a smaller sampled dataset in BEAST allowed further investigation of the genetic effect on viral load. Plotting the estimated node effect on viral load back onto the phylogeny for the 652 sampled sequences illustrates the association of closely related sequences and similar genetic effects on viral loads in transmission chains that seem to have begun differentiation around the time subtype B arrived in the UK (Hué et al., 2005). Finding viral lineages with both positive and negative genetic effects on viral load indicates that there is viral genetic variation that acts to both increase and decrease viral load relative to the mean.

The time-resolved tree produced by BEAST estimated the root of the UK subtype B epidemic at around 1985 (1983-1987). This date seems slightly late, given that the

first case of AIDS in the UK was diagnosed in 1981 (du Bois et al., 1981). There is also evidence that there were multiple distinct lineages of subtype B already circulating in the UK in the 1980's (Hué et al., 2005), and a previous analysis estimated the root of the subtype B epidemic at 1975 (1968-1980) (Leigh Brown et al., 1997). However, the sub-sample of 652 sequences used in the BEAST analysis was not entirely random, as the sequences were selected from those that retained structure after the splits in the tree with less than 90% bootstrap support had been collapsed, and this could lead to some bias in the dating of the tree.

It is important to note, however, that the estimated date of the subtype B epidemic in the UK differs greatly from the suggested root of the global subtype B epidemic, which has been estimated at somewhere between the mid-1940's (Faria et al., 2014) to the early- to mid-1960's (Gilbert et al., 2007; Abecasis et al., 2009). This is not surprising, as subtype B arrived in the UK from the US (Leigh Brown et al., 1997; Hué et al., 2005), where it had been imported from Haiti (possibly as a single variant) (Gilbert et al., 2007) years after it was originally brought to Haiti by Haitian workers returning from the DRC (Gilbert et al., 2007; Kuyu, 2008; Faria et al., 2014). These subsequent 'bottleneck' events, where very few variants founded each epidemic in Haiti, the US, and the UK, would indeed lead to a more recent common ancestor for the subtype B UK epidemic.

3.4.2 Other Effects on Viral Load

The estimates of the fixed effects influencing set-point viral load reflect previous reports identifying age (O'Brien et al., 1996; Nogueras et al., 2006) and sex (Farzadegan et al., 1998; Sterling et al., 1999; Gandhi et al., 2002) as significant, with older individuals and males having higher set-point viral loads. Ethnicity was also found to have a significant effect on set-point viral load, finding a similar estimate for the effect of Black-African ethnicity to a previous paper looking specifically at this effect (Müller et al., 2009b). Although many previous studies on the influence of ethnicity on set-point viral load suggest there is no difference between ethnic groups, or that non-white minorities have higher viral loads (Brown et al., 1997; Swindells et al., 2002; Boyd et al., 2005), differences in socio-economic status, risk-group, and access to care make the effect of

ethnicity difficult to investigate (Boyd et al., 2005). Non-white individuals are known to access care and be diagnosed at a later disease stage than white individuals (Saul et al., 2000; Burns et al., 2001; Swindells et al., 2002; Boyd et al., 2005), so controlling for disease stage at the time clinical measures are taken is important, as done in Müller et al. (2009b). Different ethnic groups can also be strongly associated with one exposure type. In both Boyd et al. (2005) and Müller et al. (2009b), as well as in my own subtype B data, black African individuals are much more likely to identify their risk group as heterosexual, while white individuals identify as MSM. This confounding of ethnicity and risk group makes it very difficult to disentangle the effect each may have on viral load. Only Müller et al. (2009b) addressed this issue, by analysing only those infected via heterosexual contacts. Given the large sample size and many factors controlled for in my analysis, and the well-controlled analysis performed by Müller et al. (2009b), it seems likely that ethnicity influences viral load, but more well-designed studies will be needed to confirm this finding.

The finding that those with longer time from HIV diagnosis to viral load testing had a slightly lower set-point viral load could reflect that individuals with lower viral load progress more slowly and therefore may be in general slower to access care, and also indicates that I am not classifying late-stage, rising viral loads as 'set-point,' which would result in the opposite effect. Finally, the fact that individuals with a more recent year of diagnosis also have a slightly lower set-point viral load could suggest that the proportion of individuals being diagnosed in late-stage infection is decreasing with time (Health Protection Agency, 2011). Overall, the analysis performed here is supported by the agreement in fixed effects with the findings of previous studies.

3.4.3 Comparison to Previous Analyses

Previous studies investigating the heritability of viral load in HIV have reported the genetic effect at between 1.3% to 60%, though disagreement in how to best measure heritability can make these numbers hard to interpret (see Chapter 7.2.2). If heritability is measured as the regression slope or by phylogenetic signal, estimates range from 10% to 60%, higher than the estimate of 5.7% obtained here. Four of the seven studies were done on cohorts infected with subtypes other than B: two on subtype C (Tang

et al., 2004; Yue et al., 2013), and two on mixed subtype populations (Hollingsworth et al., 2010; Lingappa et al., 2013), making comparisons difficult. Because virulence differs between subtypes (Kanki et al., 1999; Kaleebu et al., 2001; Kiwanuka et al., 2008), heritability estimates could be affected in studies where the cohort is infected with multiple subtypes, even when subtype is included as a variable in the model. Similarly, both the environmental and genetic variance that determines heritability can vary between populations, and may be particularly divergent between studies focusing on different demographic or risk groups. Müller et al. (2011) postulate that increased viral and host genetic variance in Africa versus Europe, as well as differences in the environmental variance found in rural African populations compared to high-income Western countries, could cause very different estimates of heritability. Considering this, some disparity in heritability estimates may not be unexpected.

Six of the previous studies used transmission pairs (n=23 to n=195) to estimate the heritability of viral load, and this could also influence the estimates obtained. As discussed in the introduction (Chapter 1.5 on page 20), the viral load in transmitting couples may not reflect the general epidemic, transmitting couples may share HLA alleles which could influence viral load, and long-term sexual partners often share confounded environmental factors (Lockett et al., 2001; Tang et al., 2004; Dorak et al., 2004).

The only previous study that utilized a phylogeny-based approach also reported a heritability estimate considerably higher than the one obtained here. Alizon et al. (2010) obtained a significant heritability estimate of around 50-60% when they used the most stringent criteria to define which samples would be taken as set-point viral load. Heritability estimates apply only to the population studied, so their estimate may be specific to this small (n=134) population of MSM individuals with exceptionally stable viral load measures. Interestingly, when they relaxed their definition of set-point viral load to include all 661 MSM who had at least three consecutive viral load measures within 1 log, the heritability estimates shrank to around 11%, much closer to my estimate. This suggests that using overly-strict criteria may result in heritability estimates that are not generalizable to the larger population.

The analysis performed here avoided issues associated with using multiple subtypes,

transmission pairs, or restricted samples by including as many cases as possible. The aim was to minimise bias, but this clearly would be expected to introduce a substantial amount of noise and depends on the availability of large datasets. In fact, twenty-fold more individuals were included than the largest previous dataset with a significant heritability estimate (Alizon et al., 2010). Nevertheless, this approach could allow some viral loads to be classified as set-point when they were actually taken during the acute stage, prior to the onset of AIDS, while on ART, or during a transient rise in viral load. The data cleaning methods utilized were able to exclude several cases that may have fallen into these categories (see Chapter 2.3), but this was difficult when there was only one pre-ART measure, as applied to approximately 20% of the dataset (1,726 cases). If many of the viral loads classified as 'set-point' are not actually set-point measurements, this could affect the estimate of heritability obtained. However, when the dataset was re-run in duplicate after removing these 1,726 cases, the heritability estimates remained significant with a mean value of 7.2% (CI 3.9–10.6%; Table 3.5), showing that any errors made in classifying sequences with just one pre-ART viral load do not significantly affect the estimate.

3.4.4 Change in Viral Load Over Time

No evidence was found that subtype B HIV is becoming more virulent in the UK. Indeed, the relatively small heritability of around 6% implies that host, environmental, and demographic effects play a much larger role in determining viral load than the virus genotype in this population, and suggests that any change in viral load due to the viral genotype would be relatively small. As mentioned in the introduction, the implications of a heritable viral load have been extensively explored, especially in the context of HIV adapting towards an 'optimal' viral load for transmission due to selection (Fraser et al., 2007; Müller et al., 2011). My findings, however, imply that selection on the viral genetic component of viral load would have very limited influence on viral evolution.

The MCMCglmm analysis estimated a small but significant decrease over time of -0.05 $\log_{10}$ copies/mL/year in the mean value of the component of viral load determined by viral genotype. At this time the change due to selection on the virus cannot be disentangled from change due to environmental effects that were not controlled for,

such as the background level of ART in the population, so it cannot be assumed that all (or even any) of this change is due to selection on the viral genome. As outlined in the Methods (see section 2.7), simulations investigating the power of the method used to detect changes in viral load due to the viral genome over time suggest that the estimate of change due to between-lineage selection if probably an underestimate. However, even if the true value is double the prediction (0.002 $\log_{10}$ copies/mL/year) its magnitude is very small compared with the change due to within-host selection and environmental effects (-0.05 $\log_{10}$ copies/mL/year) (see Figure 3.3 on page 84). Therefore, the overall change due to selection on the virus will be largely due to this within-host component. It is important to note that the estimate of a decrease in viral load of -0.05 $\log_{10}$ copies/mL/year is change due to selection acting on the virus, and thus is different from the 'year of diagnosis' effect estimated in the main heritability model (see Table 3.4), which estimates change in viral load over time due to all factors, not just the viral genome (see Chapter 2.7).

It should also be noted that though the viral genetic influence on viral load seems to be causing a decrease in viral load, this does not necessarily mean that overall viral load would be expected to decrease. With the small viral genetic contribution to the variance in viral load estimated here, changes in any of the many host and environmental factors influencing viral load could cause viral load to remain constant or even increase.

Previous cohort-based studies of viral load data have indeed estimated an increase in the phenotypic value. In an analysis based on 1,584 individuals with viral load data from the 22 CASCADE cohorts, Dorrucci et al. (2007) estimated an increase in set-point viral load of more than a log over 30 years. Herbeck et al. (2012) performed a meta-analysis based on eight previous studies investigating change in viral load, which generated a more modest estimate of 0.013 $\log_{10}$ copies/mL/year and an overall increase of 0.39 $\log_{10}$ copies/mL in 30 years. More recently, Pantazis et al. (2014) analysed 88,205 viral load measures from the CASCADE cohort and found an increase of 0.5 $\log_{10}$ copies/mL between 1980 and 2002, and a slight decrease from 2002 until 2008. Parsons et al. (2014) used viral load data from the UK Register for HIV Seroconverters, and estimated that set-point viral load had increased roughly 0.2 $\log_{10}$ copies/mL between 1997 and 2012.

These changes have led to suggestions that the virus may have evolved to become more virulent (Dorrucci et al., 2007; Herbeck et al., 2012; Pantazis et al., 2014), but this was not directly analysed and is clearly not the case in my study. However, a much larger fraction of the phenotypic value of viral load in my model is determined by the fixed effects including sex, age, and time from diagnosis to first viral load, which have certainly not remained constant over the course of the epidemic, so the observations by no means necessarily conflict. The time period used to estimate these changes also differs between studies. Though my time-dated phylogeny of subtype B goes back from 2009 to the mid-1980's (Figure 3.2 on the next page), most of the sequences were taken within ten years of the most recent sample. Thus, the power to reliably estimate changes in viral load due to the viral genome rapidly decreases as the phylogeny goes backwards in time, where fewer branches are present in the tree, making it hard to postulate on the deeper history of change in viral load. The estimate obtained is most reliable near the tips, however, which would agree with Pantazis et al. (2014)'s finding of a decrease in viral load between 2002 and 2008.

The studies included by Herbeck et al. (2012) range from -0.013 to 0.056 $\log_{10}$ copies/mL/year in their estimates, with the largest study reporting a significant decline of -0.013 $\log_{10}$ copies/mL/year. This suggests that changes in viral load are difficult to quantify and may be quite population specific, with different environmental effects and selection pressures working in each. Herbeck et al. (2014) investigated this with simulated epidemics and found that the direction of change in viral load over time was highly dependant on the viral load of the founding lineage and the age of the epidemic.

My findings indicate that the genotype of HIV subtype B in the UK has a small but significant effect on viral load, and suggest that the virulence of HIV has not increased. I found the new phylogenetic method to be robust to multiple methods of phylogenetic construction and also to potential errors in the phylogeny, indicating the method could allow investigation of trait heritability in many datasets where sequence data is available.

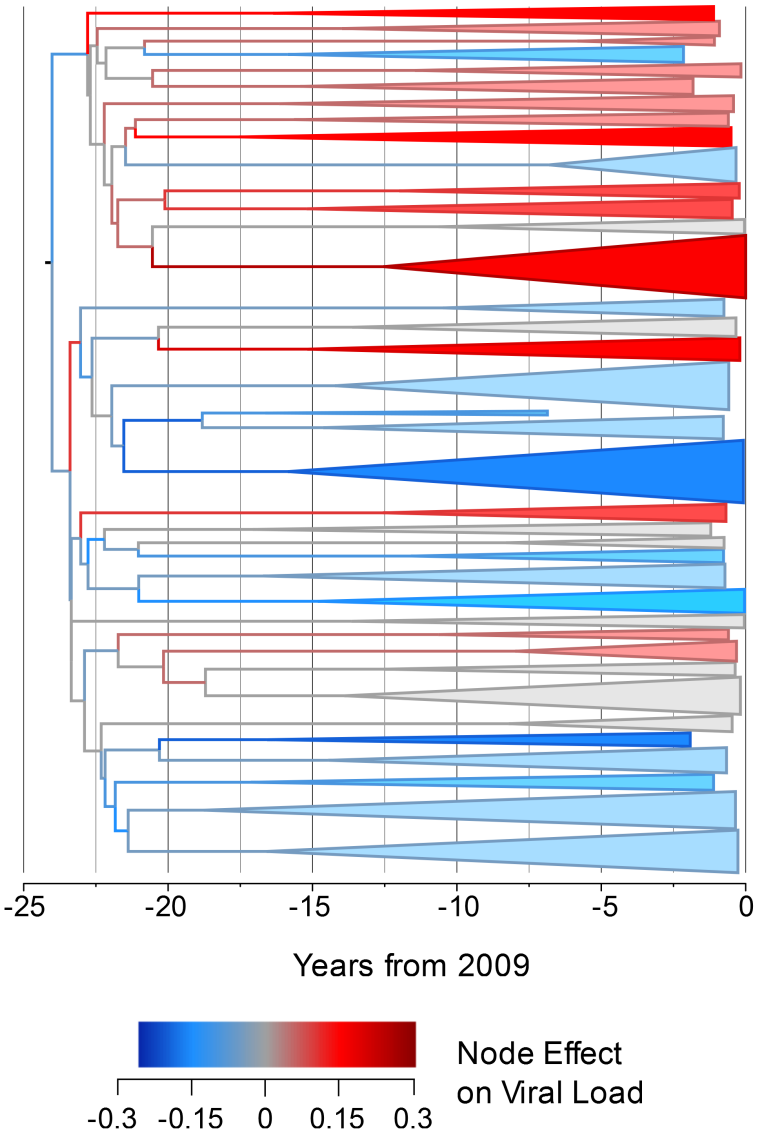

Figure 3.2: **The estimated node effect plotted onto the phylogeny - Subtype B.** The estimated phylogenetic effect of each node on $\log_{10}$ viral load plotted back onto the phylogeny from the 652 sequence sample BEAST analysis. The axis shows the time in years from the most recent sequence, which was taken in 2009. Branches have been coloured by the scale of the effect. Cluster of branches have been collapsed to improve readability, and are coloured by the average tip effect within each cluster. As the number of bifurcations in the tree reduces at around 17.5 years before 2009, this was used as the threshold for collapsing. Nodes that have a similar effect on viral load cluster together, as expected if some of the variation in viral load is heritable

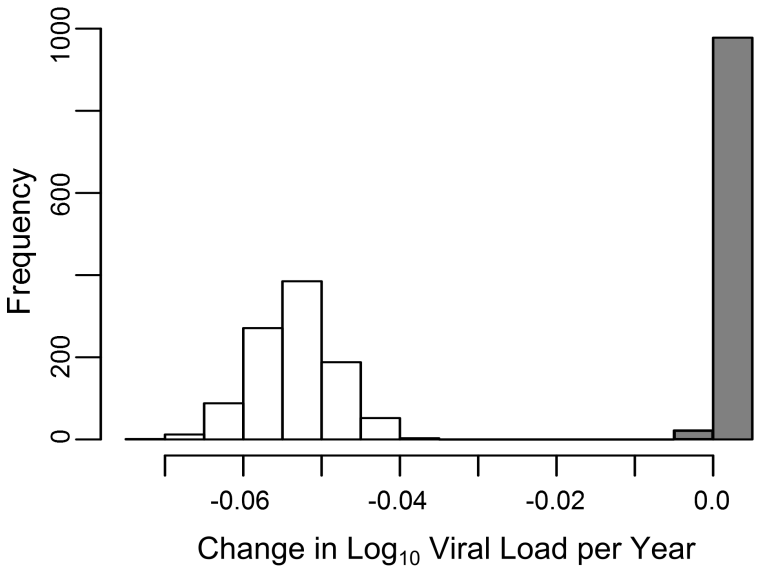

Figure 3.3: The estimated $\log_{10}$ change in viral load over time due to between-lineage selection (grey) and within-host selection and environmental effects (white). The change due to between-lineage selection was not significantly different from what could be explained by drift.

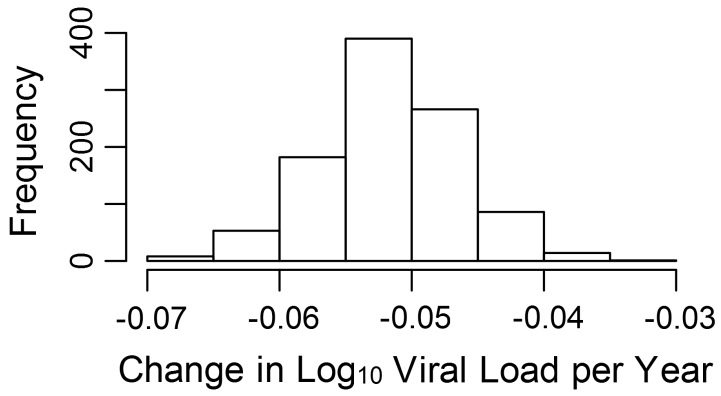

Figure 3.4: The estimated $\log_{10}$ change in viral load over time due to the combined effect of between-lineage selection and within-host selection and environmental effects.

C

Estimates from the 100 Bootstrapped Alignments, Subtype B

Table C.1: Estimates of the viral genetic influence on set-point viral load in the subtype B dataset, using 100 bootstrapped alignments in RAxML

Dataset	**Method**	***N***	**Bootstrap Replicate**	**Viral h^2**	**(Conf. Interval)**	**Sig.[1]**
Full Dataset	RAxML	8,483	1	6.5%	3.4–9.7%	***
			2	6.3%	3.4–9.3%	***
			3	5.5%	2.8–8.2%	***
			4	4.7%	1.9–7.5%	***
			5	4.9%	2.1–7.7%	***
			6	5.4%	2.5–8.4%	***
			7	5.4%	2.5–8.3%	***

Continued on next page...

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$

Table C.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			8	6.9%	3.7–10.0%	***
			9	5.8%	2.7–8.9%	***
			10	5.9%	2.7–9.1%	***
			11	5.2%	1.7–8.6%	***
			12	4.0%	1.5–6.4%	***
			13	5.8%	3.0–8.5%	***
			14	5.4%	2.7–8.1%	***
			15	7.0%	3.6–10.4%	***
			16	5.8%	2.8–8.7%	***
			17	7.1%	3.5–10.7%	***
			18	5.2%	2.4–8.1%	***
			19	5.7%	2.7–8.7%	***
			20	4.5%	1.7–7.2%	***
			21	5.4%	2.3–8.5%	***
			22	6.1%	3.1–9.0%	***
			23	3.2%	0.8–5.6%	***
			24	5.9%	2.9–8.9%	***
			25	6.1%	3.1–9.0%	***
Full Dataset	RAxML	8,483	26	6.2%	3.1–9.2%	***
			27	6.4%	3.4–9.3%	***
			28	4.9%	2.1–7.8%	***
			29	5.2%	2.1–8.3%	*
			30	5.4%	2.4–8.3%	***
			31	6.8%	3.8–9.8%	***
			32	4.7%	2.1–7.2%	***
			33	5.4%	2.3–8.5%	***
			34	6.2%	3.2–9.1%	***
			35	4.6%	1.8–7.5%	***
			36	5.7%	2.4–9.0%	***
			37	5.2%	2.4–8.1%	***
			38	5.2%	2.4–8.0%	***
			39	6.3%	3.3–9.3%	***
			40	4.7%	2.1–7.3%	***

Continued on next page...

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$

Table C.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			41	6.4%	3.0–9.9%	***
			42	6.6%	3.5–9.8%	***
			43	7.2%	3.9–10.6%	***
			44	4.7%	2.2–7.3%	***
			45	5.7%	2.9–8.6%	***
			46	5.7%	2.9–8.4%	***
			47	6.9%	3.5–10.3%	***
			48	5.3%	2.6–8.0%	***
			49	5.9%	2.8–9.0%	***
			50	5.5%	2.7–8.4%	***
			51	3.8%	1.4–6.2%	*
			52	4.1%	1.4–6.7%	*
			53	5.9%	2.9–9.0%	***
			54	5.8%	2.8–8.8%	***
			55	6.4%	3.2–9.6%	***
			56	6.2%	3.0–9.3%	***
			57	4.0%	1.4–6.6%	***
			58	6.6%	3.5–9.7%	***
Full Dataset	RAxML	8,483	59	5.5%	2.7–8.3%	***
			60	5.5%	2.5–8.5%	***
			61	4.5%	1.8–7.1%	***
			62	4.8%	2.2–7.5%	***
			63	6.0%	2.8–9.1%	***
			64	5.6%	2.7–8.4%	***
			65	5.9%	2.9–9.0%	***
			66	5.6%	2.5–8.6%	***
			67	4.5%	2.1–6.9%	***
			68	5.9%	2.8–9.0%	***
			69	6.7%	3.4–9.9%	***
			70	5.8%	2.9–8.6%	***
			71	5.4%	2.6–8.3%	***
			72	4.2%	1.4–6.9%	*
			73	6.0%	3.3–8.7%	***

Continued on next page...

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$

Table C.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			74	3.9%	1.6–6.1%	***
			75	6.2%	2.9–9.4%	***
			76	4.2%	1.5–6.8%	***
			77	5.9%	2.9–9.0%	***
			78	5.2%	2.5–7.8%	***
			79	5.3%	2.6–8.0%	***
			80	6.2%	3.2–9.2%	***
			81	5.7%	3.0–8.4%	***
			82	4.9%	2.2–7.6%	***
			83	4.8%	2.0–7.6%	***
			84	5.2%	2.6–7.8%	***
			85	5.0%	2.3–7.7%	***
			86	5.7%	2.8–8.6%	***
			87	4.8%	2.2–7.5%	***
			88	4.9%	2.2–7.7%	***
Full Dataset	RAxML	8,483	89	8.0%	4.6–11.4%	***
			90	6.9%	3.1–10.8%	***
			91	4.6%	2.1–7.0%	***
			92	4.7%	2.2–7.3%	***
			93	5.8%	2.8–8.7%	***
			94	6.2%	3.2–9.1%	***
			95	5.8%	2.9–8.7%	***
			96	5.2%	2.4–7.9%	***
			97	8.1%	4.8–11.4%	***
			98	4.8%	2.2–7.3%	***
			99	3.8%	1.4–6.2%	***
			100	5.5%	2.8–8.2%	***

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$

"The Universe is under no obligation to make sense to you."

Neil deGrasse Tyson

" 'I do have a dreadful love for understanding,' Alma admitted."

Elizabeth Gilbert - 'The Signature of All Things' (2013)

4

Virulence in Subtype C

4.1 Introduction: Subtype C in the UK

Though HIV-1 subtype B remains the most common subtype in the UK HIV epidemic, the number of non-B subtype infections has increased dramatically over the last ten years. The growth in non-B subtype infections reflects the changing face of the HIV epidemic in the UK. Historically, non-B subtypes were primarily acquired outside of the UK, but in the last few years non-B subtype heterosexual infections acquired within the UK have become more common (Yin et al., 2014) (see Introduction Section 1.1.2).

Because HIV subtypes are often closely associated with particular demographic, ethnic, and risk groups, comparing disease progression between subtypes is often difficult (see Introduction Section 1.2.1). Despite this, multiple studies have identified differences in time to AIDS or time to death between subtypes, particularly in subtypes D and A (Kanki et al., 1999; Kaleebu et al., 2001; Kiwanuka et al., 2008). Given that subtype C is the second most common subtype in the UK, I wished to utilize the new method to investigate the viral genetic influence on viral load in subtype C, and compare these results with what was found in subtype B.

Table 4.1: Number of Subtype C Sequences Discarded During Data Cleaning

Number of Subtype C sequences with at least 1 viral load measure before ART			1,849
1.	(a) Potential acute-stage viral load	-1	
2.	(a) Viral load potentially taken on ART	-7	
	Acute stage or AIDS patient starting unreported ART	-4	
	Missing RT or protease	-14	
	Identical sequences removed	-2	
Final number of records and sequences			1,821

4.2 Methods: Subtype C Data from the HIV DRB

In the UK HIV RDB, 1,849 initial subtype C sequences had at least one viral load measure before starting ART available from UK CHIC. The sequences were aligned and set-point viral load defined as described in the methods, with rules applied to exclude viral loads taken during the acute phase, AIDS, or on unreported ART (see Introduction Section 1.4 on page 18 for issues around choosing 'set-point' viral load values, and Methods Section 2.3 on page 32 for details of how set-point viral load was chosen).

When only one viral load was available and was found to be very high or low, $CD4^+$ count was used to determine if a patient was in acute stage, AIDS, or on unreported ART when viral load was taken, and excluded if so (see Methods Section 2.3.2). If large portion of the sequence were missing, or the sequences were identical to other sequences, samples were also excluded (Methods Section 2.4 on page 39). A total of 28 sequences and data were excluded from the subtype C analysis due to these rules; the number excluded due to each rule can be found in Table 4.1 (the numbered items correspond to the rule list in Methods Section 2.3.2 on page 36).

To make the distribution of set-point viral loads approximately normal, all measures were $\log_{10}$ transformed before analysis. In 90% of the patients the viral load measurement used as set-point was taken within 3 years of HIV diagnosis, suggesting that most viral loads were indeed taken during the chronic phase of the disease. More information about the dates of HIV diagnosis and the dates of the viral load tests used as set-point can be found in Table 4.2, and a histogram showing dates of HIV diagnosis

Table 4.2: Median, quartiles, and range of HIV diagnosis date, set-point viral load test date, and the number of days between HIV diagnosis and set-point viral load testing.

	1st Quartile	Median	3rd Quartile	Range
HIV Diagnosis	30-Jun-2001	22-Jun-2004	1-Jul-2006	1-Jun-1985 to 30-Apr-2009
Date of Set-point Viral Load Test	26-Mar-2002	10-Dec-2004	31-Oct-2006	20-Nov-1995 to 30-Apr-2009
Days between HIV Diagnosis & Set-Point Viral Load Test	1.0	13.0	55.3	-4,246 to 8,509

and set-point viral load testing is shown in Figure 4.1.

The final analysis, after exclusion rules were applied, included 1,821 subtype C sequences and viral loads. As discussed previously, because only the first sequence available for each patient was used, which is taken before ART is started, it is unlikely that many drug-resistance mutations are present. However, sequences that share the same drug-resistant mutations could cluster more closely in a phylogeny due to false but apparent shared ancestry. To prevent this, sequences were stripped of codons in positions associated with drug resistance mutations (Rhee et al., 2003; Liu and Shafer, 2006) before analysis (see Methods Section 2.4). The analysis was also performed on the original, non-stripped, sequences, but no significant difference in the heritability estimates were observed (Table 4.5).

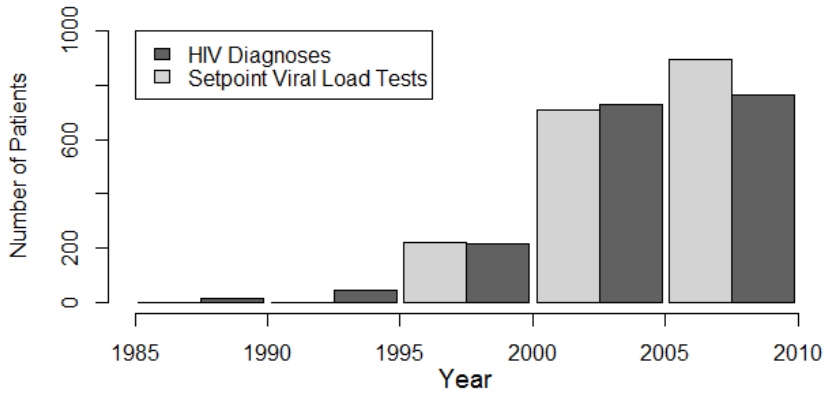

Figure 4.1: Histogram of the dates of HIV diagnosis (dark grey) and set-point viral load tests (light-grey) in the subtype C dataset.

As detailed in the methods, (see Methods Section 2.5.3) 38 subtype reference *pol*

sequences (subtypes A-K) from the Los Alamos HIV Database (www.hiv.lanl.gov) were used as an outgroup, and RAxML (Stamatakis, 2006; Stamatakis et al., 2007) was used to create and bootstrap a maximum-likelihood phylogeny. To make a comparison with a different maximum-likelihood based method, ten phylogenies were also created with the program FastTree (Price et al., 2009, 2010).

TreeCollapseCL 4 was used to root the phylogeny and calculate an average root-to-tip distance to standardize the heritability estimates from ASReml (see Methods Section 2.6). For the main subtype C analysis, the average root-to-tip distance was 0.18 substitutions per site per year, but the root-to-tip distance was calculated independently for each analysis run and used to standardize the estimates from that analysis only.

4.2.1 Choosing Fixed and Random Effects

As with subtype B, the fixed effects to include in the model were identified by preliminary runs in ASReml (see Chapter 2.6.1 on page 51), with age at the set-point viral load sample date, sex, ethnicity, time from HIV diagnosis to set-point viral load date, and year of HIV diagnosis (as a continuous effect) included. Unlike in subtype B, where all these effects were significant, ethnicity was not found to be significant in subtype C (p=0.2; all others p<0.05). However, there is no reason to assume that ethnicity would influence viral load in subtype B but not subtype C, and it is possible that the lack of significance is due to the much smaller sample size of subtype C. Because including ethnicity in the subtype C may help absorb variance that could otherwise be partitioned incorrectly, I decided to keep ethnicity as a fixed effect in the subtype C model.

Country of origin, year of HIV diagnosis (as a categorical effect), and the phylogeny were included as random effects, and year of HIV diagnosis as a continuous fixed effect for the reasons described in Methods Section 2.6.1.

4.2.2 Analysis Strategy

All analyses were run against a 'null' model that excluded the phylogeny in order to test the significance of the model fit, using a log-likelihood ratio test with one degree

of freedom to see if including the phylogeny significantly improved the model.

To investigate how uncertainty in the phylogeny may be influencing the heritability estimates obtained from ASReml, the original analysis was repeated using TreeCollapseCL 4 to collapse splits with bootstrap-support values less than 90% down to polytomies (see Methods Section 2.5.3). As with subtype B, one hundred bootstrapped trees were generated in RAxML using the rapid bootstrap search (Stamatakis et al., 2008) and analysed to further investigate the effects of uncertainty in the tree.

When using RAxML, sequences were run through RAxML and the analysis pipeline twice, so that each analysis was performed in duplicate. When using FastTree, ten independent phylogenies were created and run through the pipeline, and when using BEAST, only one time-scaled phylogeny was created. Bonferroni correction was used to adjust the significance threshold for the number of replicates in each run.

Finally, a time-scaled phylogeny was produced in BEAST (Drummond et al., 2012) to further examine the phylogenetic effects on viral load (see Methods Section 2.5.2), and investigate the change in set-point viral load over time as described in the methods (see Methods Section 2.7). As with subtype B, the complexity of the BEAST analysis means that the number of samples included in the run had to be limited to achieve a reasonable run-time.

Initially, a random subsample of 350 sequences from the subtype C dataset were chosen for the BEAST analysis. Runs in BEAST were performed in duplicate, with a relaxed log-normal clock and the SkyRide population model for 100,000,000 steps, sampling every 10,000 steps. The SRD06 nucleotide substitution model was used, which consists of the HKY substitution model with four categories of site heterogeneity (γ+4) and partitions the first and second codon positions from the third to parametrize separately.

After the initial 100,000,000 steps, which ran in approximately two weeks, convergence had not been reached by either duplicate, and so each run was re-started using the population size, ucld.mean, CP1+2.kappa and CP3.kappa parameters and tree from the last sampled step of the first run. The second runs were also for 100,000,000 steps with sampling every 10,000 steps, and again completed in about two weeks. Neither of the two runs had achieved optimal mixing (exploration of parameter space) but one of

Table 4.3: Demographics of Patients whose Samples were Analysed

		Subtype C (n=1,821)
Age at Set-point (years) (mean, range):		34.9 (17 - 77)
Log_{10} Set-point Viral Load (mean, SD):		4.367 ±0.92
Sex	Female:	1074 *(59.0%)*
	Male:	747 *(41.0%)*
Risk Group	Homo/Bisexual:	133 *(7.3%)*
	Heterosexual:	1605 *(88.1%)*
	IDU:	16 *(0.9%)*
	Other/Unknown:	67 *(3.7%)*
Ethnicity	White:	272 *(14.9%)*
	Black:	1402 *(77.0%)*
	Asian:	41 *(2.3%)*
	Other/Unknown:	106 *(5.8%)*

the duplicates was had converged and was relatively well-mixed, and so was used. 10% burn-in was removed from the resulting files and the two run stages of the first duplicate were combined using an in-house script. A summary tree was generated using the BEAST program TreeAnnotator, and then run through the ASReml pipeline to obtain heritability estimates.

4.3 Results: Initial Findings

4.3.1 Estimates of the Effect on Viral Load

After data cleaning, where incomplete sequences and records with uncertain disease or treatment status were removed, 1,821 subtype C sequences and clinical records remained. 1,797 (98.7%) records had a known ethnicity, 1,772 (97.3%) had a known risk group, and 1,755 (96.4%) had information on both risk group and ethnicity. The demographics of the dataset show that of those with known ethnicity and risk group, 76% (1,336 individuals) are black with a heterosexual risk group, and 49% (858 individuals) are black heterosexual women, in line with the historic association of the C subtype with non-white women and heterosexual spread (Arnold et al., 1995; Tatt et al., 2004) (Table 4.3).

The preliminary ASReml runs on both the B and C subtype determined that sex, ethnicity, country of origin, age when the set-point viral load was taken, year of HIV

diagnosis, and time from HIV diagnosis to the date when set-point viral load was taken were all included in the final model (all effect estimates given in Table 4.4). Set-point viral load was found to decrease with a more recent year of diagnosis and with a longer time period between HIV diagnosis and viral load testing, but increase with age. Women were found to have decreased set-point viral load measures compared to men. For many of the estimates of the effect of ethnicity on viral load, the standard error was found to be larger than or almost as large as the estimate itself, reflecting the fact that ethnicity was not found to be significant in preliminary runs, but was included to help correctly partition variance. The exception to this is having a Black-African ethnicity, which is associated with a higher viral load than having a White ethnicity. However, given the ethnicity imbalance in the dataset (Table 4.3) and the overall non-significance of ethnicity's effect on viral load, it is difficult to interpret this result. The total variance in set-point viral load (V_P) was 0.963 (SE 0.0660). The variance explained by the random effects year of HIV diagnosis and country of origin was estimated at 7.83×10^{-4} and 1.93×10^{-7}, respectively. The variance explained by the viral genome (V_A) was 0.288 (SE 0.0912), and the residual variance was estimated at 0.675 (SE 0.0403).

Using RAxML and ASReml, bootstrapped phylogenetic trees were reconstructed and analysed in duplicate. Each run was done both including and excluding the phylogeny so that the resulting log-likelihood values could be compared to estimate significance. Both replicates produced highly significant ($p<0.01$) heritability estimates of 34.1% (CI 19.5–48.6%) and 25.4% (CI 10.0–40.7%; Table 4.5). Ten phylogenies were also created using FastTree and analysed in the same way. None of the resulting heritability estimates from the FastTree trees were found to be significant at the Bonferroni-corrected p value of 0.005, but four were significant at $p<0.05$ and had a mean heritability of 18.9% (CI 3.6–34.2%). The mean heritability estimates from the six non-significant FastTree phylogenies was 9.8% (CI -3.4–22.9%; Table 4.5).

4.3.2 Phylogenetic Uncertainty

As detailed in the Methods (see Methods Section 2.5.3), phylogenies constructed from population-level samples using the *pol* gene often have poorly-resolved internal nodes

Table 4.4: Mean Fixed Effect Estimates of $\log_{10}$ Set-point Viral Load Influence

Effect	**Units**	**Estimate**	**Standard Error**
Intercept		4.014	0.330
Age at viral load test*	per year	0.012	2.47E-3
Year of HIV Diagnosis*	per year	-0.024	7.10E-3
Time from HIV Diagnosis to VL test*	per day	-1.41E-4	3.79E-5
Sex	Female	0	0
	Male	0.237	0.045
Ethnicity	White	0	0
	Black-African	0.120	0.065
	Black-Caribbean	-0.060	0.135
	Black-other	-0.054	0.147
	Asian/Oriental	-0.352	0.264
	Indian/Pakistani/Bangladeshi	-0.115	0.198
	Other/Mixed	0.144	0.115
	Other	0.144	0.115
	Not known	0.366	0.224

*Values were means-adjusted before analysis

due to the star-like nature of the tree. To investigate the effect poorly-supported splits may have on the heritability estimate, splits with bootstrap-support values below 90% were collapsed, which removed 87% of internal nodes from the phylogeny. Despite this, the heritability estimates remained significant for both duplicates, with estimates of 32.9% (CI 20.7–45.0%) and 29.5% (CI 17.1–41.9%; Table 4.5)).

Given that a significant heritability estimate could be obtained after removing the majority of the structure from the phylogeny by collapsing nodes with support <90%, I attempted to examine the importance of having some structure in the phylogeny in order to estimate heritability by collapsing the entire subtype C tree, leaving only branch length information. Surprisingly, this resulted in a significant heritability estimate with a mean of 95.0% (CI 38.3–151.7%; $p<0.01$), differing greatly from all other estimates obtained. However, the plot of the residual values (the difference between the predicted value and the observed value) and the predicted values shows that the analysis is suspect (Figure 4.2). Residual values should be randomly associated with predicted values as the error associated with predicting the values should be random, and so the

Table 4.5: Estimate of Viral Genetic Influence on Set-Point Viral Load in HIV Subtype C in the UK

Dataset	**Method**	***N***	**Replicate**	**Viral h^2**	**Standard Error**	**Sig.[1]**
Full dataset *(Resistance -codons-stripped)*	RAxML	1,821	1	34.1%	0.0741	***
			2	25.4%	0.0784	***
Full dataset *(Resistance -codons-included)*	RAxML	1,821	1	26.3%	0.0766	***
			2	22.6%	0.0806	***
With bootstraps <90% collapsed	RAxML	1,821	1	32.9%	0.0621	***
			2	29.5%	0.0634	***
Full dataset	FastTree	1,821	1	12.8%	0.0745	*
			2	4.3%	0.0697	
			3	11.2%	0.0779	
			4	17.5%	0.0761	*
			5	15.7%	0.0823	
			6	28.6%	0.0670	*
			7	8.2%	0.0779	
			8	16.7%	0.0730	*
			9	11.0%	0.0619	
			10	8.1%	0.0523	
BEAST Sub-Sample	BEAST	350	1	6.2%	0.0821	
Sequences with only 1 VL removed	RAxML	429	1	24.9%	0.1389	***
			2	38.2%	0.1420	***

[1]'***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$; ' ' indicates non-significance. Significance is determined by a Log-Likelihood Ratio Test against the same model with no genetic relationships, as described in Chapter 2.6.1.

residual vs predicted values plot should be a 'cloud' of values without a pattern (Figure 4.2 (a)). In the residual vs predicted values plot for the completely collapsed tree it is apparent that the error associated with predicting the values is clearly non-random, forming a line (Figure 4.2 (b)).

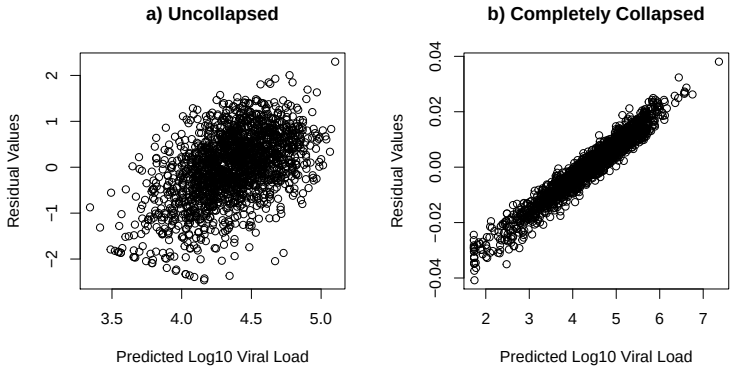

Figure 4.2: Plots showing the association between the residual values (the difference between the predicted value and the observed value) and the predicted values produced by ASReml in an un-collapsed subtype C run (a) and a completely collapsed subtype C run (b). Residual vs predicted value plots should be random, as the error associated with predicting the values should be random. When the plot has a clear pattern, as in (b), this is a sign of an unreliable analysis.

To further examine the effect of uncertainty in the tree, one hundred bootstrapped phylogenies were analysed. Only nine of the estimates were significant according to the Bonferroni-corrected p value of 0.0005, and these gave a mean heritability estimate of 29.7% (CI 15.8–43.5%). Fifty estimates (including the nine just mentioned) were significant at $p<0.05$, with a mean estimate of 25.6% (CI 9.3–37.8%). The remaining fifty estimates were not significant and had a mean heritability of 6.8% (CI -4.0–17.5%). The mean heritability of all one hundred estimates was 15.2% (CI 2.6–27.7%; see Appendix D on page 115).

4.3.3 Phylogenetic Effect on Viral Load and Change over Time

The time-resolved phylogeny generated from the initial 350 sequences randomly sub-sampled from the subtype C dataset generated a non-significant heritability estimate of 6.2% (CI -9.9–22.3%; p=0.55). This non-significance could be due to the very small sample size of the analysis, so a random sub-sample of 650 and 910 sequences were selected to run in BEAST with the same parameters as the initial 350 sample run.

Each sample was started in triplicate to improve mixing (exploration of parameter space) and chances of convergence. After the initial 100,000,000 steps, none of the triplicates of either sample showed signs of convergence. In the 650 sample run, the second and third triplicates were restarted as described for the 350 sample run (see section 4.2.2), and in the 910 sample run, the first and third triplicates were restarted. After a further 100,000,000 steps, neither run of either sample size showed signs of convergence, and so both runs of both samples were again restarted for a second time for a further 100,000,000 steps. After this third run, none of the runs had converged. The two 910 sample runs were restarted a final time, but still showed no signs of convergence. Finally, after finding that the 429 sub-sample of only sequences with multiple viral loads produced a significant heritability estimate when run in RAxML (see section 4.4.5), the same 429 sub-sample was run in BEAST, in triplicate, under the same parameters as used previously. After one restart and a month of run-time, none of the three runs showed signs of convergence.

Unlike the 350 sample run, which took approximately 2 weeks to run 100,000,000 steps, running 100,000,000 steps in the 650 sample run and 910 sample run takes approximately 1 month and 1.5-2 months, respectively, meaning that in total 8 calendar months and two years of computation time was spent attempting to get a time-resolved tree in BEAST. As it seemed increasingly infeasible that a larger sample size was going to converge, the 350 sample time-resolved tree was used calculate the estimate of the phylogenetic effect of each node on viral load, and these values were plotted back onto the tree, despite the lack of a significant heritability estimate from this tree. This allows the phylogenetic effect on viral load across the tree and over time to be visualised, and shows that some viral lineages are associated with both positive and negative effects on viral load, relative to the mean (Figure 4.3 on page 113). BEAST estimated the root of the subtype C 350 sequence sub-sample to be 45.8 years before the most recent sample (95% highest posterior density (HPD): 36.9-57.3 years), placing the root at around 1963 (1952-1972).

Change in Viral Load Over Time

In order to investigate the change in set-point viral load over time as described in the methods (see Methods Section 2.7), a time-resolved tree with a significant heritability estimate is needed. As only the 350-subsample BEAST run had converged, but did not produce a significant heritability estimate, and none of the other sub-sampled BEAST runs showed signs of convergence after extensive runs, a new method for dating phylogenies was attempted. Least-squares-dating (LSD) uses least squares data fitting approaches, which are used to minimise the sum of squares of the errors made when there are more equations than unknowns, and applies this to estimate the substitution rate and dates of ancestral nodes in phylogenies with dated tips (To et al., 2015). By taking advantage of these computationally efficient algorithms, LSD is significantly faster than BEAST, running in linear time when the root is not estimated, and quadratic time when the root is estimated. LSD requires a topology as input, which it does not change (apart from re-rooting, if requested by the user), instead calculating only the branch lengths and the dates of internal nodes.

In an effort to take advantage of this new method, one of the two original 1,821 sequence RAxML trees was run through LSD. First, LSD was not allowed to choose the root, forcing it to use the root designated by RAxML based upon the outgroup sequences, which allows the run to complete in seconds. This run produced a very unrealistic tree, placing the root at approximately 1924, and causing 16% (574 of 3640 branches) of the branch lengths to be negative, with a mean negative branch length of -2.5 years. This high frequency of very negative branch lengths makes the tree difficult to interpret, as the implication of large amounts of 'backwards' evolution (see Figure 4.4 on page 114) undermines the reliability of the tree.

In an effort to obtain a more realistically-dated tree, the LSD run was replicated, allowing LSD to choose the root, which extends the run time to about an hour and a half. The resulting tree placed the root at approximately 1954 and 16% (578 of 3640 branches) of the branch lengths were negative. Despite this more reasonable root date, the negative branch lengths estimated were severe enough that the earliest estimated ancestral node was dated at 1935 – twenty years prior to the root – and 7 other internal

nodes also pre-dated the root. However, given the more reasonable estimate of the overall root, the tree was run through the heritability estimation pipeline. Because negative branch lengths (and zero-length branch lengths) cannot be processed by MCMCglmm, all negative branch lengths were converted to very small positive numbers (0.000001) before the analysis was run. The resulting heritability estimate of 0.9% (CI -2.5–4.3%) was non-significant (p=0.53).

The subtype C sequences included in the analysis date no earlier than 1997, and three of the four subtype C reference sequences date to 1986, 1992, and 1995. As the dearth of earlier dated sequences may cause problems in resolving the dates of deep ancestral nodes, I decided to include publicly available subtype C sequences from earlier time points. Gonzalo Yebra provided 23 subtype C *pol* sequences from Botswana and Zimbabwe dating from 1989 to 1996, curated from the Los Alamos HIV Database (www.hiv.lanl.gov). These were aligned to the resistance-site-stripped alignment of the 1,821 UK HIV DRB subtype C sequences and the 38 subtype reference sequences, and RAxML was used to produce a phylogeny including these sequences. The resulting tree showed the additional 23 sequences clustering tightly into two separate clades amongst the UK HIV DRB sequences.

The LSD analysis was repeated using the new phylogeny, first forcing LSD to maintain the original root based on the outgroup, then allowing LSD to choose the optimal root. When forcing the root to remain the same, LSD placed the date of the root at approximately 1919 and 15% of the branch lengths were negative (569 of 3692 branches). When allowing LSD to choose the optimal root, LSD placed the date of the root at approximately 1946 and 15% of the branch lengths were negative (570 of 3692). As before, though the root date was more reasonable when allowing LSD to select the root, the earliest ancestral node was dated at 1927, and a total of 7 internal nodes pre-dated the root. As before, the resulting tree was run through the heritability estimation pipeline, with negative branch lengths being replaced by very small positive numbers before the analysis was run. The resulting heritability estimate of 0.9% (CI -2.1–3.8%) was non-significant (p=0.43).

As a time-resolved tree with a significant heritability estimate could not be obtained after numerous attempts, I was unfortunately unable to investigate the change in set-

point viral load over time due to selection on the virus.

4.4 Discussion: The Genetic Basis of Set-Point Viral Load in Subtype C

4.4.1 Heritability Estimates and Phylogenetic Uncertainty

The analysis on the subtype C dataset revealed that the viral genotype has a significant effect on set-point viral load in subtype C in the UK, with an estimated mean heritability of 29.7% (CI 14.8–44.7%). When the analysis was repeated using ten trees constructed using FastTree, none of the resulting heritability estimates reached Bonferroni-corrected significance, though four were significant at $p<0.05$, with a mean heritability of 18.9% (CI 3.6–34.2%), not significantly different from the estimates obtained from RAxML.

HIV phylogenies often have a star-like structure that limits the resolution of internal nodes and results in low bootstrap-support values (see Methods Section 2.5.3). To investigate the impact of these weakly-supported splits, poorly-supported nodes were collapsed and one hundred bootstrapped phylogenies were analysed. Collapsing poorly-supported nodes revealed that the heritability estimates and significance were not dependant on spurious splits. Of the one hundred bootstrapped phylogenies produced in RAxML, fifty were significant at $p<0.05$, with only nine of those being significant at the Bonferroni-corrected $p<0.0005$ (see Appendix D on page 115).

The rapid-bootstrap search algorithm used by RAxML (Stamatakis et al., 2008) is not as thorough as the full maximum-likelihood search used by RAxML to find final ML trees. Similarly, FastTree uses fewer optimization steps and considers fewer moves than RAxML, sometimes producing less-accurate trees (Price et al., 2010). Both of these faster methods returned phylogenies where approximately half of the heritability estimates reached significance at $p<0.05$ and the estimates did not differ significantly from the original RAxML estimates, emphasising the importance of being able to accurately reconstruct the HIV phylogeny in order to estimate the heritability.

In subtype B, ninety-six of the one hundred bootstrapped phylogenies (Appendix

C on page 85) and all of the FastTree replicates were significant at the Bonferroni-corrected p value (Table 3.5), with both giving mean heritability estimates nearly identical to the estimates given by the original RAxML runs. The subtype B sample is almost five times larger than the subtype C sample, and the additional information available to reconstruct the subtype B phylogenies may mean that even with less-thorough phylogenetic methods, an accurate tree can be found. However, it's also possible that the very different histories of the two subtypes in the UK create different challenges in reconstructing phylogenies (see further discussion in Section 4.4.6).

To test the importance of having some structure present in the phylogeny, the entire subtype C tree was collapsed, excepting the split to the outgroup sequences, so that only branch length information remained, and then run through the ASReml pipeline. Unlike in the subtype B dataset, where complete collapsing of the tree failed to produce a significant heritability estimate, in the subtype C dataset the resulting estimates were significant, with a mean of 95.0% (CI 38.3–151.7%; $p<0.01$). This heritability estimate is considerably higher than all other significant heritability estimates obtained for the subtype C dataset, and has very large confidence intervals. While this result seems to imply that branch length information alone can perfectly predict viral load after accounting for environmental and demographic effects, the extremely non-randomly scattered residual vs predicted values plot (Figure 4.2) indicates that the analysis may be unreliable.

Under a Brownian motion model, the variance of the trait values increase linearly with time. With no information but the branch lengths (i.e. ignoring tree topology), estimates of the phylogenetic variance exploit any relationship between the variance of trait values and the branch lengths they are associated with. When the intercept of this relationship passes through the origin (i.e. zero variance at branch lengths of zero) the data are consistent with evolution under pure Brownian motion (h^2=1, all the variation in trait value can be explained by the distance to the root). Under this scenario, predictions, conditional on the random effects, would be highly correlated with trait values.

However, these predictions are conditional on the estimated heritability value being the true value. In reality, extrapolating what the variance would be at a branch length

of zero from the observed variation in branch lengths is likely to be subject to large sampling variance. If the heritability is actually lower, then the branch length information would not appear to predict the trait values as well. As shown by the very large confidence intervals of (38.3–151.7%), it would seem that the true heritability could deviate substantially from the estimate of 95.0%, meaning that the predicted trait values would be much less accurate.

4.4.2 Phylogenetic Effect on Viral Load

The genetic effect on viral load was further investigated using the time-resolved phylogeny produced in BEAST with a 350 sequence sub-sample of the subtype C dataset. The estimated node effects on viral load were plotted onto the time-resolved tree and coloured by the size and direction of the effect (Figure 4.3 on page 113), and shows that some viral lineages are associated with positive genetic effects on viral load, relative to the mean, while others are associated with negative genetic effects on viral load, as was seen in subtype B (Figure 3.2 on page 83). The node effects on viral load in the subtype C tree are less pronounced than those observed in the subtype B tree, which is likely due to the heritability estimate from the subtype C time-resolved tree being non-significant.

The root of the subtype C sub-sampled sequences was estimated by BEAST to be around 1963 (1952-1972), much earlier than the subtype B estimate of 1985. Though recent analysis with sequences dating back to 1959 has speculated that subtype C emerged as a distinct subtype before subtype B, perhaps in the early 1940's, (Faria et al., 2014), analyses containing only sequences from the mid-1980's onwards estimate the root of the subtype C tree somewhere between the mid-1950's and late 1960's (Travers et al., 2004; Rousseau et al., 2007; Tee et al., 2008; Abecasis et al., 2009). Though the subtype C 350 sequence sub-sample analysis in BEAST converged, the 'mixing' (number of states explored) was not as good as in subtype B, lending uncertainty to the root date, reflected in the wide 95% HPD interval of 1952-1972. Despite this, the estimated date of 1963 from the BEAST analysis is consistent with previous estimates using sequences from after 1980.

Unlike in subtype B, where the root of the UK epidemic differs greatly from the root

of the global subtype B epidemic (see Section 3.4.1 on page 76), the estimated root of the subtype C epidemic in the UK aligns well with the global subtype C epidemic root. Subtype B underwent repeated bottleneck events when it was transmitted from the DRC to Haiti, the US, and finally, following a limited number of introductions, to the UK, causing the lineages in the UK subtype B epidemic to have a more recent common ancestor. In contrast, subtype C seems mostly to have arrived in the UK directly from Africa as individual direct introductions, as the rise in heterosexual and non-B subtype infections in the late 1990's is tied closely with a sharp rise in the number of infections acquired in South-East Africa (Health Protection Agency et al., 2003; Sinka et al., 2003) and an increase in the number of African-born immigrants to the UK (Owen, 2009). As the subtype C infections that arrived in the UK were effectively sampled from the diverse original epidemic in Africa, it is not surprising that the root of the subtype C epidemic in the UK corresponds to the root of the global subtype C epidemic. The diverse background and history of the subtype C epidemic in the UK may be part of the reason why phylogenetic reconstruction proved much more challenging with subtype C than with subtype B (see further discussion in Section 4.4.6).

4.4.3 Change in Viral Load Over Time

As stated previously, in order to investigate the change in viral load over time, a time-resolved tree that produces a significant heritability estimate is required. Obtaining a time-resolved tree for the subtype C sequences proved to be very challenging indeed. Though the 350 sequence random sub-sample run in BEAST did converge, the heritability estimate of 6.2% (CI -9.9–22.3%; p=0.55) obtained from the resulting tree was not significant, possibly due to the small sample size. Attempt to include larger number of sequences in BEAST runs failed to converge even after numerous restarts and months of run-time.

The recent development of a new method of dating phylogenies, LSD, provided another opportunity to generate a time-resolved tree from the subtype C sequences, but failed to produce reasonable trees. Even when LSD was allowed to choose the root and produced a reasonable root-date of 1946 or 1954, the resulting tree had an abundance of negative branch lengths, conveying the unreliability in the tree's dating, and failed

to produce significant heritability estimates (0.9%, CI -2.5–4.3%). The addition of 23 African subtype C sequences dating from 1989-1996 in hopes that they would provide information to help date ancestral nodes also failed to generate a more reliably-dated tree, and the resulting phylogeny also produced a non-significant heritability estimate (0.9%, CI -2.1–3.8%).

The difficulty in obtaining a time-resolved tree using the subtype C sequences contrasts sharply with the relative ease with which a time-resolved tree was generated for the 652 subtype B sequences after just one duplicate run in BEAST, and may be due to the very different history of subtype C in the UK (see further discussion in Section 4.4.6).

4.4.4 Other Effects on Viral Load

Among the fixed effects influencing viral load, age and sex were significant, with older individuals and males having higher set-point viral loads, confirming the same trend found in subtype B (see Chapter 3.3.1) and in previous studies (O'Brien et al., 1996; Farzadegan et al., 1998; Sterling et al., 1999; Gandhi et al., 2002; Nogueras et al., 2006). Though the size of the effect of sex on viral load was almost identical between the B and C subtypes, the size of the effect of age differed, with a larger effect size estimated for subtype C (Subtype C: Table 4.4; Subtype B: Table 3.4). As noted previously, interpreting the fixed effect estimates for ethnicity is difficult, as it was not found to be significant in the subtype C model, and the estimates for most of the ethnicity classes have very large standard errors. The effect of Black-African ethnicity on viral load was the one estimate with a comparatively reasonable standard error, and was estimated as increasing viral load by 0.120 $\log_{10}$ copies/mL relative to those of White ethnicity. This stands in stark contrast to subtype B, where having a Black-African ethnicity was estimated to reduce viral load by 0.260 (Table 3.4), a finding confirmed in a study on the Swiss HIV Cohort (Müller et al., 2009b).

It is difficult to interpret this apparent difference in the effect of ethnicity on viral load in subtypes B and C, particularly as the demographics of the populations infected by each subtype diverge significantly. Among the individuals included in the analysis of subtype B with a known ethnicity, only 1.7% are Black-African. In subtype C, 76.1%

of individuals with a known ethnicity were listed as Black-African. There is also a difference in the country of origin of those identifying as Black-African between the two subtypes: in subtype B, 25.5% of Black-Africans with a known country of origin originated from the UK, but only 15.0% of Black-Africans infected with subtype C reported the UK as their country of origin. (These numbers rise to 28.7% and 15.2%, respectively, if one considers individuals who originated from any EU country.) Despite this, there are no significant differences in the age at diagnosis or age at viral load test between the two groups (Kolmogorov-Smirnov test, p=0.36 and p=0.48, respectively), implying that originating from a country outside the UK does not seem to significantly delay the age at which individuals are diagnosed as HIV positive or begin receiving treatment. It is possible that Black-Africans originating from countries outside the UK or the EU could be infected with HIV at an earlier age, so that even though diagnosis happens at the same age the disease has progressed further, leading to slightly higher viral loads. As there is no information available about when individuals become infected with HIV in the UK DRB, this theory cannot be investigated with the data available, but multiple other studies have reported that HIV-infected Black-Africans do present with more advanced disease (Saul et al., 2000; Burns et al., 2001; Swindells et al., 2002; Boyd et al., 2005).

Three studies that set out to look at the effect of ethnicity on viral load found that there were no differences between black or African individuals and white individuals (Brown et al., 1997; Swindells et al., 2002; Boyd et al., 2005), though none of these studies included information on HIV subtype. Because of the numerous differences in HIV subtype, risk-group, socio-economic status, and access to care (Boyd et al., 2005), separating out the effect of ethnicity on viral load is fraught with difficulty.

As in subtype B (see Chapter 3.3.1), subtype C individuals with a longer time from diagnosis to viral load testing had a slightly lower set-point viral load, which may indicate that those with a lower viral load progress more slowly and thus initiate treatment later. The magnitude of this effect differed greatly between subtypes B and C, however, with subtype B having a reduction in viral load of 6.18×10^{-5} per day, and subtype C having a reduction of 1.41×10^{-4} per day. There is a significant difference in the length of time between diagnosis and viral load testing between the

subtypes (Kolmogorov-Smirnov test, $p<2.2 \times 10^{-16}$), with subtype B having a mean time of 459.5 days (median: 20, Q1: 3, Q3: 245) and subtype C having a mean time of 220.6 days (median: 13, Q1: 1, Q3: 56). Given that at least 71.6% of the subtype C population is Black-African, this shorter time period between diagnosis and viral load testing may reflect the previous finding that Black-Africans are more likely to present in more advanced stages of HIV (Saul et al., 2000; Burns et al., 2001; Swindells et al., 2002; Boyd et al., 2005). The estimate that subtype C individuals diagnosed more recently in time have a lower set-point viral load supports the fact that the proportion of individuals diagnosed in late-stage infection has declined (Health Protection Agency, 2011), but the much larger magnitude of this effect in subtype C than subtype B (-0.024 and -0.004, respectively) suggests that diagnosing more Black-Africans before late-stage HIV infection has had a much larger effect on set-point viral load in subtype C.

4.4.5 Comparison to Previous Analyses

The estimated mean heritability of 29.7% (CI 14.8–44.7%) compares favourably with most previous estimates of the heritability of viral load in HIV. In the two previous studies where at least 95% of participants were infected with subtype C, heritability measured as the regression slope ($\hat{b}$) was estimated at 36% (Tang et al., 2004) and 27% (Yue et al., 2013), very close to my own estimates. Both of these studies used transmission pairs, which could introduce some bias into the heritability estimates obtained due to confounded environmental factors, shared HLA alleles, and viral load measures that do not represent the general epidemic, as discussed in the introduction (Chapter 1.5 on page 20).

Five other papers had participants that were either mostly subtype B or a mix of multiple subtypes, making their estimates of the heritability of viral load, which ranged from 2-60%, more difficult to compare (Hecht et al., 2010; Hollingsworth et al., 2010; van der Kuyl et al., 2010; Alizon et al., 2010; Lingappa et al., 2013).

As in subtype B (Chapter 3.2), as many samples as possible were included in the analysis to avoid problems associated with restricted and selected sample size. Though this is expected to introduce noise, it allowed inclusion of over four times as many samples as the largest subtype C transmission-pair study (Yue et al., 2013). Stringent

data cleaning methods were put in place to avoid analysing viral loads as set-point when they may actually have been taken on ART, during the acute stage, or after the onset of AIDS (Chapter 2.3). However, as noted in the methods (Chapter 2.3.2), determining whether a viral load should be included becomes very difficult when only one pre-ART viral load is available, and it is possible that some of these viral-load measures were not actually set-point measurements. Unlike in the subtype B dataset, where only 20% of the dataset had just one pre-ART viral load available, in the subtype C dataset, 1,392 individuals (76.4%) only had one pre-ART viral load. After removing these individuals, the dataset was re-run in duplicate with the remaining 429 samples. Both runs provided significant estimates of heritability, with a mean heritability of 31.6% (CI 4.0–59.1%), very similar to the original heritability estimate obtained. As might be expected from the extreme reduction in sample size and thus power, the confidence intervals on the estimates were much wider than estimates using the full dataset, and one overlapped zero (Table 4.5). Despite this, the ability to obtain a significant and similar heritability after removing the cases where only one viral load sample was available gives confidence that any mis-classification errors made when choosing 'set-point' viral load are not significantly biasing the result.

4.4.6 Phylogenetic Reconstruction and Dating in Subtype C

The analysis of the subtype C dataset, as presented here, was intended to be a natural extension of the analysis performed on the subtype B dataset (Chapter 3), allowing a comparison of the heritability estimates, phylogenetic effects, epidemic history, and change over time between the two largest HIV subtypes circulating in the UK. However, were the subtype B dataset analysis was fairly straightforward, the subtype C dataset presented unexpected challenges.

As well as the analyses performed on phylogenies produced using the full ML search in RAxML, two less thorough ML-based methods (FastTree and rapid-bootstrapping in RAxML) were also used to reconstruct phylogenies and repeat the heritability estimation. Unlike in subtype B, where all of the estimates from FastTree and ninety-six of the one hundred estimates from the rapid-bootstrapping in RAxML were significant, only half of the subtype C heritability estimates resulting from these methods

were significant at $p<0.05$. None of the estimates from FastTree phylogenies reached Bonferroni-corrected significance of $p<0.0005$, and only nine of the one hundred rapid-bootstrapped phylogenies produced by RAxML reached Bonferroni-corrected significance.

Totally collapsing the phylogeny down to a polytomy also led to unpredicted results in subtype C. In subtype B, this collapse, causing the loss of all structure in the tree, meant that a heritability signal could no longer be detected. In subtype C, this unexpectedly resulted in a mean heritability estimate of 95.0%. The residuals and the confidence interval on the estimate indicate the run is not reliable, and is instead ASReml imprecisely exploiting any detectable relationship between the variance of trait values and the branch lengths they are associated with. However, the fact there is apparently some relationship between branch length and the variance of the trait values to detect in subtype C, when there was none in subtype B, is interesting.

Finally, the difficulty in obtaining a time-resolved tree with the subtype C data, when one was obtained with relative ease using the subtype B data, is another sign of the apparent divergence between the two datasets. A total of 8 calendar months and two years of computation time was spent running and restarting different subsamples of the subtype C dataset in BEAST with the hope of producing a time-resolved tree with a significant heritability estimate. When this failed, the new LSD method provided another chance to obtain a dated phylogeny, but also proved unsuccessful after numerous runs.

The 350 sample BEAST run did produce a converged time-resolved tree with a realistic root date. Though it did not yield a significant heritability estimate, it did highlight the fact that the subtype C sequences circulating in the UK seem to have a similar root date as the subtype C sequences gathered globally, a very different scenario from the UK subtype B, with its more recent root date estimate relative to subtype B globally (Leigh Brown et al., 1997; Hué et al., 2005; Gilbert et al., 2007) (see Chapter 3.4.1). As described earlier, the rise in non-B subtype infections in the UK is closely tied to the rise in the number of infections acquired in south-east Africa (Health Protection Agency et al., 2003; Sinka et al., 2003) and the increase in African-born immigrants in the late 1990s (Owen, 2009), suggesting that the subtype C infections in

the UK are effectively 'sampled' from the diverse original epidemic in Africa. Despite the much smaller sample size (n=1,821) than subtype B (n=8,483), it seems likely that the subtype C datasets is much more diverse, and that the ancestors of the sequences gathered in the UK are far deeper and more distant than the ancestors of the sequences in subtype B.

When subtype C is considered as sub-sample of the African epidemic, with roots going back roughly 60 years, it might seem unsurprising that phylogenetic methods have struggled to accurately reconstruct the history of subtype C from the relatively small sample. As the relationships between many samples is probably fairly distant, lineages will converge deeper in the tree. Most of the information in a phylogeny lies in its tips, with events further from the tips being less certain due to the loss of information that occurs as one moves back through time. In the subtype C dataset, it seems reasonable that many lineages will converge deeper in the tree, where there is less information to correctly predict the phylogeny. This might well explain why less thorough methods of phylogenetic reconstruction (FastTree and the rapid-bootstrapping in RAxML) had difficulty obtaining trees where significant heritability signal could be detected.

This may also explain why BEAST struggled to converge on a time-resolved phylogeny, as all dating information is also only at the tips of the tree. With such an old root, so many deep internal nodes, and no older dated sequences to help calibrate the model, it is perhaps not unexpected that BEAST would struggle to produce a reliable result. The new LSD method likely failed for similar reasons.

Though the deep and more complex history of subtype C has caused problems with some parts of the attempted analyses here, it does provide an interesting dataset for further work. Including publicly available subtype C sequences sampled from around the world, particularly Africa, and creating a new phylogeny would allow investigation of the diversity and possible origins of the subtype C sequences now found in the UK. These sequences may also 'fill in the gaps' between some of the more divergent UK sequences, providing more information to reconstruct a more reliable tree. Global sequences with sample dates could also potentially be located and help improve the chances of obtaining a time-resolved tree. The subtype C epidemic in the UK has never had its phylogenetic history extensively catalogued, and work to better characterise the

history and diversity of the subtype C sequences would be valuable.

However, the UK subtype C epidemic may not be ideal dataset from which to obtain a heritability estimate, given its diverse background and 'sampling' from Africa. Through links with the Africa Centre, it may be possible in the future to obtain viral load and subtype C sequences directly from a cohort in South Africa. These sequences are likely to provide a better dataset from which the heritability of subtype C could be more reliably estimated.

My findings through this analysis of the subtype C epidemic in the UK suggest that the genotype of HIV subtype C has a significant impact on viral load, which is much larger than the genetic effect UK HIV subtype B has on viral load. As with subtype B, I found the new phylogenetic method to be robust to collapsing poorly-supported nodes and significant down-sampling to use only sequences with multiple viral loads available. Unlike in subtype B, significant heritability estimates were less reliably obtained when using less-stringent ML-based reconstruction methods (FastTree and RAxML rapid-bootstrapping algorithm), and obtaining a time-resolved phylogeny that produced a significant heritability estimate proved infeasible, which highlights the importance of being able to reliably reconstruct the phylogeny when attempting to estimate heritability using this method. The difficulties in obtaining phylogenetic trees for the subtype C dataset may in part be due to the much smaller sample size of the data, but are almost certainly also due to the very different and complex history of the subtype C epidemic in the UK.

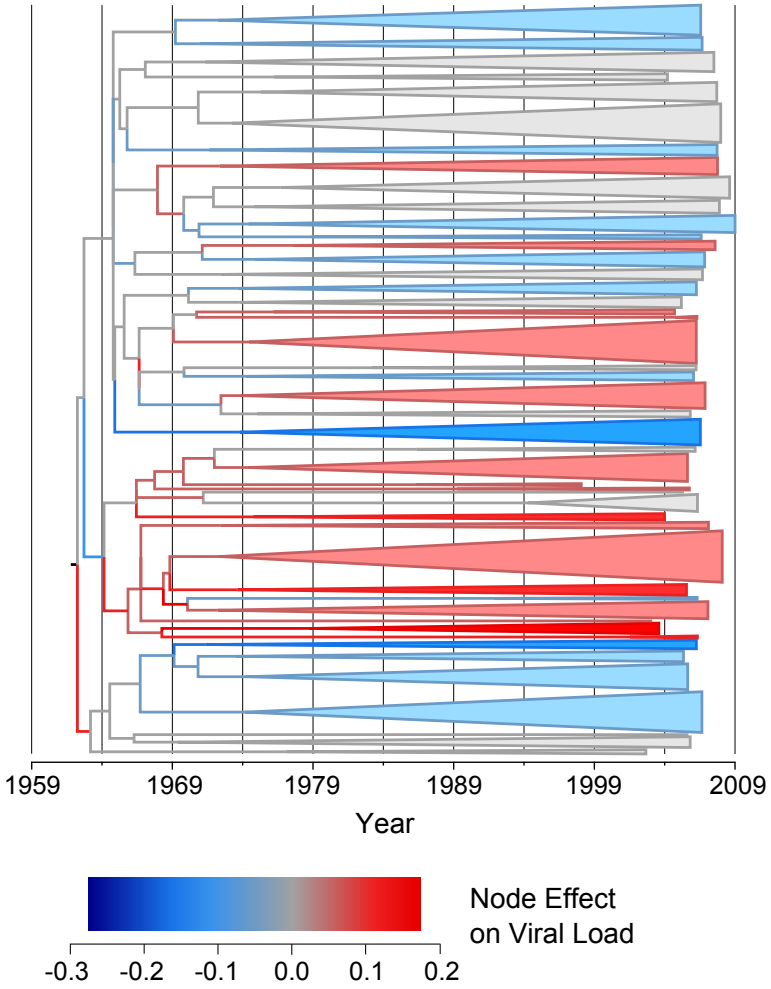

Figure 4.3: **The estimated node effect plotted onto the phylogeny - Subtype C.** The estimated phylogenetic effect of each node on $\log_{10}$ viral load plotted back onto the phylogeny from the 350 sample BEAST analysis. The axis shows the time in years from the most recent sequence, which was taken in 2009. Branches have been coloured by the scale of the effect. Clusters of branches have been collapsed to improve readability, and are coloured by the average tip effect within each cluster. As the number of bifurcations in the tree reduces at around 1974, this was used as the threshold for collapsing. Nodes that have a similar effect on viral load cluster together, as expected if some of the variation in viral load is heritable. The node effect on viral load in this tree are less pronounced than in the subtype B tree (Figure 3.2), which is likely due to the heritability estimate from the subtype C time-resolved tree being non-significant.

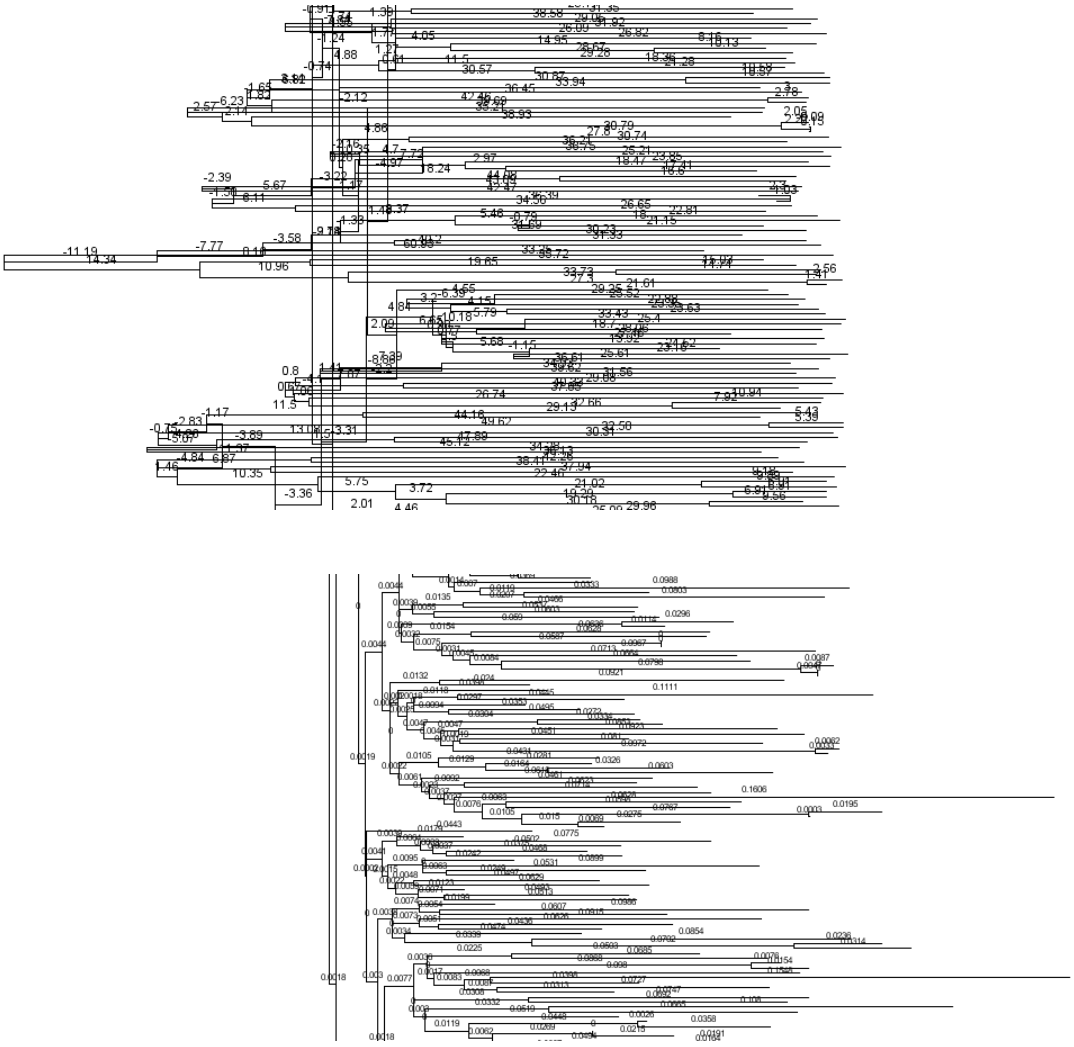

Figure 4.4: The effect of negative branch lengths on a phylogeny generated by least-squares-dating (LSD) (top) compared with the same section of the original RAxML-generated phylogeny (bottom), which only has positive branch lengths. Branch lengths are printed above the branch they refer to. Rather than a conventional phylogeny where each internal node is dated later in time than its ancestral 'parent' (as seen in the bottom phylogeny), negative branch lengths imply situations where the ancestral 'parent' node is younger than its child (as seen in the top phylogeny). Though impossible, it makes the resulting phylogeny hard to interpret, and undermines the reliability of the tree.

D

Estimates from the 100 Bootstrapped Alignments, Subtype C

Table D.1: Estimates of the viral genetic influence on set-point viral load in the subtype C dataset, using 100 bootstrapped alignments in RAxML

Dataset	**Method**	***N***	**Bootstrap Replicate**	**Viral h^2**	**(Conf. Interval)**	**Sig.[1]**
Full Dataset	RAxML	1,821	1	1.2%	-3.9–6.4%	
			2	3.2%	-5.3–11.6%	
			3	23.5%	9.6–37.3%	*
			4	5.6%	-4.0–15.2%	
			5	21.6%	7.6–35.7%	*
			6	23.1%	8.8–37.4%	***
			7	1.4%	-5.4–8.2%	

Continued on next page...

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$; ' ' indicates non-significance

Table D.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			8	17.4%	3.7–31.1%	*
			9	21.9%	8.0–35.9%	*
			10	16.5%	3.1–29.9%	
			11	7.0%	-4.2–18.2%	
			12	12.1%	0.0–24.3%	*
			13	15.3%	1.9–28.6%	
			14	9.0%	-4.9–22.8%	
			15	26.1%	9.6–42.6%	*
			16	0.0%	0.0–0.0%	
			17	27.8%	12.3–43.2%	*
			18	10.3%	-3.9–24.4%	
			19	0.0%	0.0–0.0%	
			20	5.2%	-4.3–14.6%	
			21	26.2%	12.4–40.0%	***
			22	2.5%	-4.1–9.0%	
			23	10.2%	-2.4–23.0%	
			24	13.3%	-5.2–31.7%	
			25	1.7%	-5.4–8.8%	
Full Dataset	RAxML	1,821	26	28.7%	13.2–44.3%	*
			27	28.2%	12.6–43.8%	*
			28	13.0%	0.6–25.3%	*
			29	19.1%	5.1–33.0%	*
			30	11.8%	0.0–23.6%	
			31	19.4%	6.3–32.5%	*
			32	0.0%	0.0–0.0%	
			33	3.5%	-6.3–13.3%	
			34	18.6%	6.0–31.3%	***
			35	7.3%	-4.3–18.9%	
			36	18.3%	3.4–33.2%	*
			37	11.0%	-0.9–22.8%	*
			38	10.0%	-2.1–22.1%	
			39	3.4%	-5.1–12.0%	
			40	24.7%	10.9–38.5%	*

Continued on next page...

[1] '***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$; ' ' indicates non-significance

Table D.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			41	16.5%	3.4–29.6%	*
			42	26.8%	11.9–41.8%	*
			43	26.2%	11.5–40.9%	*
			44	11.3%	-2.2–24.9%	
			45	10.3%	-1.8–22.5%	
			46	4.7%	-4.4–13.9%	
			47	7.0%	-2.9–16.8%	
			48	17.0%	4.1–29.9%	*
			49	44.7%	30.3–59.2%	***
			50	15.4%	2.5–28.2%	*
			51	36.3%	21.6–51.0%	*
			52	36.7%	22.3–51.0%	***
			53	1.7%	-5.7–9.1%	
			54	16.7%	1.6–31.8%	
			55	14.4%	1.3–27.4%	*
			56	12.5%	-1.4–26.5%	
			57	4.4%	-5.7–14.5%	
			58	13.5%	-1.2–28.2%	
Full Dataset	RAxML	1,821	59	21.2%	6.4–35.9%	*
			60	23.4%	8.2–38.7%	*
			61	23.4%	9.6–37.3%	*
			62	8.5%	-2.8–19.8%	
			63	11.4%	-1.7–24.4%	
			64	27.5%	12.5–42.5%	*
			65	20.4%	5.9–34.9%	*
			66	3.8%	-6.5–14.0%	
			67	5.3%	-4.5–15.0%	
			68	9.1%	-3.6–21.9%	
			69	0.5%	-5.3–6.3%	
			70	28.8%	15.2–42.5%	***
			71	30.6%	16.4–44.8%	***
			72	36.1%	20.3–51.8%	*
			73	3.7%	-5.0–12.4%	

Continued on next page...

[1]'***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$; ' ' indicates non-significance

Table D.1 – Continued from previous page

Dataset	Method	*N*	Bootstrap Replicate	Viral h^2	(Conf. Interval)	Sig.[1]
			74	17.1%	3.3–30.8%	*
			75	2.3%	-4.6–9.1%	
			76	5.2%	-4.7–15.1%	
			77	10.7%	-1.7–23.0%	
			78	4.2%	-4.7–13.1%	
			79	8.3%	-3.7–20.2%	
			80	34.0%	19.4–48.6%	*
			81	28.0%	12.9–42.9%	*
			82	17.9%	4.0–31.7%	*
			83	23.0%	8.3–37.6%	*
			84	17.2%	4.7–29.8%	*
			85	5.0%	-4.8–14.8%	
			86	27.2%	12.7–41.6%	*
			87	29.1%	13.8–44.2%	*
			88	21.1%	6.5–35.8%	*
Full Dataset	RAxML	1,821	89	8.1%	-2.9–19.2%	
			90	12.1%	-1.2–25.5%	
			91	3.4%	-5.6–12.4%	
			92	3.8%	-4.3–11.9%	
			93	20.0%	5.9–34.2%	*
			94	4.1%	-3.7–11.9%	
			95	19.1%	4.3–33.8%	*
			96	33.7%	19.5–47.8%	***
			97	19.5%	3.8–35.3%	*
			98	7.8%	-4.3–19.8%	
			99	24.5%	11.5–37.6%	***
			100	20.7%	5.7–35.6%	*

[1]'***' indicates significance at the Bonferroni-corrected p value; '*' indicates significance at $p<0.05$; ' ' indicates non-significance

"Furthermore, he quit drinking coffee, and naturally, his brain stopped working."

Orhan Pamuk - 'My Name Is Red' (2001)

"My work, Sonmi, can be taxing and hazardous, but dull? Never."

David Mitchell - 'Cloud Atlas' (2004)

5

Modelling HIV Epidemics: The Discrete Spatial Phylo Simulator

As the 'true' heritability of a trait in a population cannot be directly measured, there are few opportunities to evaluate a method for estimating heritability. My goal Chapter 5 was to develop a way of simulating a realistic HIV epidemic which could be used to assess the performance of phylogenetic methods in estimating disease parameters, and importantly, to test my own new method for estimating the heritability of viral load. Thus, I developed the Discrete Spatial Phylo Simulator (DSPS), which I describe here. First, I outline a brief history of infectious disease modelling, and introduce the basic DSPS as a way to simulate epidemics. Next, I give an overview of modelling sexually-transmitted diseases, focussing on HIV. The majority of the chapter then describes in detail the modifications made to the DSPS to turn it into an HIV-specific model, and how the output from the simulations generates realistic viral phylogenies and sequences, which can be used to assess phylogenetic methods such as the new heritability estimation pipeline.

5.1 Introduction to Modelling of Infectious Diseases

Mathematical modelling has long been used as a way to better understand and predict the dynamics of disease. For the purposes of this thesis, I will focus only on methods relevant to the modelling of infectious diseases, such as HIV. Mathematical

modelling provides a way to evaluate estimates of disease parameters. An appropriate well-parametrized model should be able to reproduce the infection dynamics observed, and may provide a way to estimate further unknown parameters. Modelling can also be predictive, allowing exploration of possible outcomes given different scenarios.

5.1.1 Compartmental Models

One of the earliest types of infectious disease models divided hosts into 'compartments' depending on their infection status (Kermack and McKendrick, 1927), and is thus known as the 'compartmental model.' Arguably the most famous compartmental model is the 'S-I-R' model, which divides individuals into **S**usceptible, **I**nfected, and **R**ecovered (or **R**emoved) compartments. Here, susceptible individuals are able to become infected, infected individuals are actively infected and can transmit the disease, and recovered or removed individuals are no longer actively infected but also cannot be re-infected (they have life-long immunity). Most compartmental models are 'deterministic,' meaning they use ordinary differential equations (ODEs) to describe the rate at which individuals can move from one compartment to another. These equations are usually parametrized to reflect what is known about rate of infection and recovery for the pathogen being modelled, and are sometimes dependent on the number of individuals in another compartment. For example, the rate at which individuals move from 'susceptible' to 'infected' may be dependent upon the number of individuals in the infected compartment already, to reflect how contact rate influences transmission. On the other hand, the rate at which individuals move from 'infected' to 'recovered' is often not dependent on the number of individuals in another compartment, and is thus parametrized as a constant rate. Because deterministic compartmental models do not track individuals, but rather the overall population, they are computationally efficient, and allow complex epidemiological models to be run on limited computational resources.

Different parameters and ODEs can be used to recreate a variety of assumptions about disease progression, such as whether transmission rate is linked to population size (frequency-dependent, also called mass action) or not (density-dependent, also called pseudo mass action). Compartmental models can also be extended to model

diseases with different phases: S-I, for diseases with terminal infection; S-I-S, when the host does not develop life-long immunity; S-E-I-R for diseases where there is a non-infectious 'exposed' period before the individual becomes infectious; and models where immunity wanes ('recovered's go back to being 'susceptible') and where individuals can become 'carriers' after infection rather than recovering. Compartmental models can be extended further still by dividing individuals in the model into different groups. In a simple example there might be two 'S' and two 'I' compartments, one 'S-I' modelling 'high risk' individuals and the other 'S-I' modelling 'low risk' individuals. Infected individuals of both high and low risk could influence the infection rate of both high and low risk susceptible individuals. In a more complex models, individuals could move between being high and low risk. Multiple levels of 'S-I' models may be used to model further divisions of the population, such as by age. Alternatively, multiple 'E' or 'I' categories may be added to model hosts progressing through different disease stages, perhaps with different levels of infectiousness. There is no limit on the number of compartments that can be included in a model, but each one must be parametrized correctly. Every time another compartment is added, error is introduced around the parametrization of that compartment, and modellers must avoid a situation where the model ends up containing so much error it may give inaccurate predictions. There are methods to evaluate error in deterministic models, such as Latin Hypercube Sampling, described by Blower and Dowlatabadi (1994).

Though deterministic compartmental models are very flexible and commonly used, being extremely fast to compute, they do have limitations. While compartmental models can be adjusted to create different scenarios, the dynamics and result of a deterministic model will always be the same for the set of parameters and structure used, which makes them very useful for describing how disease dynamics reach equilibria or move between equilibria. Compartmental models also assume a very large population size and homogeneous mixing between connected compartments (though more compartments could be added to model some level of heterogeneous mixing). This means that contact and transmission rate is not based on actual 'contact' events between individuals, but that the two are parametrized together. Finally, compartmental models do not model the actions of individuals, just of the population as a whole. Because of

this, the effects of individual-level variation cannot be modelled, and the behaviour of a particular individual or group of individuals within the population cannot be traced or further inspected.

5.1.2 Agent-Based Models

Early ideas about agent-based models (sometimes also called individual-based models) can be traced back to the middle of the 20th century (von Neumann, 1963; Gardner, 1970), when systems of 'cellular automatons' were proposed as ways of studying complex interactions based on simple rules governing individuals. In the field of population genetics, Robertson (1978) implemented a very early and basic agent-based model in 1978 to investigate the effect of selection on allele frequency. However, it wasn't until the 1990s, when access to increased computational resources became available, that agent-based models grew in popularity and began being regularly used to model epidemiological scenarios, initially focussing on the spread of livestock diseases (Jalvingh et al., 1999; Mangen et al., 2002; Bates et al., 2003).

Agent-based models differ from compartmental models in that each individual in the population is simulated as a discrete 'object' according to parameters that are unique to that individual. They are stochastic rather than deterministic, meaning that they have inherent randomness due to the variability of parameters associated with the individuals in the model. There may still be an underlying 'S-I-R' model, as there must still be rules about how individuals progress through disease stages. However, these rules are now influenced by the properties of the individual. For example, an individual's susceptibility to infection may be related to their age or previous events in their 'life.' Similarly, an individual's ability to infect others may be dictated by a combination of factors like age and amount of time they've been infected. In agent-based models, transmission rate and contact rate are now no longer coupled together. Contact rate is modelled independently, and may be dictated by a network structure that provides rules describing which individuals contact which other individuals, and at what frequency. The stochasticity of agent-based models means that models starting with the same parameters generate different dynamics and outcomes. (It is important to note that there are ways of introducing an element of stochasticity into ODE compartmental

models, commonly through event-driven approaches (selecting the next event based on the probability of all possible events) implemented with Gillespie's Direct Method (Gillespie, 1977). This is not covered here, but an excellent overview can be found in Chapter 6 of Keeling and Rohani (2008).)

Like compartmental models, agent-based models are open to giving unreliable predictions if there is too much error associated with the parameters used in the model to describe the interaction between the individual traits and the underlying disease model. Unlike compartmental models, they do not require large population sizes or the assumption of homogeneous mixing. Because each individual is tracked through the model, the behaviour and history of one individual, or group of individuals, can be closely inspected, and the the model can produce 'line-lists' or 'transmission trees' that record disease spread through the population from one individual to the next.

Bonabeau (2002) highlighted two main benefits of agent-based models: their ability to provide a natural description of a system and their flexibility. Because agent-based models produce results based on the interactions of individuals within the model, they can produce and describe counter-intuitive and unpredictable scenarios that are hard to reduce down to a general description of a system. Agent-based models can also provide more natural descriptions of a system. Bonabeau (2002) gives the example of modelling movement in a supermarket – it might be much easier to come up with rules that dictate the behaviour of individual shoppers than to come up with equations that describe the density of shoppers in the supermarket. Finally, the inherent properties of agent-based models make them very flexible. Adding more individuals to the population or slightly tweaking the contact network or parameters that influence infectiousness is easily done.

One of the biggest drawbacks to using agent-based models is the computational resources required due to the complex nature of the model. The traits of every individual must be stored and accessed while also tracking the overall state of the population. Each event must be generated and evaluated independently, as no one rate can be used to describe what is happening. With the computational systems currently available, agent-based models are becoming easier to run, even for fairly complex implementations. However, as the size of the population and the complexity of the model increase, so do the computational requirements and processing time needed to run the model,

creating practical limits on what can be modelled.

5.1.3 Modelling in Time

Both compartmental and agent-based models can be modelled in either discrete or continuous time. Compartmental models involving ODEs are always in continuous time, as the model calculates the total number of individuals in each compartment for all values of the time over which the model is run. Compartmental models can also be designed that run in discrete time, where the number of individuals in each model is only calculated at the specified time intervals. At any time point in between these discrete time points, the numbers will be the same as at the previous time point. For computational reasons, some programs convert continuous-time compartmental models to discrete-time models with very small time intervals to execute the model more efficiently.

Agent-based models can also run in both discrete and continuous time. In discrete-time agent-based models, there may be a set number of contacts per time 'step,' for example. For each time interval in the model, individuals are chosen to have contacts or events, and the outcomes of these events are considered to happen simultaneously during the unit of time. Agent-based models can also run in continuous time by generating discrete events that happen at non-standard time intervals governed by an event rate. A common way of determining the time when events should take place in a continuous-time agent-based model is by using the Gillespie algorithm (Gillespie, 1976, 1977) (described in Section 5.2.3).

5.1.4 The Discrete Spatial Phylo Simulator

The Discrete Spatial Phylo Simulator (DSPS) (Lycett et al., 2015) is an individual-based stochastic model that runs in continuous time. The base code for the DSPS as a more general infectious disease simulator was written by Samantha Lycett, but I have developed the DSPS extensively to enable simulation of realistic HIV transmission scenarios.

The purpose of the DSPS is to simulate complex, realistic epidemics along highly customizable contact networks. Input is in the form of an XML file, which outlines the

parameters of the model, and output is in the form of a full event log and a phylogenetic tree of any samples taken. I will first describe how the basic DSPS model works, as designed by SL, and then explain the modifications I made to convert the DSPS to a realistic and flexible HIV epidemic simulator (section 5.3).

5.2 Basic DSPS

Within the DSPS, 'host' objects (agents) are organized into 'demes,' which can be connected together to form contact networks. All hosts within a deme have the same parameters (for example, recovery rate), but each deme can have its own discrete parameters that describe the hosts within. The hosts and demes can be configured for a variety of settings: there can be just one deme with multiple hosts, to model a fully-connected free network; multiple demes can represent geographic locations or species, with hosts transmitting or moving between demes; or multiple demes could be used to represent organisms, with each organism having their own population of 'hosts,' representing a viral, bacterial, or parasitic population within the host.

5.2.1 Input

The input for the DSPS is provided in an XML file with four sections. The 'General' section gives the random number seed to be used in the run, the path and root file name for the output files, and the number of replicates. The 'Sampler' section allows the user to specify the type of sampling to use, whether to sample just before recovery (how much time before recovery can be specified), or to take a random sample at each time point. The 'Deme' section (or sections) is where the number of demes can be specified, as well as what parameters each deme should have, including number of hosts, type of infection (migration of the infected individuals, or infection only over the contact network) and infection, exposure, recovery, and migration parameters (depending on the model being implemented). It should be noted that if migration of infected individuals is allowed, the 'migration' parameter is the probability of migrating to another deme, whereas if infection is allowed over the contact network, the 'migration' parameter is the probability of an infected host attempting to an infect a host in a neighbouring

deme rather than their own deme. Alternatively, some deme parameters can be specified in the 'Population Structure' section, if the user wishes the same parameters to be applied to every deme. This is also where the network type (fully connected, connected in a line, connected in a star, randomly connected with a user-specified probability, or connected according to user-specified links) and model type (SI, SIR, or SEIR) are defined.

5.2.2 Output

The original DSPS produced five output files: a population log, showing the number of individuals in each state at every step in the simulation time; an event log showing all infection, exposure, and recovery events; and three tree files, including a 'full' transmission tree (containing the identities of the infecting individuals, and sampled and unsampled individuals), a 'pruned' tree containing only sampled individuals, and a 'binary' tree in standard phylogenetic tree format where all internal nodes are forced to be bifurcating (any internal nodes with just one 'child' are removed) (see Figure 5.1).

5.2.3 Algorithm

At the beginning of the simulation, the XML file provided is read in, and the corresponding host and deme objects are constructed, and connected in the manner specified. A random deme from the simulation is chosen, and the first host in that deme is infected to begin the simulation. In order to avoid too many simulations in which the epidemic dies out immediately, the first event generated is always an infection event (rather than a recovery event), and this event is added to the Scheduler's event list. Events in the event list are then run in a loop, each event generating a new event (except sampling events, which do not generate new events), until a stop condition is reached, or there are no more events in the event list.

Generating Events

Events are created by surveying the current state of the population (for example, the number of susceptible and infected individuals), and generating a weighted distribution

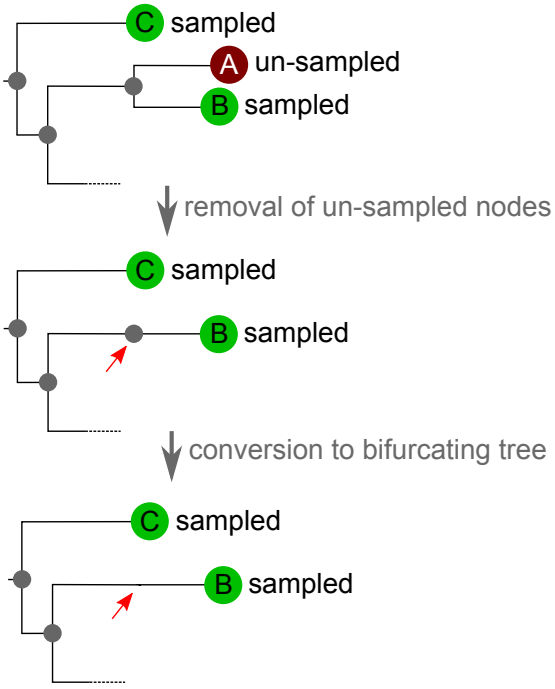

Figure 5.1: Illustration of how the three output trees are generated from the DSPS. The 'full' tree (top) includes all sampled and unsampled infected individuals. Unsampled individuals are then removed to create a 'pruned' tree (middle), which still contains internal nodes reflecting past transmission events (red arrow). Finally, a 'binary' or bifurcating tree is created (bottom) by removing all internal nodes with only one child node.

for the likelihood of each deme to have any event, and then the likelihood of each type of event within each deme.

The amount of time until the next event is determined by selecting a random number between 0 and 1 from a uniform distribution, taking the natural log, dividing by the sum of the Total Hazards, and converting to a positive number. This essentially controls the number of events that take place in a unit of time – the greater the sum of Total Hazards of the demes (the greater the likelihood that event will happen), the more events are generated in a shorter time period. The time interval generated in this manner is added to the current time, and becomes the time the event being generated will occur. (This is an implementation of the Gillespie algorithm (Gillespie, 1976, 1977).)

When generating a new event, the population surveys each deme, and asks it to report on its 'Total Hazard,' or the total likelihood of something happening. The Total Hazard of a deme is the sum of the likelihood that the hosts in the deme: move from 'exposed' to 'infected,' infect another individual, migrate to another deme, recover from

Table 5.1: How the 'hazard' or probability of each event is generated for each deme in the population in the DSPS

Hazard	Condition	Formula
Exposed-becomes-Infected	*only if* SEIR	'exposed' parameter × # of exposed hosts
Infect-Another	SIR/SEIR/SI	'infection' parameter × # of infected hosts
Migrate	*only if* migration of hosts is possible	'migration' parameter × # of hosts
Recovery	*only if* SIR or SEIR	'recovery' parameter × # infected hosts
Total Hazard		Sum of all of the above

being infected, are born, or die. In any given simulation, not all of these events may be possible. For example, in an SIR simulation, there is no 'exposed' category, and so the chance of moving from exposed to infected will always be zero. The 'hazards' for all these events are generated using the user-specified parameters, as described in Table 5.1. (For further explanation of population growth, see Section 5.3.2 and 5.3.7.)

Once the Total Hazard for each deme has been obtained, each deme's Total Hazard is put into a vector, and the sum of all of the Total Hazards is obtained (Figure 5.2 on the facing page (a) and (b)). To chose which deme will perform the event, another random number between 0 and 1 is chosen from a uniform distribution, and multiplied by the sum of the Total Hazards (Figure 5.2 on the next page (c)). The resulting number lies somewhere between zero and the sum of the Total Hazards. The individual deme Total Hazards are then summed in a cumulative fashion, so that an 'upper' and 'lower' boundary for the hazard of each deme is calculated, with the difference between the boundaries being equal to the Total Hazard of each deme (Figure 5.2 on the facing page (d)), so that demes with the largest Total Hazard have the widest boundaries. Whichever of these demes the multiplied random number falls between is the deme chosen to have an event (Figure 5.2 on the next page (d) and (e)).

The chosen deme now generates the type of event that will occur, from its own list of hazards for each event. In a similar manner to how demes are chosen (illustrated in Figure 5.2 on the facing page), the boundaries of each type of event are calculated and a random number is generated to choose which event will occur. If the event is 'exposed-becomes-infected' or 'recovery' an exposed or infected host (respectively) is

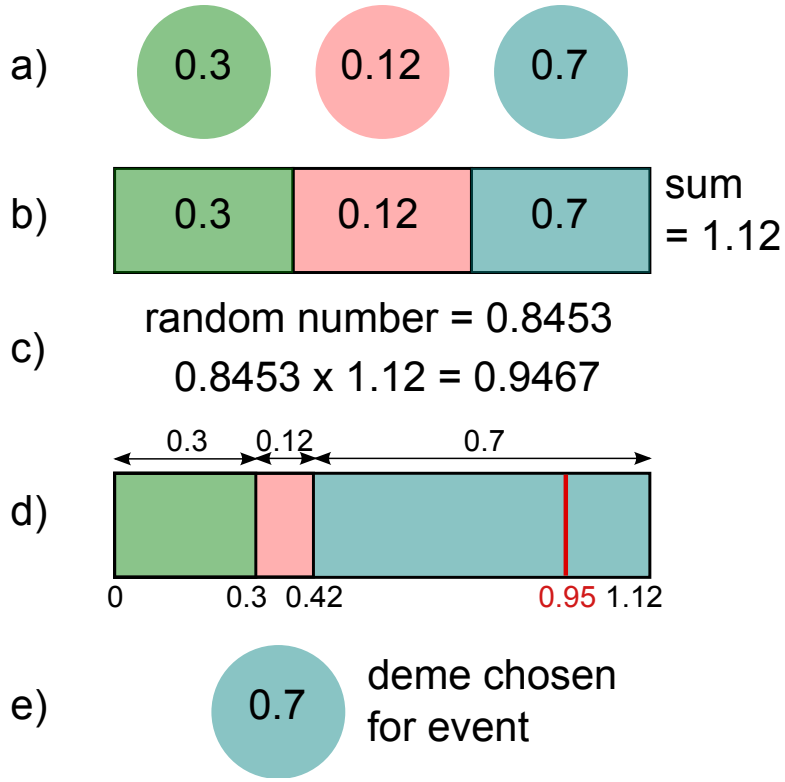

Figure 5.2: Illustration of how a deme is chosen to perform the next event. Three demes with different Total Hazards are available (a), and the Total Hazards are inserted into a vector and summed (b). A random number between 0 and 1 is chosen and multiplied by the sum of the Total Hazards (c). 'Upper' and 'lower' hazard boundaries are calculated for each deme based on their Total Hazard, so that the deme with the largest Total Hazard has the largest boundaries (d). Whichever of the boundaries the multiplied random number falls between is the chosen deme ((d) and (e)).

chosen at random from the deme to perform the event.

If it is an 'infection' event (or an 'exposure' event in a SEIR model), an infected host from within the deme is chosen, then a different host to infect must be chosen. If the model does not allow infection over the network, then another host from within the deme is chosen at random. If the model allows infection over the network, then the migration parameters between the focal deme and its neighbours are used to decide which deme the other host will be drawn from. Each neighbouring deme has a migration parameter associated with it; if these sum to one, a host will always be chosen from a neighbour deme, and never from the focal deme. Otherwise, a host could be chosen from either the focal deme or a neighbouring deme. To choose, the migration parameters are cumulatively summed to form 'boundaries' for the risk of transmission to each deme (Figure 5.3 on the next page). A random number is then generated between 0 and 1

from a uniform distribution. If the random number falls within the risk boundaries of any neighbouring deme, a host is randomly selected from that deme. If the migration parameters do not sum to one and the random number is larger than the sum of the migration risks, a host is randomly selected from the focal deme (Figure 5.3 (b)).

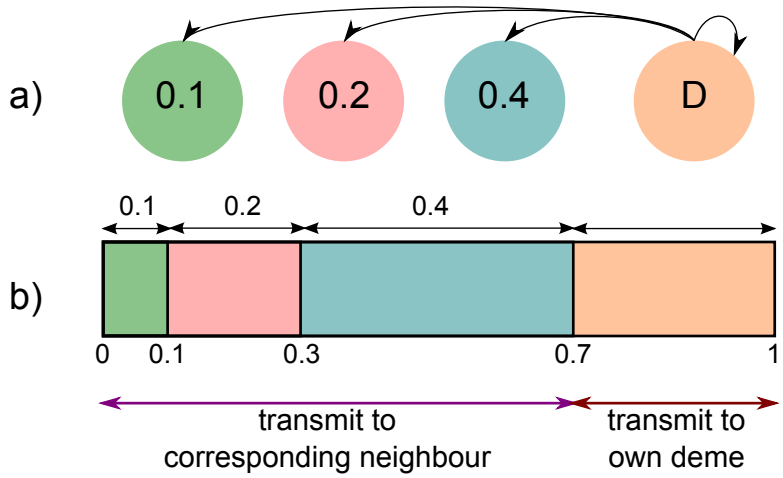

Figure 5.3: Illustration of how a deme is chosen as the recipient of an infection attempt, if infection is allowed along the network. Focal deme D has three neighbouring demes with varying migration parameters (risks of transmitting to these demes) (a). 'Upper' and 'lower' hazard boundaries are calculated for the migration parameters of each deme. A random number is generated between 0 and 1, and where it falls between these boundaries will determine which deme the focal deme attempts to transmit to. If the migration parameters do not sum to one, the focal deme will transmit to itself if the random number is larger than the total sum of the migration parameters (b).

If the event is a migration event, the migration parameters will be used in a similar fashion to choose a neighbouring deme to migrate to (Figure 5.3). Once the required host or hosts have been selected, the event is generated and added to the scheduler list, which is then sorted by the time each event should occur, so that all events happen in the appropriate order.

Performing Events

When the scheduler executes an event, the event is passed to the deme where the event will occur. The success of the execution of the event depends on the type of event. An 'exposure' (in SEIR) or 'infection' (in SI & SIR) event will only be successful if the second (recipient) host is susceptible, otherwise it will fail. 'Exposed-becomes-infected' (SEIR), 'recovery' (SEIR & SIR), and 'migration' events will always succeed.

After an event is performed, sampling events are generated depending on the sampler chosen by the user. If sampling is happening randomly, either a random selection of the active hosts (determined by the probability set by the user) will have sampling events generated for them at set time intervals, or after infection there is a set probability that the newly infected host will be sampled. If sampling is set to occur after recovery, a sampling event will be generated for the host if the event just performed was a recovery.

If an event is performed successfully the event is logged in the log file, and added to the transmission tree if appropriate. If the event was a transmission, a new 'internal' node is created for the transmitting host, connected to the most recent node of the transmitting host by a branch length determined by the time the transmission occurred. Next, a new node is created for the newly infected host, and is connected to the node representing the transmitting host. If the event was a sampling event, a new node is created for the node being sampled, connected to the next most recent node of that host by a branch corresponding to the time of the sampling. This new node is marked as a sampled node, and so will be preserved when the trees is pruned of all unsampled individuals.

Finally, at each cycle of the scheduler, the entire population state is written out to a log file. For each deme in the population, the number of hosts in each state (for example, infected, exposed, recovered) is recorded and printed to output. Thus for every time point at which an event happens, a 'snapshot' is taken of the population after that event.

Finishing the Run

The simulator run can finish in a number of ways. If an event does not complete successfully for some reason, a new event may not be generated, and thus the scheduler will be empty, causing the run to end. For example, if the initially infected individual dies without infecting anyone, no new events will be generated and the simulation will end. Alternatively, the simulation can be set to finish when all individuals are infected or all individuals are recovered.

When the scheduler has stopped running events for any of the reason above, the

DSPS finalizes the output files before completing. The 'full' transmission tree is written out in Newick format, then all un-sampled nodes are pruned, and the 'pruned' tree is written out. Finally, any internal nodes with only one child node are removed, so that the tree is in commonly accepted bifurcating format (see Figure 5.1), and this tree is written out in Newick format.

5.3 Turning the DSPS into an HIV-Specific Model

5.3.1 Modelling Sexually-Transmitted Diseases and HIV

Before mathematical modelling techniques were extended to investigate the dynamics of HIV, they had already been put into practice investigating other sexually-transmitted diseases (STDs), laying the foundation of many of the principles that would later be used and expanded in HIV-specific models. In particular, modelling realistic sexual contact structures proved (and often still proves) challenging. Hethcote and Yorke (1984)'s work on modelling gonorrhoea transmissions through the 1970s and 1980s using compartmental models eliminated many misconceptions about the disease's spread, and was one of the first studies showing the importance of considering variation in sexual activity in STD models. Later studies investigating how different contact models influence transmission demonstrated the 'protective' role of long-term partnerships in curbing epidemic spread of STDs (Dietz and Hadeler, 1988; Waldstätter, 1989).

However, it was recognised that neither random contacts nor serial monogamy accurately reflect the real-life complexity of sexual interaction. 'Concurrency,' when an individual is in two stable relationships that overlap in time, was recognised as potentially being an important driver for STDs, but early models were criticized as allowing either too little or too much concurrency (Dietz and Tudor, 1992; Altmann, 1995).

Stochastic agent-based models started becoming more common in the late 1990s as more powerful computational facilities became readily available, and were used to attempt to address this issue. Studies began implementing individual-based models to investigate more complex sexual mixing and network structure in STD spread and prevention (Kretzschmar et al., 1996; Morris and Kretzschmar, 1997). Despite advances in computational power, these early agent-based models were still criticised for being

limited in the size and complexity of the simulation that could be run, and for modelling sexual contact networks so complex that the specific influences on the epidemic could not be disentangled (Ferguson and Garnett, 2000). New compartmental models were also being developed to model ever-more realistic contact structure, including concurrency, heterogenetity in sexual behaviour, and partner-switching (Garnett and Anderson, 1996; Ferguson and Garnett, 2000).

Modelling HIV

Anderson et al. (1986) recognised that while previously developed mathematical models for STD transmission provided valuable insights, some of the unique features of HIV, such as the lengthy and variable asymptomatic phase, lack of immunity, and (at the time) the still vague definition of 'AIDS' meant that a new model specifically for HIV would be required to study the new disease (Anderson et al., 1986). These first compartmental HIV models provided valuable information about the 'incubation period' (the chronic phase) of HIV and the importance of heterogeneity in aiding the spread of the virus (Anderson et al., 1986; May and Anderson, 1987), even though some of the initial assumptions about the virus were incorrect, such as the parametrization that only 30% of those infected progress to AIDS (May and Anderson, 1987).

Despite the fact that much less was known about HIV at the time, models in the late 1980s and early 1990s correctly predicted that HIV epidemics could turn positive population growth negative (Anderson et al., 1988, 1989, 1991, 1992), explored how assortative and disassortative mixing could lead to very different patterns of HIV epidemic (Jacquez et al., 1988; Gupta et al., 1989), and showed how between- and within-village contacts could explain how the HIV epidemic began to spread widely (May and Anderson, 1990).

Though earlier deterministic compartmental models of HIV transmission had previously confirmed that concurrency of partners could cause the same fast spread after initial infection as previously found in other STDs (Watts and May, 1992), a stochastic agent-based model also investigating concurrency a few years later was able to expand on that finding by showing that stochastic simulations set up with the same concurrency parameters will often, but not always, demonstrate the fast initial growth previously

observed (Morris and Kretzschmar, 1997), demonstrating that the importance of concurrency may vary between populations and outbreaks.

Models have not only been used to investigate population-level transmission dynamics, but also to examine the within-host dynamics of HIV infection. Multiple compartmental models have been used to explore HIV infection on a viral and cellular level by simulating the interaction between HIV, $CD4^+$ cells, and other pathogens (McLean et al., 1991; McLean and Nowak, 1992; McLean, 1993) and investigating the role of immune system escape mutations (Fryer et al., 2010; Palmer et al., 2013).

Both compartmental and agent-based mathematical models have continued to become more complex and realistic, and have been used in HIV research to model vaccine trials (Adams et al., 1998), to investigate the impact of ART, condom use, and other interventions (Blower et al., 2000; van Vliet et al., 2001; Nagelkerke et al., 2002), and to reconstruct the history of the HIV epidemic in the UK (Phillips et al., 2007).

Agent-based and compartmental models both continue to be used today in HIV research, and both still contribute important insights. The continued increase in available computing power has released agent-based models from many of the early constraints around complexity, size, and replicates. However, it's still possible for very large and complex agent-based models to be constrained by the equipment available. And though compartmental models have continued to expand to encompass more complex situations, they cannot reach the individual-level specification possible with agent-based models.

As often is the case where two tools are available, the question of which to use comes down to the job at hand. It's unlikely that enough information about individual cell properties is available to make it reasonable to code an agent-based model to investigate HIV and $CD4^+$ cell interaction, an area where compartmental models have proven highly effective and efficient (McLean, 1993; Fryer et al., 2010; Palmer et al., 2013). On the other hand, agent-based models have proven particularly valuable when individual-level variables are being modelled and assessed, as done in Phillips et al. (2008)'s paper investigating which clinical measures were most beneficial to outcomes in HIV positive patients in resource-limited settings.

Choosing an Agent-Based Model: the DSPS

Here, I wished to use simulation to generate individual-level viral load data and a corresponding viral phylogeny from runs parametrized with different heritability values in order to test my method of heritability estimation. I also wanted the model to allow the generation of simulated sequence data from different epidemic scenarios that could be used to test different phylogenetic methods (see Chapter 6.1). These two requirements meant that an agent-based model was the ideal choice, and the base of the DSPS was used as a starting point from which a HIV-specific simulator could be developed to model complex, realistic epidemics along highly customizable contact networks.

To make the DSPS more appropriate for modelling HIV epidemics, a number of changes were made to the base code. The most important of these were: implementing population growth, birth, and death; adding viral load, which determines transmission risk and disease progression; allowing more complex networks to be specified; adding gender and orientation; and implementing treatment. Some changes influenced only the basic functioning of the DSPS. For example, by changing options in the XML file, the transmission trees can now be output in a different file format and a run time can be set, so that the simulator stops after a specified number of time units. All simulations are run so that infections can occur across the network (between demes), but migration of hosts themselves between demes was not permitted. Thus the 'migration parameters' referred to from here are only used to determine the likelihood of transmitting to another deme, not the likelihood of migrating.

In order to be useful in assessing any phylogenetic method, an HIV epidemic simulator must be able to generate a reasonably realistic phylogenetic tree from the transmission history of the virus, and this tree must trace back to an ancestral root. To accomplish this, the simulator must have only one initial introduction, so that all subsequent infection can be traced to a single ancestor and sequences can be generated that reflect this history. However, in order to get a epidemic of a useful size from just one introduction, the simulation must be allowed to run for a relatively long time, unlike simulators that start with multiple introductions. This is reflected in the DSPS runs

below, where the epidemic does not peak until 30 to 40 simulated years have passed.

Alongside these modifications, some more substantial additions were necessary to create a realistic HIV simulator.

5.3.2 Birth, Death, and ~~Taxes~~ Population Growth

One of the first major additions to the DSPS were birth and death events. Birth events are host-density-dependent within a deme, but are either positively or negatively correlated with the host density, depending on the type of population growth selected. Though a simplification of all possible population growth scenarios, I initially developed just two types of population growth: 'growth' and 'stable.' When 'growth' (default) is specified, the probability of a birth event happening in a deme is the birth parameter (specified in the XML along with the SIR parameters) multiplied by the number of hosts already in the deme – births are positively correlated to host density in a deme. This means that over time, if a deme becomes empty due to death, it can never be repopulated. Conversely, demes that have many births are increasingly likely to have even more births (Figure 5.4 on the facing page).

When 'stable' is specified, the probability of a birth event happening in a deme is the birth parameter multiplied by the difference between the initial number of hosts in the deme and the current number of hosts in the deme – births are negatively correlated to host density in the deme. This means that emptier demes are more likely to have birth events (demes where all the hosts died can be 're-populated'), and no deme will gain more individuals than they had at the start of the simulation, keeping the population relatively 'stable' (Figure 5.4 on the next page).

To simulate a realistic HIV epidemic scenario, all initial simulations attempting to recreate an HIV epidemic used the 'stable' population growth to maintain two people per deme at the maximum.

'Death' events, as first implemented, are part of a random process that equally affects all hosts, regardless of disease/infection status. As might be expected, the probability of a 'death' event in a deme is the death parameter multiplied by the number of hosts in a deme. When generating a 'death' event, a random host is selected from within the deme performing the event. The probability of both birth and death

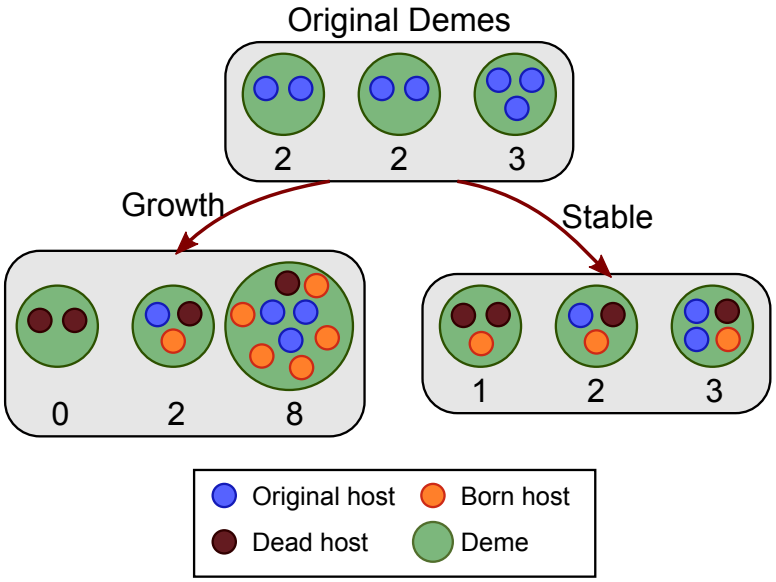

Figure 5.4: Illustration of the difference between 'growth' and 'stable' population growth in the DSPS. Under 'growth,' new births are more likely in demes with a higher host density, and demes where all hosts have died cannot be re-populated. Under 'stable,' new births are more likely in demes with lower host density. Demes where all hosts have died can be repopulated, and no deme can gain more hosts than it had at the start of the simulation. The number of live hosts is shown beneath each deme.

events are included in the calculation of the Total Hazard of a deme, thus influencing how likely that deme is to have an event.

5.3.3 Viral Load, Transmission Risk, and Disease Stages

Viral Load and Transmission Risk

The most substantial change made to the DSPS in order to make it a realistic HIV simulator was the addition of viral load parameters that influence transmission risk and disease progression. To implement this, a 'Virus' class was created in the DSPS, which is an object held by infected hosts. Virus objects hold information like the time the host became infected, the host who 'owns' the virus, and most importantly, the set-point viral load. Viral load information is read in from a text file that contains a distribution of $\log_{10}$ viral loads from a population.

When the first individual is infected in the simulation, the code can be changed so that the initial viral load is either selected randomly from the population distribution, or specified by the user. When a virus is transmitted from one host to another, a

new virus object is created for the newly infected host, with a viral load influenced by the transmitting host's virus (see Subsection 5.3.4 on page 145 for information on the implementation of heritability of viral load).

The chance of having a contact, or attempting a transmission, is not affected by the viral load, as it is dictated by the network structure, infection parameters, and migration rate. Thus, the generation of 'infection' events is no different from before. However, how a deme performs an infection event must change drastically. Previously, the success of an infection event leading to an infection depended only on the second host being susceptible; if this was the case, the infection occurred. With the implementation of viral load, the transmission risk should now be tied to this, and thus not guaranteed to occur in every contact with a susceptible host.

Fraser et al. (2007) provide an equation that relates viral load to per-year transmission risk. This equation is ideal for a simulation as it provides a continuous relationship between the two without the need to partition into viral load 'groups.' However, I wished to check that the estimates given by Fraser et al. (2007) correspond to what has been found in previous studies. Further, while Fraser et al. (2007) gives the probability of transmission per year, in the simulation a probability per event is needed, so a way must be found to convert between the two.

Gray et al. (2001) followed 174 heterosexual monogamous couples in Rakai from 1994 to 1998, of whom 38 couples reportedly infected their partner. The mean number of sexual contacts was 8.9 per month, giving an average of 106.8 per year. Transmitting individuals were split into four groups by viral load: <1,700, 1700-12,499, 12,500-38,500, and >38,500 to estimate transmission risk per sexual act (Gray et al., 2001). I used Gray et al. (2001)'s estimate of approximately 100 contacts per year to convert Fraser et al. (2007) and Quinn et al. (2000)'s per-year estimates to per-act estimates.

Quinn et al. (2000) followed 415 heterosexual sero-discordant couples from Uganda for 30 months, during which time 90 converted. Quinn et al. (2000) found an overall transmission rate of 12 per 100 person-years. They divided transmitting individuals into five categories by viral load: <400, 400-3499, 3500-9999, 10,000-49,999, and ≥50,000 copies/mL, and found the transmission rate per 100 person-years for each category. To convert the transmission risk to risk per act, I first divided by 100 to get risk per year.

Then, based on Gray et al. (2001) I assumed 100 acts per year, and converted to risk per act.

Fraser et al. (2007) re-analysed previously gathered data from Zambia, where 317 sero-discordant couples were monitored at 3-month intervals with viral load at CD4+ cell levels and 109 couples transmitted during the observation period (Fideli et al., 2001). Fraser et al. (2007) fit a flexible parametric model to the data in order to find a relationship between the transmitting partner's viral load and the risk of transmission. Because of their concern about the unreliability of previous reports about the frequency of unprotected sex acts, Fraser et al. (2007) preferred to estimate the transmission risk as a function of time rather than per act. The difficulty in estimating the frequency people engage in unprotected sex is an obvious concern for the DSPS. However, the stochastic nature of the model means that a per-act transmission rate is necessary. As with Quinn et al. (2000), I assumed 100 acts per year to convert the per-year estimate given by Fraser et al. (2007)'s equation to a per-act probability.

I plotted the transmission risk measured by Gray et al. (2001) and Quinn et al. (2000) and predicted by Fraser et al. (2007) on one graph, to see how well these independent estimates corresponded (Figure 5.5 on the next page). As Quinn et al. (2000) and Gray et al. (2001) divided subjects into groups by viral load, I plotted the the transmission risk of that whole 'range' of viral loads as a line (Figure 5.5). As Fraser et al. (2007) provides an equation, I can plot a curve to represent the transmission probability for any range of viral loads (Figure 5.5). To better show how this equation corresponds to the other two paper's estimates, I highlighted the predicted transmission probability at the viral load values that 'divide' Quinn et al. (2000) and Gray et al. (2001)'s subjects into groups by plotting them as empty circles. Encouragingly, the graph shows a very strong correlation between the estimates obtained by observation and by Fraser et al. (2007)'s equation. Given the strong agreement of the three independent studies, I decided to use Fraser et al. (2007)'s equation to calculate transmission potential from the set-point viral load in the DSPS simulation.

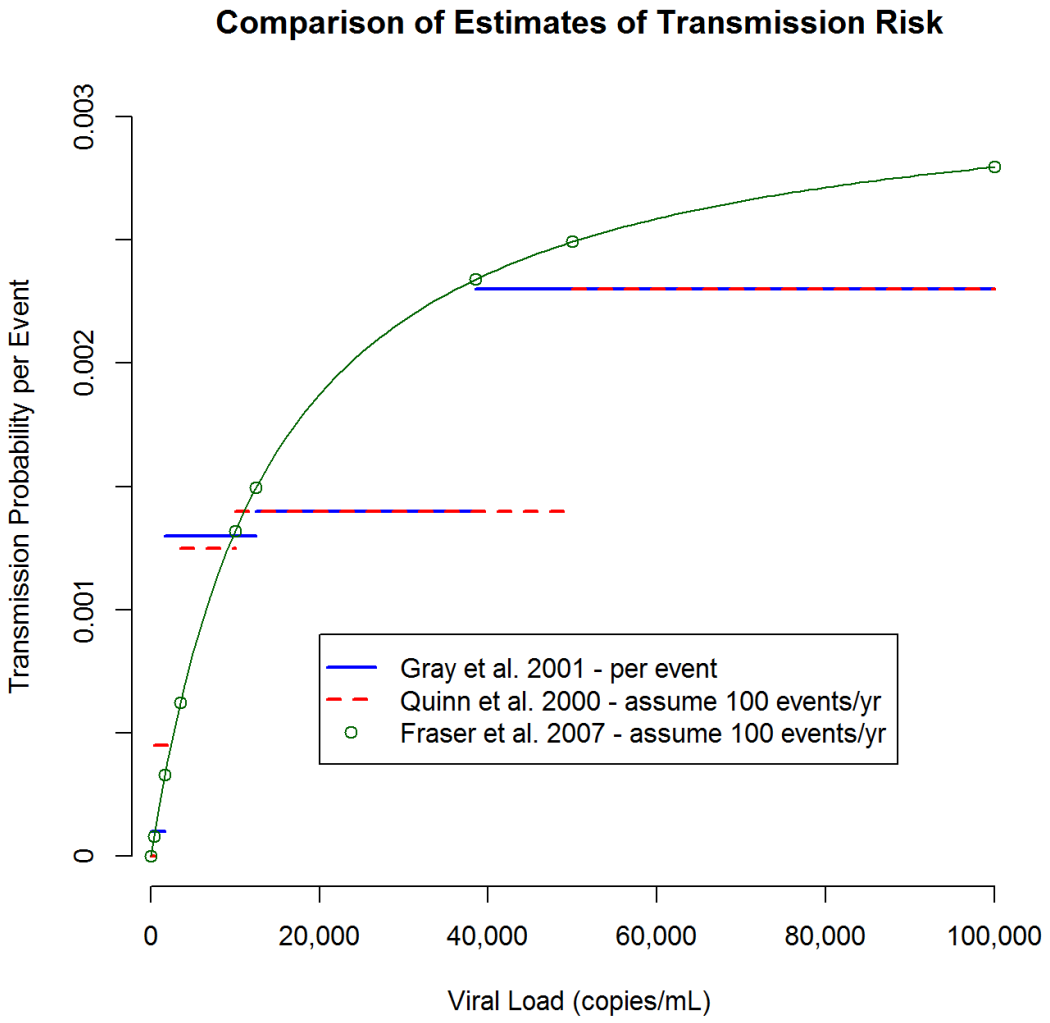

Figure 5.5: A comparison of three different estimates of the transmission potential of individuals with varying viral loads. As Gray et al. (2001) and Quinn et al. (2000) divided transmitting individuals into groups by viral load, the entire range contained in a group was plotted as a line to show the corresponding transmission risk. Fraser et al. (2007)'s equation is plotted as a curve, with the viral loads dividing Gray et al. (2001) and Quinn et al. (2000)'s groups highlighted as empty circles.

The equation given by Fraser et al. (2007) to relate viral load to transmission risk is:

$$\beta(V) = \frac{\beta_{max} V^{\beta_k}}{[V^{\beta_k} + (\beta_{50})^{B_k}]}$$

Where V represents the viral load (unlogged), β_{max} represents the maximum infection rate per annum, β_{50} represents the viral load at which infectiousness is half its maximum, and β_k represents the steepness of the increase in infectiousness as a function of viral load. Fraser et al. (2007) found the values that maximised the likelihood of this equation: β_{max}=0.317/yr, β_{50}=13,938 copies/mL, and β_k=1.02. The resources and data were unavailable to investigate whether different values would better fit my own dataset, so the same values found by Fraser et al. (2007) were used in the DSPS.

With the implementation of viral load, when transmission events are performed, the transmitting host's virus object is queried, and the transmission probability given the set-point viral load is calculated. A random number between 0 and 1 is then generated from a uniform distribution, and if this random number is below the transmission probability, the transmission is successful.

Viral Load and Disease Stage

As well as controlling the risk of transmitting HIV to others, viral load is highly correlated with disease progression, or time to AIDS. HIV has three stages: acute, during which the viral load spikes during early infection; chronic when the viral load is steady (the 'set-point' viral load); and AIDS, when the viral load again spikes as the immune system fails, but the individual is likely very sick and may be less likely to engage in activity that would transmit the virus (see Chapter 1.2).

Chronic and AIDS Stages

Fraser et al. (2007) also provided an equation relating viral load to the length of the chronic phase, and an estimate of the transmission risk during the acute and AIDS stages of infection. The equation is:

$$D(V) = \frac{D_{max}(D_{50})^{D_k}}{[V^{D_k} + (D_{50})^{D_k}]}$$

Where D_{max} represents the maximum duration of the chronic phase in years, D_{50} represents the viral load at which the duration is half its maximum, and D_k represents the steepness of the decrease in duration as a function of viral load. Again, Fraser et al. (2007) found values to maximise the likelihood of this function: D_{max}=25.4 years, D_{50}=3,058 copies/mL, and D_k = 0.41. As previously, the resources and data were not available to independently evaluate these values, the numbers provided by Fraser et al. (2007) were used.

This equation was used to turn the 'recovery' event into a 'removal' event, where the host has progressed to AIDS and is sufficiently sick that they no longer have any contacts through which to transmit HIV. 'Recovery' events were previously completely random, though mediated by the recovery parameter specified for each deme. Since the length of the chronic phase is now known, 'recovery/removal' events now become 'removal attempts.' Removal attempts are still random, affecting any infected individual at any time.

When a removal attempt event is performed by a deme, Fraser et al. (2007)'s equation is now used by the virus object of the host performing the action to calculate how long after infection (the infection 'age') when they should be removed, based on their viral load and when they were infected. Fraser et al. (2007)'s equation gives no error term, and there is obviously variability surrounding when someone will progress to AIDS due to environmental and host factors. To incorporate this into the simulation, a normal distribution with a mean at the calculated removal age and a standard deviation of 1.5 years is generated. The probability of removal at the current time is the cumulative density of the normal distribution at the current age. A random number between 0 and 1 is then generated from a uniform distribution, and if it is smaller than the probability of removal at the current time, the removal is successful. (See Figure 5.6 on page 160 for an example)

Introducing a way to move out of the asymptomatic ('chronic') phase and into AIDS raises a new problem, however. Without disease-mediated death, individuals will now progress to AIDS and stay in 'AIDS' phase for unrealistic amounts of time, until they are killed off by chance. In 'stable' population growth, new 'susceptible' individuals cannot move or be born into the simulation until AIDS-stage hosts have died, so the

simulation reaches an unrealistic point where the HIV epidemic is constrained by too many hosts in AIDS-stage who cannot infect others or be infected.

The solution to this problem was to introduce disease-related mortality after hosts progress to AIDS, on top of the random death that equally affects all individuals in the simulation. After each successful removal event, a death event is generated to guarantee that AIDS-stage hosts will only survive a realistic amount of time. (Fraser et al., 2007) estimated approximately 1.58 years were spent in AIDS before death, which I decided to round to an average of 2 years in AIDS prior to death. To calculate how long the host will live with AIDS before dying, a normal distribution with a mean of 2 and a standard deviation of 0.5 is generated to represent the range of time period between AIDS and death. A random number between 0 and 1 is drawn from a uniform distribution, and this is used to calculate the inverse cumulative probability of the normal distribution (the number generated is the 'cumulative probability,' and this is used to generate the time until death from the corresponding x-value on the normal distribution). A check was put in place to ensure that the time until death must be greater than zero – an individual cannot die of AIDS before they have progressed to AIDS. The death event is then added to the scheduler to occur at the chosen amount of time after the current time. Thus the vast majority of individuals live between 1 and 3 years with AIDS before dying.

There is a small chance that the host will die due to a random death event before their 'scheduled' death. In this case, the simulator recognises the situation and disregards performing the death event. 'Scheduled' death events are also marked as such in the scheduler, and so do not generate new 'random' events after being performed.

Acute Stage

To complete a realistic implementation of the stages of HIV infection in the DSPS, I wished to implement an acute stage. As part of their own models, Fraser et al. (2007) concluded that the height of the 'spike' in viral load seen during acute infection differs little between individuals, and is not related to what the set-point viral load will be during the chronic phase. Thus, they estimated a constant transmission risk for all individuals for the acute phase of 2.76 per year, with the acute phase lasting 0.24 years

(Fraser et al., 2007; Hollingsworth et al., 2008). Assuming 100 contacts per year, as before, this gives a per-act transmission risk of 0.0276. In order to maintain realistic transmission dynamics, no transmissions are allowed to take place during the first 2 weeks (0.038 years) of infection.

Implementing this in the DSPS was relatively simple. When a transmission event is attempted, the virus calculates the transmission probability from the set-point viral load, as specified previously. A simple check was added, to see if transmission was being attempted fewer than 0.24 years after infection. If so, the usual transmission risk equation is not used, and instead the transmission risk of 0.0276 is returned, regardless of viral load.

Updating the Input and Log Files

With the introduction of transmission and removal being dependent on viral load, not all events that are attempted complete successfully. Thus, the function of the 'infection' parameter had to change. This previously represented the average number of infections per time unit, but now that not all transmissions will be successful, the 'infection' parameter represents the number of attempted transmissions, or contacts, per time unit. As discussed above, I implemented a contact rate of 100 unprotected sexual acts per time unit (year).

In the original event log file format, only successful events were recorded, but this no longer gives a complete picture of what is happening in the DSPS. Hundreds or even thousands of events are now attempted but are not successful, particularly with regard to transmission events. I added an option to the XML file input where the user can specify whether the DSPS should record only successfully events, or record all events. If 'all events' is chosen, then all attempted events are recorded alongside successful ones, with the success status for each event recorded as 'true' or 'false.' This allowed me to get an idea of how many events were attempted and what fraction of them were successful in order to assess the parameters being used.

With the increase in the number of events per time period and the larger population sizes being implemented, it became computationally impractical to produce the population log file that printed out the number of hosts in each state after every attempted

event. The population log files grew to over 3GB in size, and were taking 50% of the simulation's computational time to produce. Thus, I disabled the production of the population log file in the DSPS.

5.3.4 Heritability of Viral Load

Given that one of the main drivers for producing an HIV-specific simulator was to enable the possibility of testing my phylogeny-based method for estimating heritability of virulence (Chapter 2 on page 23) on simulated data, incorporating heritability of viral load into the DSPS was vital.

Alizon et al. (2010) developed a simulation to verify their own phylogeny-based method, using an SIR model to generate twenty phylogenies over thirteen generations, with a probability of transmission of 0.75 (modelled by branching in the tree), and a probability of death of 0.25. When transmission occurred, one 'child' branch retained the viral load of the 'parent,' while the viral load of the other 'child' branch was determined by the relationship:

$$x_{a+1} = \zeta x_a + (1 - \zeta)y$$

where x_{a+1} represents the viral load of the new 'child,' x_a represents the viral load of the 'parent,' y represents a random value drawn from the empirical distribution of viral loads in the population, and ζ represents the heritability of viral load. (In passing, it is worth pointing out that under a Brownian motion model (assumed in the main heritability estimation analysis) both 'child' nodes would assume new viral load values according to these equation, which is not what happens in Alizon et al. (2010)'s simulations.)

I utilized the same equation in the DSPS, but use the square root of the heritability and one minus the heritability in order to prevent the loss of variance that occurs at each transmission if heritability is used directly. Thus the equation used is:

$$x_{a+1} = \sqrt{\zeta} x_a + \sqrt{(1 - \zeta)}y$$

However, this equation is again imperfect – it causes viral load values to increase to unrealistic levels over time. This is because while ζ and $1-\zeta$ sum to one, $sqrt\zeta$ and $sqrt1-\zeta$ sum to more than one, meaning the new viral load will always be larger than the parental viral load. To prevent the viral load increasing to unrealistic values, all viral loads are standardized by the population mean (z), giving the equation:

$$x_{a+1} = z + [\sqrt{\zeta}(x_a - z) + \sqrt{(1-\zeta)}(y - z)]$$

Unfortunately this equation also created problems in generating heritable viral loads, which is discussed fully in Chapter 6.2.3 on page 187.

The value of heritability can be specified by the user in the input XML file. Whenever a transmission is successful, the set-point viral load for the newly infected virus is calculated using the above equation. The set-point viral load of the transmitting virus is used as the 'parent' viral load (x_a), and a random viral load (y) is drawn from the distribution of viral loads read in at the beginning of the simulation (Subsection 5.3.3 on page 137). Along with the specified heritability, these viral loads are used to calculate a new set-point viral load (x_{a+1}), which is assigned to the new virus object now owned by the newly infected host.

5.3.5 Implementing More Complex Networks

DemeGroups in the XML

In the original implementation of the DSPS, connections to neighbouring demes and the risk of transmitting across those connections had to be explicitly specified for every deme in the network (Code 5.1). For a small number of demes, this is an acceptable way of connecting neighbours, but as the number of demes in the simulation grows problems arise. First, it becomes nearly impossible to create the XML by hand – there is simply too much typing to specify every neighbour for every deme. For demes allowed to transmit to themselves, the migration parameters of the neighbours should sum to one minus the probability of the deme transmitting to itself, and calculating and dividing the migration parameters for all neighbours, especially if some neighbours should have different connections, becomes a complex process. The file also begins to grow to an

unwieldy size, making it difficult to open in text editors and time-consuming to read in to the simulator. Finally, the simulator itself must store these thousands of connections in memory, slowing down the entire simulation process.

Code 5.1: Original DSPS XML code illustrating how neighbouring demes and the probability of transmitting to them has to be explicitly specified

```
<DSPS>
<General>
   <parameter id="Seed" value="12345"/>
   ...
   <parameter id="RecordAllEvents" value="true"/>
</General>
...
<Demes>
   <Deme>
      <parameter id="DemeUID" value="0"/>
      <parameter id="DemeName" value="House1"/>
      <parameter id="NumberOfHostsPerDeme" value="2"/>
      <parameter id="InfectionParameters" value="100,0.8"/>
      <parameter id="Neighbours" value="HouseHigh2,HouseLow3,HouseLow4 ...
          HouseSW100"/>
      <parameter id="MigrationParameters" value="0.005,0.002,0.002 ... 0.007"/>
   </Deme>
```

To overcome these problems and make the specification of networks easier, I implemented 'DemeGroups.' DemeGroups assumes a network is fully connected, but that there are different types of connections between different types of demes. By assigning each deme to a DemeGroup and then describing the connections between the different DemeGroups, a complex network can be specified with much less XML code (Code 5.2). To use DemeGroups, a parameter is added to the 'General' XML section listing the DemeGroup names that will be used. Then, when each deme is defined in the XML code, the DemeGroup it belongs to is specified. 'NeighbourDemeGroups' is used to specify which DemeGroups it has connections to, and 'MigrationParameters' is used to describe the probability of migrating/transmitting to demes in each of these groups.

A typical use of DemeGroups in the HIV simulation runs using the DSPS, is shown in Code 5.2. Each deme is assigned to one of three DemeGroups: 'High,' 'Low,' and 'SW,' designating demes with high risk, low risk, and as a 'sex worker' deme. The majority of the demes in the simulation are two-person 'household' demes and will be assigned to the 'High' or 'Low' DemeGroup, with 'High' demes having more sexual

contacts outside their household than 'Low' demes. There is one 'sex worker' deme holding 200 hosts who have higher contact rates than normal households demes. The first deme in Code 5.2 is assigned to DemeGroup 'High,' and has connections to all three other DemeGroups. The risk of transmitting to other demes in the 'High' DemeGroup is 0.2, to demes in the 'Low' DemeGroup is 0.1, and the risk of transmitting to the 'SW' DemeGroup is 0.2. As these sum to only 0.5, the risk of transmitting to its own deme is $1 - 0.5 = 0.5$. The connections between DemeGroups are very flexible, with one-way connections and connections with differing risk in different directions possible.

Code 5.2: Modified DSPS XML code illustrating how neighbours are grouped into different DemeGroups and transmissions between these DemeGroups is specified

```
<DSPS>
<General>
   <parameter id="Seed" value="12345"/>
   ...
   <parameter id="RecordAllEvents" value="true"/>
   <parameter id="DemeGroups" value="High,Low,SW"/>
</General>
...
<Demes>
   <Deme>
      <parameter id="DemeUID" value="0"/>
      <parameter id="DemeName" value="HouseHigh1"/>
      <parameter id="DemeGroup" value="High"/>
      <parameter id="NumberOfHostsPerDeme" value="2"/>
      <parameter id="InfectionParameters" value="100,0.8"/>
      <parameter id="NeighbourDemeGroups" value="High,Low,SW"/>
      <parameter id="MigrationParameters" value="0.2, 0.1, 0.2"/>
   </Deme>
```

DemeGroups in the Code

With DemeGroups implemented, the underlying functionality of the code changes surprisingly little, as the much of the code simply treats the DemeGroups the same as actual neighbouring demes. If DemeGroups is turned on by specifying the DemeGroups in the 'General' part of the XML, each deme that is added to the population is stored in a list of demes held in a hash table, where the key is the DemeGroup name. Thus, using just the DemeGroup name, all the demes in that group can be accessed.

If a transmission event is being generated, each of the DemeGroups connected to the focal deme is treated as if it was an actual deme, allowing a DemeGroup (or the

focal deme) to be selected exactly the same as before (Figure 5.3 on page 130). If a DemeGroup is selected to be transmitted to, a random deme (which is not the focal deme) is selected from the chosen DemeGroup, and a random host selected from that deme.

Gender and Orientation

Most connections in real sexual networks are determined primarily by two things: gender and orientation. In implementing these two attributes, the simulator automatically begins restricting contacts across the network. In a simulation without gender and orientation, an infected individual can infect anyone else they have a connection to, but a heterosexual infected host in a gender-balanced population will only have contacts with half the available population.

To implement gender, there first had to be a way for the user to specify the gender of hosts within a deme in the XML file. A new deme parameter, 'NumberOfMaleFemaleHosts' was added, where users can specify the number of males and females in the deme, separated by a comma. 'Gender' must be turned on in the 'General' part of the XML for NumberOfMaleFemaleHosts to work. If 'Gender' is on but NumberOfMaleFemaleHosts is unspecified, genders are assigned at random to the hosts in the deme. If the number of male and female hosts specified is fewer than the number of hosts in the deme, the number of males and females specified is added to the deme, and all remaining hosts are randomly assigned gender. If the number of male and female hosts specified is more than the number of hosts in the deme, then the specified genders are allocated to the hosts, males first, until all hosts have been assigned a gender. (The onus is on the user to prevent a potentially very male-biased population by ensuring that the number of each gender specified and the total number of hosts match.)

Once the simulation is running, any new hosts born into demes must be assigned gender in a sensible way. If NumberOfMaleFemaleHosts was never specified for a deme, then the gender is randomly assigned, as before. Gender is also randomly assigned if there are both fewer males and fewer females than was specified, or if there are enough (or too many) of both males and females. If there is a shortage of just one gender, however, the new host will be assigned to that gender. If the user wants to ensure that

a deme contains only one gender, they can use the 'GenderOnly' option to specify which gender any new individuals born to the deme should be. Normally, a deme originally created with 200 females under 'Growth' population growth would add new hosts as female until 200 females were present, and then start assigning new hosts random gender. Using 'GenderOnly' set to 'Female,' even when 200 females are present, new hosts will only be assigned female gender.

Sexual orientation is assigned in a similar fashion. It first must be turned on in the 'General' part of the XML, but will only be implemented if 'Gender' is also on – otherwise an error is thrown, as orientation without gender is not possible. 'OrientationChoice' can be specified at the deme level (all hosts in a deme have the same orientation), or at the population level, which is applied to all demes in the population. Orientation can be set to heterosexual, homosexual, or bisexual. If 'Gender' is on but 'Orientation' is not, or hasn't been specified, all individuals are bisexual by default. Since Orientation is specified at the deme or population level, any new individuals born into a deme will take on the Orientation of that deme.

Putting gender and orientation into action when choosing an appropriate transmission partner is slightly more tricky. Previously, any host connected to the focal deme could be selected as a transmission partner, and so a random host only had to be selected once. Now, the host must be of a gender and orientation that is complementary to the transmitting host. Because a suitable host might not be located on the first try, a loop was added so that a deme and host can be randomly selected until a suitable match is found, up to 200 attempts. If no suitable partner has been found after 200 attempts, an error is thrown and the simulation is stopped so the user can examine whether the network they have created is realistic.

To evaluate the suitability of a random host, the new host is given to the transmitting host object. Strict rules are implemented to decide who will have contact with whom. I found it was easier to create rules about who hosts would **not** have sexual contact with, even though this is slightly counter-intuitive. The rules are shown in Table 5.2 on the next page.

Table 5.2: DSPS sexual contact rules based on gender and orientation. The table indicates what kind of hosts the transmitting host will **not** have sexual contact with

If the transmitting host is...		*They will **not** have sex with...*		
Orientation	**Gender**		**Orientation**	**Gender**
Heterosexual	Male		Any	Male
		or	Homosexual	Female
	Female		Any	Female
		or	Homosexual	Male
Homo	Male		Any	Female
		or	Heterosexual	Male
	Female		Any	Male
		or	Heterosexual	Female
Bisexual	Male		Homosexual	Female
		or	Heterosexual	Male
	Female		Homosexual	Male
		or	Heterosexual	Female

5.3.6 Treatment

One very important potential use for an HIV simulator is to look at the effect of different types of treatment. Implementing treatment in the DSPS allows users to implement many different treatment strategies, including different treatment roll-out speeds, percentage of the population receiving treatment, and when in the epidemic treatment should be started.

There are currently some limitations in the implementation of treatment in the DSPS. The largest is that the user must decide which households will receive treatment when creating the XML file. This largely relies on the assumption that demes are equally likely to be infected. So, if treatment should reach 20% of those infected, the user must select 20% of the demes as 'eligible' to receive treatment, and assume that the infection will spread randomly, and that 20% of infected demes will be those who can receive treatment. If there are some demes that are more or less likely to become infected, the user must decide whether or not to include these demes as 'eligible' for treatment. In a population with a small number of demes, there is always a chance that either the infection is only in 'treatment-eligible' demes (so 100% of the infected individuals are treated) or that the infection is not in any 'treatment-eligible' demes (so 0% of the infected individuals are treated). The network originally created to run

the HIV simulations contains more than 4,000 two-person households, which is a large enough size that choosing 20% of the households at random to be 'eligible' for treatment does approximate to treating 20% of the population, but it is recognised that in future updates to the DSPS this is a major problem that needs addressing.

Specifying Treatment

To use treatment in a simulation, any deme 'eligible' for treatment should have a TreatmentParameter added to the deme description in the XML. The value of the treatment parameter is the 'risk' of treatment, which translates in practice to how quickly the deme will receive treatment. The treatment parameter becomes part of the Total Hazard of a deme when demes are being chosen to generate events, and influences the likelihood of a treatment event occurring in a deme.

The presence of the TreatmentParameter variable is not enough to implement treatment, however. In the 'PopulationStructure' part of the XML, the 'TreatmentTimer' parameter must also be present. If TreatmentTimer is set to 0, treatment will be available immediately when the simulation starts. If the TreatmentTimer is set to some other number, treatment will be 'turned on' when the simulation reaches the time specified. If or while treatment is 'turned off,' the treatment hazard, or likelihood of a deme having a treatment event, is returned as zero for every deme. When treatment is turned on at the specified time, the treatment hazard is returned as the treatment parameter multiplied by the number of infected hosts in the deme. Demes without a treatment parameter will always have a treatment risk of zero.

If a deme has a larger treatment parameter, it is more likely to start treatment soon after treatment is turned on, or sooner after infection if treatment is already on, while a lower treatment parameter will cause the hosts in that deme to be slower starting treatment.

How Treatment Works

All treatment events, once generated, are successful. Treatment itself is also completely successful and 100% effective, with no resistance or inconsistent adherence. Treatment works in a very simple way: the host remains infected, but the viral load is reduced

to 50 copies/mL. Because transmission risk and disease progression are based on viral load, this one step both makes transmissions from treated individuals very rare, and increases their life span by increasing the chronic phase of the disease according to Fraser et al. (2007)'s equation (Subsection 5.3.3 on page 141). In a host where the set-point viral load was 4.5 $\log_{10}$ copies/mL, treatment will lengthen the calculated chronic stage from 7 to 21 years.

Because treated hosts are still infected, they are still counted as infected hosts for the purposes of generating and performing transmission and removal events. Hosts on treatment have as many sexual contacts as infected hosts not on treatment, but they have a much lower risk of transmission. They are also just as likely to attempt a removal event (progressing to AIDS) – though it is unlikely to be successful until near the end of the calculated chronic stage.

When hosts on treatment do successfully undergo a removal event, they are treated in the same manner as infected hosts who are not on treatment. After removal they are considered to have AIDS, and so have no contacts, and a death event is generated to take place a mean of two years after removal. Treating hosts on treatment as if they progress to AIDS is another limitation of treatment as currently implemented. However, without any kind of age or age-associated death in the DSPS, it ensures that individuals on treatment do not live on indefinitely.

With the implementation of treatment, a new sampler was created in order to sample 'just before treatment and recovery.' This sampler extends the original 'just before recovery/removal' sampler so that a sample is also taken just before treatment is started, as would often be done in a real-world scenario. Since patients on treatment will also progress to removal, they will be sampled twice.

5.3.7 Exponential Population Growth

In a realistic setting, very few populations are 'stable' and have no growth. Using the birth parameter in the 'Growth' setting of population growth, exponential growth, which is found in most populations, could probably be approximated, but would mean allowing demes to grow very large, rather than keeping them as two-person households. A means to achieve exponential growth while retaining the overall structure of the

network was needed. An illustrative example of how exponential growth works is shown in Figure 5.7 on page 161.

Demes and Exponential Population Growth

The DSPS has no way of creating or destroying demes, and though adding a way to do this would be one way of addressing the problem, creating and connecting new demes appropriately in the middle of the simulation would be a complex task. Instead, I decided to create the entire network that would be needed at the populations largest size, and instead limit the amount of the network that was active.

Using scripts in R, an XML file can be created so that given a starting population size and a growth rate, the network can be created as it would be at the end of the simulation. Each deme now has another parameter, 'DemeStatus,' which determines whether or not a deme is 'on.' All of the demes in the XML are turned off except a random selection matching the starting population size. When a deme is off, all the hazards of that deme, including the Total Hazard, are zero, so the deme will generate no events and be subject to no events. If 'DemeGroups' is active, demes that are off are not added to the DemeGroups. The deme also contains no hosts, though the maximum number of hosts allowed in the deme and the gender of those hosts can be specified for when the deme is turned on.

Exponential growth is turned on by setting the 'BirthDeathType' to 'exponential' in the 'PopulationStructure' section of the XML, and specifying a growth rate using 'ExpGrowthRate' in the 'General' part of the XML. When exponential growth is on, the simulator will pause at the end of every time unit (year) and attempt to randomly turn on as many demes as are expected from the growth rate and the current number of demes. If there are not enough 'off' demes available to turn on to grow the population by the required amount, the simulation will throw an error and stop.

When a deme is turned on, it is added to its respective DemeGroup (if applicable), and the deme's risk hazards behave as normal for an empty deme. The deme is not automatically populated, but is it now available to have 'birth' events. If the maximum number of hosts and the desired gender of the hosts was specified for a deme, it will only be populated to these parameters.

Exponentially Growing Demes

In some cases, however, it's possible that the user does want a deme to grow exponentially by adding more hosts at an appropriate rate. For example, in the now exponentially-growing population of 'High' and 'Low' risk two-person households, the 'sex worker' deme will remain the same size. For this deme to also grow at a controlled exponential rate, some new parameters must be implemented.

To show that the sex worker deme should grow at the same exponential rate as the rest of the population, the maximum number of hosts for the deme is set to 'exp' in the XML file. This sets the maximum number of hosts in the deme to the initial number of hosts in the deme (in this case, 200), and also causes the deme to be added to a special list of all demes that should be 'grown' at an exponential rate. At the end of each year, after the simulator finished turning on demes, each exponentially growing deme has the maximum number of hosts allowed in their deme incremented by the number of hosts in the deme times the exponential growth rate. As with the demes that have just been turned on, new hosts are not automatically added to the deme, but they now have an increased chance of being subject to birth events.

5.4 Generating Viral Phylogenies and Simulated Sequences

Though the DSPS does generate phylogenies based on the sampling scheme selected, this type of sampling may not be very realistic in some situations. Because the sampling scheme cannot be changed part-way through the simulation, samples are taken throughout the run, going back to the initial stages of the epidemic. In reality, it's very rare that samples would go back so far, and more likely that samples cover a discrete time period. In developed countries this might be a time period of 10-15 years, starting from when resistance testing became wide-spread; in other circumstances it might cover a short period during a clinical trial or cohort study.

To allow more realistic sampling and generation of viral phylogenies and sequences, an alternative pipeline was developed. An illustrated example of the pipeline can be seen in Figure 5.8 on page 162. An in-house R script was written to generate random samples of the hosts that were alive and infected during a specified time period of the

run. To maintain realistic sampling parameters, sampling is not allowed until 3 months (0.25 years) after infection. The R script generates a 'line list' of all infection events, the time they occurred, and the transmitting and seroconverting host, with the generated sampling events added to the list at the appropriate time.

A Java program called 'VirusTreeSimulator' was developed specifically by Matthew Hall to generate viral phylogenies based on the line lists generated by the above R script from DSPS output. The VirusTreeSimulator simulates within-host phylogenies such that variation in the effective population size within each host obeys the logistic function:

$$N_e(t) = \frac{N_0(1 + e^{-rT_{50}})}{1 + e^{-r(T_{50}-t)}}$$

Where $t = 0$ represents the point of infection and t increases as time moves into the past . As a result, the entire infection of a host takes places in 'negative' or 'backwards' time. N_0 represents the effective population size (N_e) (as population size $\times$ generation time) at the point of infection, r the growth rate during the exponential phase, and T_{50} the time at which N_e is equal to half its final asymptotic value.

After viral phylogenies are generated, a program called piBUSS (Bielejec et al., 2014) can be used to simulate sequences down the phylogenies. piBUSS uses ancestral sequences as a starting point, and models how mutations arise at random along the tree branches according to the HKY substitution model, allowing different mutational parameters to be used for the 1st and 2nd codon position and the 3rd codon position. To do this, appropriate ancestral sequences must be provided by the user. These can be generated in BEAST from a collection of samples from the subtype and population being modelled. (See Chapter 6.1.3 for an example of how this is done.)

5.5 Future Developments

5.5.1 Near Future

Though the modified DSPS is already a flexible and realistic HIV simulator, there are several areas where it could be improved, allowing even more realistic simulation.

Currently, the only difference between males and females is who they have contacts

with when their sexual orientation is specified – a network of bisexual males and females is essentially the same as a genderless network. In future iterations of the DSPS, however, it will be possible to specify differences in behaviour, disease progression, and transmission risk by gender and the type of contact. For example, male-male contact could be most risky, and heterosexual contact risk could differ depending on which partner is infected.

While the addition of DemeGroups allows complex networks to be specified, including both long-term partners (in two-person households) and random sexual encounters, there is currently no way to implement concurrency or parametrised partner switching, which is recognised as an important driver of HIV epidemics (Watts and May, 1992; Morris and Kretzschmar, 1997). Implementation of concurrency and partner-switching would allow better representation of real sexual network structure, and could be incorporated into the DSPS in future.

The DSPS currently moves individuals from 'infected' to 'removed' as a representation of moving from the chronic stage of HIV to pre-AIDS and AIDS (see Section 5.3.3). Once 'removed' individuals have no contacts, as they are considered to be too ill to participate in behaviour that would lead to the spread of HIV. This is a rather simplistic view, and one that should be re-evaluated in future DSPS iterations to more closely model what may be occurring in real-life.

Though care was taken to evaluate Fraser et al. (2007)'s equation associating viral load with transmission risk by comparing the results with other independent studies (see Section 5.3.3), the same comparison was not done for Fraser et al. (2007)'s parameters dictating transmission risk during the acute phase (section 5.3.3) and the length of the chronic phase (section 5.3.3). To ensure that the equations and parameters being used in the DSPS accurately represent what is supported by HIV research, further work should be done to evaluate these assumptions against other studies.

Finally, there is currently no way to stop or change treatment regimes or roll-out speed during the simulation, but this is something that could be implemented in future updates of the DSPS. Some low level of treatment failure due to resistance and the effects of inconsistent adherence could also be incorporated. Treatment also currently has no effect during the acute stage of infection, and patients with 'AIDS' are not

eligible to receive treatment and move back into the 'chronic' phase of the disease, but future modifications to the DSPS hope to address these issues.

5.5.2 Potential Future Use

As detailed earlier (Section 5.3), compartmental models have been successfully used to investigate within-host dynamics of HIV and its interaction with other pathogens and $CD4^+$ cells (McLean et al., 1991; McLean and Nowak, 1992; McLean, 1993; Fryer et al., 2010; Palmer et al., 2013). HIV is thought to persevere in 'reservoirs' of long-lived cells that permit low levels of viral replication (sometimes called 'cryptic viremia') even while on effective ART (Finzi et al., 1997; Chun et al., 1997, 2008), and compartmental models have also been used in exploring the impact of these reservoirs on within-host dynamics and the three stages of HIV infection (Hadjiandreou et al., 2007; Cardozo et al., 2012; Hernandez-Vargas and Middleton, 2013; Knorn and Middleton, 2014). HIV viruses originating from these reservoirs can be genetically distinct from the most common viruses circulating in the blood and those found in other reservoirs (Potter et al., 2004), and may differ again from viruses found circulating after discontinuing ART (Zhang et al., 2000). As more next-generation sequencing is done on HIV samples, more information will become available about the low-level variants circulating within patients.

The DSPS has the ability to be re-parametrized as a within-host model, with 'demes' representing reservoirs or different tissues and 'hosts' as infected cells. Though much has been learned through compartmental models about the dynamics of within-host infection, agent-based models have the ability to directly track the virus as it infects different types of cells and invades reservoirs, possibly providing valuable insights about how genetically distinct variants of HIV diverge, proliferate, persist, and re-emerge within patients. A model of this complexity would take considerable work in re-parametrizing to ensure realistic behaviour, but could prove very useful as more information about the full diversity of the HIV population in a patient becomes available.

5.6 Use of the DSPS

With the modifications made, the DSPS is now a highly flexible yet specialised simulator that can be used to model HIV epidemics in a variety of simulations. The potential applications of the DSPS are numerous. The DSPS is already being used to evaluate phylogenetic analysis tools on their ability to detect changes in incidence and prevalence and reconstruct time-dated trees (see Chapter 6), and is also being used to investigate how best to reconstruct accurate contact networks. In the future, the DSPS could simulate different treatment strategies and how different assumptions about contact patterns and transmission risk influence the effectiveness of different treatment roll-out.

In the context of this thesis, however, the DSPS, with its realistic integration of viral load and heritability, allows the simulation of HIV epidemics with varying heritability, which can then be used to objectively test the method described here.

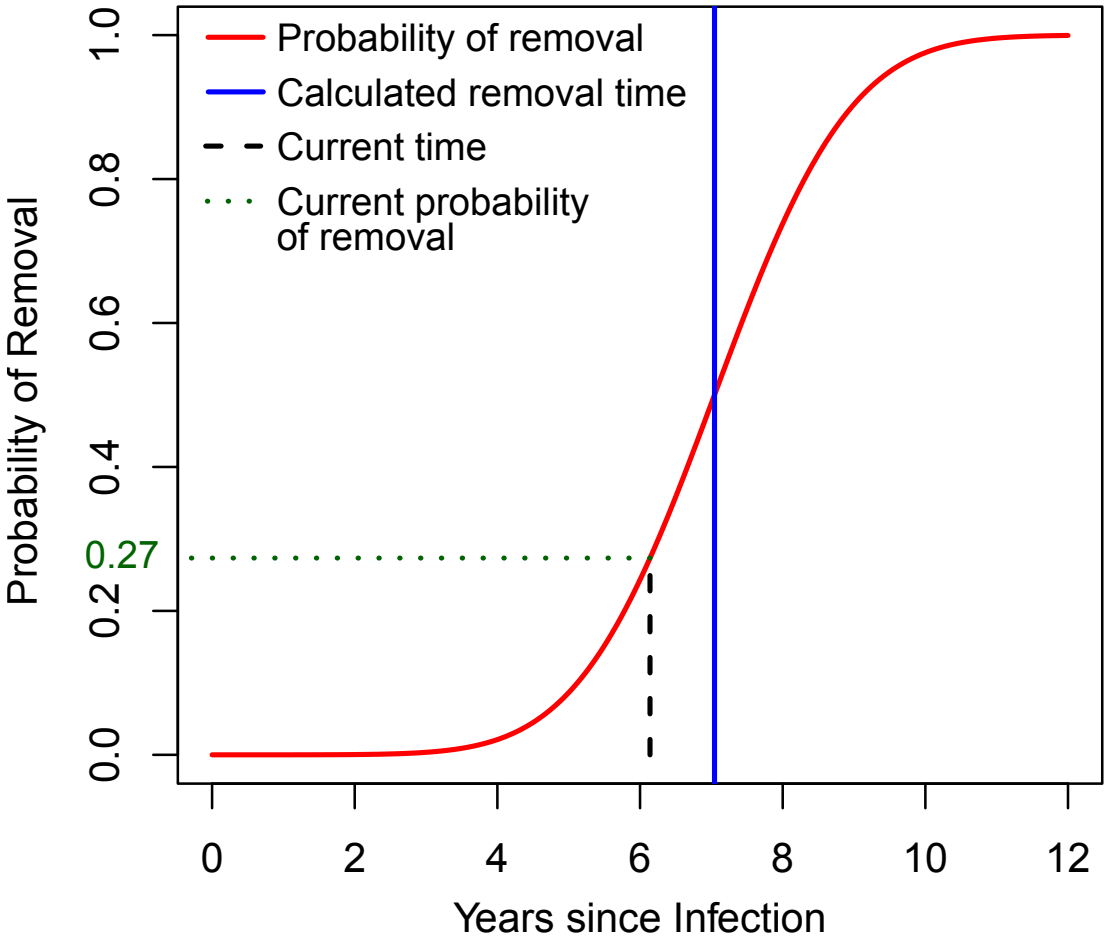

Figure 5.6: An example of how probability of successful removal is calculated. An example host has a viral load of 31,623 copies/mL, and has been infected 6.14 years at the time the removal event is performed. The calculated length of the asymptomatic phase is 7.04 years, so a normal distribution with mean of 7.04 years and standard deviation of 1.5 years is generated. The cumulative density of this distribution at the current infection length of 6.14 years is calculated as 0.27 – the probability of removal at the current time. A random number between 0 and 1 is generated from a uniform distribution, and if it is smaller than the removal probability at the current time, the removal event will be successful.

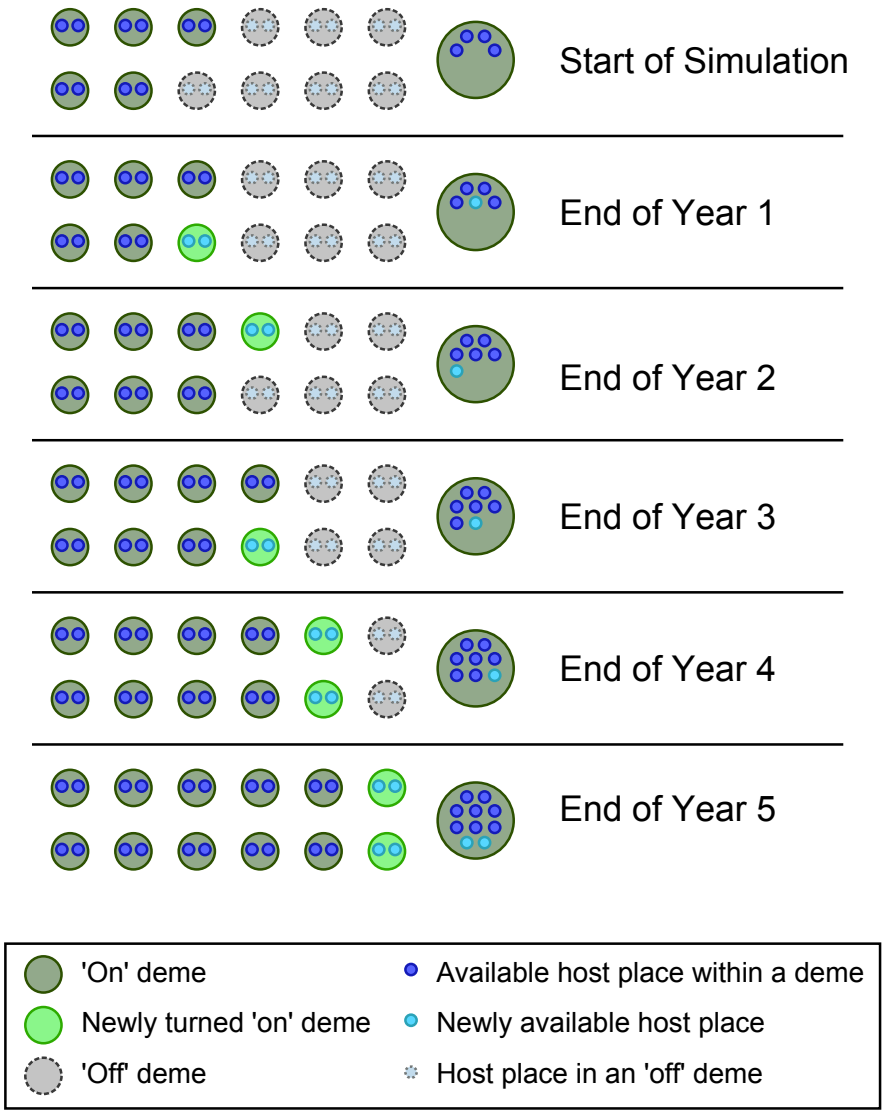

Figure 5.7: An example of how exponential growth is implemented in the DSPS. In this example the simulation starts with five two-person demes and one 'sex-worker' deme containing four hosts. The population is specified to grow exponentially at a rate of 20%. At the start of the simulation, a network is created containing all the demes that will be needed for a run of five years, but all but the starting five two-person demes are 'off' (grey). The sex-worker deme is at its starting size of four hosts. At the end of every year, the current number of active demes (dark green) multiplied by the growth rate informs the simulator of how many demes to turn 'on' (light green). The DSPS also adds new host spaces to the sex-worker deme at the end of each year, determined by on the current number of available host spaces multiplied by the growth rate (light blue hosts being added to the largest deme). By the end of the simulation, all demes are 'on' and the maximum number of hosts spaces has been made available.

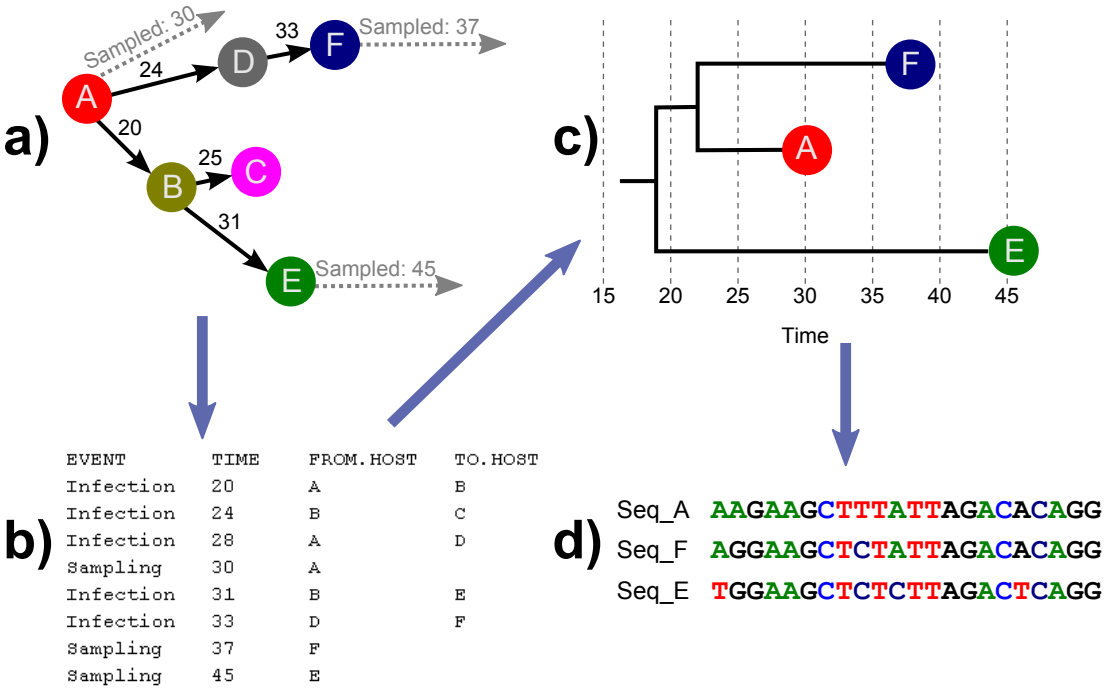

Figure 5.8: An illustration of how the DSPS can be used to generate realistic viral phylogenies and sequences. When a simulation is running, the virus is transmitted between individuals in the simulated population, and infected individuals are sampled. In a), the virus is transmitted between individuals, starting with A, with transmission time marked in black next to the arrow indicating direction. Sampling events are also marked in grey. All events during the simulation are stored in a 'line-list' (b) which records the type of event, time it took place, and individuals involved. The line-list is then used to generate a viral phylogeny with sampled individuals at the tips (c). The viral phylogeny takes into account within-host evolution, so two lineages may diverge before a transmission event, reflecting the standing diversity of viruses within a host. Finally, the viral phylogeny is used to generate viral sequences (d), starting with an ancestral sequence at the root, and using realistic mutation rates to produce sequences that correspond to the transmission relationships from the simulation.

"Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."

Jules Verne - 'Journey to the Center of the Earth' (1864)

6 Using the Discrete Spatial Phylo Simulator

In Chapter 5 I described how the Discrete Spatial Phylo Simulator (DSPS) works and what modification were made so that it can be used to model HIV dynamics. Here, I describe how the DSPS has been used to simulate two different kinds of epidemics.

First, through my work with PANGEA_HIV (described further in Section 6.1), the DSPS has simulated complex heterosexual HIV epidemics in a 'small African village' scenario. Second, to investigate the reliability of the heritability estimation pipeline described in this thesis (Chapter 2), the DSPS has been used to simulate varying degrees of heritability in a simplified approximation of a Western MSM (men who have sex with men) population.

6.1 The DSPS and PANGEA_HIV

PANGEA_HIV (Phylogenetics and Networks for Generalized HIV Epidemics in Africa; http://www.pangea-hiv.org) is a major initiative funded by the Bill and Melinda Gates Foundation with the goal of generating large volumes of next-generation HIV sequences from generalized epidemics in sub-Saharan Africa in order to use epidemiological and phylogenetic techniques to better characterise the epidemic and assess the impact of treatment and prevention efforts. 20,000 HIV-1 genomes will be generated from samples gathered from study sites and centres in Botswana, Zambia, Uganda, and South Africa and sequenced at the Sanger Institute and the Africa Centre.

With such an immense dataset and the constraints of a time-limited project, it is imperative that as sequences become available they are analysed using the most efficient and accurate methods. There are currently a great number of phylogenetic-based methods available, both published and unpublished, that have been developed to estimate disease and epidemic-related parameters such as R_0 (reproductive number, the number of infections one infection causes, on average), prevalence (the proportion of individuals in a population infected with a disease), and incidence (the number of susceptible individuals infected in a set time period). However, very few of these methods have been tested in datasets where the true values being estimated are known, allowing the estimates to be evaluated. There is also often little information on what kind of data a method might work best on. Some analyses might requires hundreds of sequences, some might only work in settings where certain information is available, and others might only work in a small population where sampling density is high. Finally, few methods have been tested on full-length HIV sequences like those that will become available through PANGEA_HIV. Thus, the extra computational resources required to perform analyses on full-length data are unknown, as is the amount of extra information gained from using more than just one region of the HIV genome.

Using simulated data, current phylogenetic methods can be tested on a dataset where all parameters are known in order to evaluate their performance in estimating disease parameters and detecting changes in epidemic dynamics. To provide such datasets for testing, and thus be able to recommend efficient and appropriate methods to be used on the PANGEA_HIV sequences, 'work package 4' (WP4) was set up to generate simulated HIV epidemics and assess the results of analyses performed by external groups using different phylogenetic methods.

Two independent simulators were adopted to develop sub-Saharan Africa-like datasets that could be released for evaluation: the modified DSPS and the HPTN 071 simulator developed for the PopART study (Cori et al., 2014). The HPTN 071 simulator would be used to generate HIV epidemics in a 'regional' setting, with a large population and low sampling coverage. The modified DSPS would be used to simulate a 'village' scenario, with a smaller population size and higher sampling density.

6.1.1 Generating HIV Epidemics in an African Village Population

To approximate an African village, the DSPS was configured so that only one male and one female host occupied each 'deme,' forming 'households' or long-term sexual relationships. All hosts were set to be heterosexual in orientation. As discussed in Chapter 5.3.3, the contact rate, or average number of sexual contacts per year, was set at 100. Initially, all demes were connected with the same parameters, with the probability of having a contact within the household being 50%, and the probability of having a contact with a host from any other household being 50%. Heritability was held at one to simplify the simulation and avoid complex dynamics due to viral load evolving, but the starting viral load was selected at random from the population distribution. For all HIV simulations, the viral load data from my subtype B dataset (Chapter 3 on page 67) is used as the population distribution of viral loads, simply because it is the largest. For the datasets generated for PANGEA_HIV, viral load is not provided to participants, and the population distribution of viral loads is essentially unused (see next paragraph). Birth and death parameters were set to 0.1 and 0.01, respectively, from experiments with early versions of the modified DSPS, and the 'removal' (progression to AIDS) parameter was set at 0.8 (this generates removal 'attempts' that are more likely to succeed as the time when the asymptomatic phase has been calculated to end gets closer). Preliminary runs confirmed that a removal parameter of 0.8 generated simulations where the average length of time infected hosts stayed in the asymptomatic phase was equal to the time predicted by the viral load in the population (Chapter 5.3.3). The transmission risk during the acute phase was the same as during the rest of the asymptomatic phase (the acute phase was 'turned off'). The growth of the population was set to 'stable' to avoid more than two individuals populating a deme (see Chapter 5.3.2) and to allow re-population of demes where all hosts have died.

It quickly became apparent that starting an HIV epidemic from one infected individual was not as predictable as might be assumed. As the viral load was randomly drawn from the population distribution, many simulations started with the initially infected individual having such a low viral load that they never infected anyone, or such a

high viral load that they progressed to AIDS and death very quickly, without infecting anyone. To reduce this occurrence, simulations proceeded with the initial viral load set to the population mean of 4.5 $\log_{10}$ copies/mL. As previously stated, heritability was set to one to reduce the complexity of the simulation by eliminating unpredictable epidemic patterns caused by variation in viral load over time. Thus, all individuals in the population had the same viral load of 4.5 $\log_{10}$ copies/mL.

Though this decreased the chance that the epidemic ended before it even got started with the death of the initially infected individual, epidemics failed to grow to infect any significant proportion of the population. Figure 6.1 shows the result of a simulation with 5000 households, each containing a male and female host. As mentioned before, the chance of having a contact within the household was 50%, and the chance of having a contact outside of the household was 50%. Though unrealistic, this seemed necessary to generate any kind of sustained infection. Though the epidemic often ran for many years, it rarely infected many people.

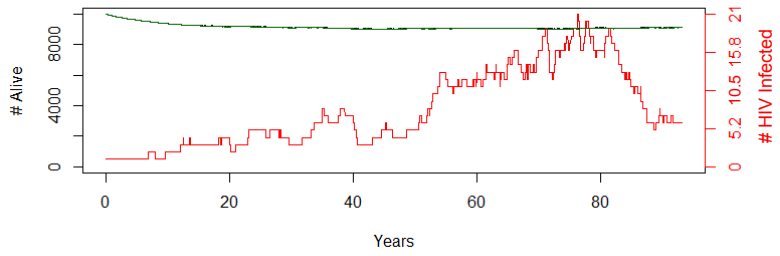

Figure 6.1: Illustration of the results of an early run with the modified DSPS. 5,000 two-person demes were created containing one male and one female each. A random individual was initially infected with a viral load of 4.5 $\log_{10}$ copies/mL. The graph shows the number of people alive in black and the number of people susceptible in green, both corresponding to the left-hand axis. In this simulation the numbers are so similar the lines overlap. The number of individuals infected with HIV is shown in red, and corresponds to the right-hand axis (also in red). The simulation ran for 100 years, but only a maximum of 31 people out of 10,000 were infected at any given time.

To try and make the simulator more realistic and more likely to generate an epidemic, the idea of having different 'risk groups' was introduced (this lead to the development of 'DemeGroups,' see Chapter 5.3.5). A 4,000 two-person deme simulation was run, with each deme containing one male and one female. 2,000 demes were assigned to be 'high risk,' and 2,000 assigned to be 'low risk.' In high risk demes, the risk of a

contact within the household was 50%, the risk of a contact to another high risk deme was 30%, and the risk of a contact to a low risk deme was 20%. In low risk demes, the risk of a contact within the household was 80%, the risk of a contact to a high risk deme was 15%, and the risk of a contact to another low-risk deme was 5%. As these parameters mean the population has fewer contacts outside their households than it did before (half of the population is now having 80% of their contacts within their households), the epidemic still did not spread effectively (Figure 6.2). However, this was an important step towards creating a more complex contact network.

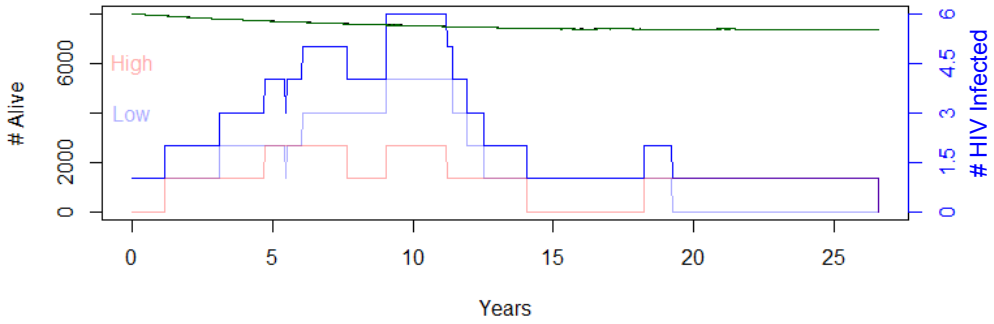

Figure 6.2: Illustration of the results of an early run with the modified DSPS. 4,000 two-person demes were created containing one male and one female each. A random individual was initially infected with a viral load of 4.5 $\log_{10}$ copies/mL. Half the demes were 'high risk,' with 50% of their contacts being outside of their household, and half the demes were 'low risk,' with 20% of their contacts being outside of their household. The graph shows the number of people alive in black and the number of people susceptible in green, both corresponding to the left-hand axis. In this simulation the numbers are so similar the lines overlap. The number of individuals infected with HIV is shown in dark blue, and corresponds to the right-hand axis (also in dark blue). The number of individuals infected in the high and low risk groups is shown in pink and light blue, respectively, and corresponds to the right-hand axis.

A new simulation was parametrised, with 6,000 two-person demes divided equally into high, medium, and low risk groups. High risk demes had 50% of their contacts within their household, 28% with other high risk demes, 17% with medium risk demes, and 5% with low risk demes. Medium risk demes had 70% of their contacts within their household, 17% with high risk demes, 10% with other medium risk demes, and 3% with low risk demes. Low risk demes has 90% of their contacts within their household, 5% with high risk demes, 3% with medium risk demes, and 2% with other low risk demes. Figure 6.3 on the following page shows the results of one of these runs. While the

number of infected individuals at the end of the simulation has gone up, still only a tiny proportion of the population is infected.

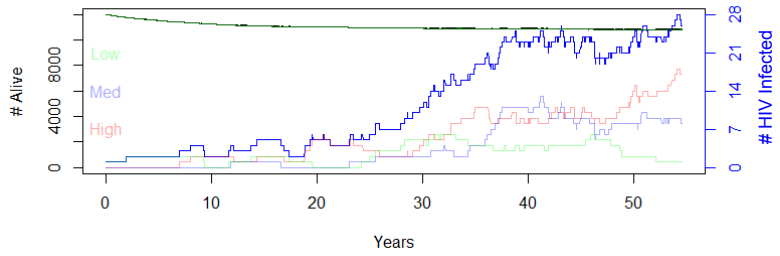

Figure 6.3: The results of an early run with the modified DSPS. 6,000 two-person demes were created containing one male and one female each. A random individual was initially infected with a viral load of 4.5 $\log_{10}$ copies/mL. 2,000 of the demes were 'high risk,' with 50% of their contacts being outside of their household, 2,000 of the demes were 'medium risk,' with 30% of their contacts being outside of their household, and 2,000 of the demes were 'low risk,' with 10% of contacts being outside of their household. The graph shows the number of people alive in black and the number of people susceptible in green, both corresponding to the left-hand axis. In this simulation the numbers are so similar the lines overlap. The number of individuals infected with HIV is shown in dark blue, and corresponds to the right-hand axis (also in dark blue). The number of individuals infected in the high, medium, and low risk groups is shown in pink, light blue, and light green, respectively, and corresponds to the right-hand axis.

Realising that a higher extra-household contact rate was needed to spread the infection, I decided to add a deme that would function as a group of female sex workers. Sex workers are common in many African villages and towns, and could be an effective and realistic route to increasing the spread of HIV, particularly as sex workers could be expected to have a much higher number of sexual contacts, and one sex worker could have contacts with individuals from many different households across different risk groups. A 4,001 deme run was created, with 4,000 of the demes being two-person households, as before, and one deme being a 'sex worker' deme. 2,000 of the two-person demes were high risk, with 50% of their contacts inside their household, 20% to other high risk demes, 10% to low risk demes, and 20% to the sex worker deme. The other 2,000 two-person demes were low risk, with 80% of their contacts being within their household, 8% being to high risk demes, 5% being to other low-risk demes, and 7% being to the sex worker deme. The sex worker deme contained 200 female hosts, who each had a contact rate of 600 contacts/year on average. As the sex workers are all female, they cannot infect each other, and can only have contacts with male hosts (half

the non-sex worker population). Sex workers have contacts with high risk demes 80% of the time, and low risk demes 20% of the time.

To aid in the initial spread of the epidemic, the DSPS was modified to ensure that the initial infection always occurred in the sex worker deme. This greatly increases the chances that an epidemic will reach exponential growth, and spread successfully through the population. Figure 6.4 on the next page shows some of the runs from this configuration. Though not all runs with different random number seeds were guaranteed to be successful, the new configuration made it much more likely that a run would experience exponential growth and that about 10-20% of the population would be HIV-infected by then end of the simulation.

This configuration, with high and low risk two-person household demes and a sex worker deme, became the basis for all future runs of the PANGEA_HIV simulations. The introduction of a sex worker deme allowed HIV epidemics to be generated much more easily, with slightly more realistic settings. It is important to remember that the simulation is stochastic, so all epidemics do depend on an element on chance. Thus, there are always some epidemics that don't 'take off' and there is plenty of variation between epidemics that do – some grow exponentially right away, whereas others have a long period of slow growth before growing exponentially.

6.1.2 Incorporating Treatment and 'Migrant' Sequences

The aim of the PANGEA_HIV data sets was to have samples from increasing, decreasing, and stable epidemics. 'Migrant' sequences were also to be incorporated, to represent HIV transmissions coming in from outside the focal population, which could potentially disrupt phylogenetic analysis by introducing divergent sequences.

Treatment

From the parameters already incorporated, sampling during increasing and stable epidemics was easily done. In order to sample during a decreasing epidemic, treatment was introduced (see Chapter 5.3.6). Treatment was started at year 35, just before the epidemic peaked in most of the simulations selected for the PANGEA_HIV release. Treatment was rolled out to 20% of the two-person household demes, randomly se-

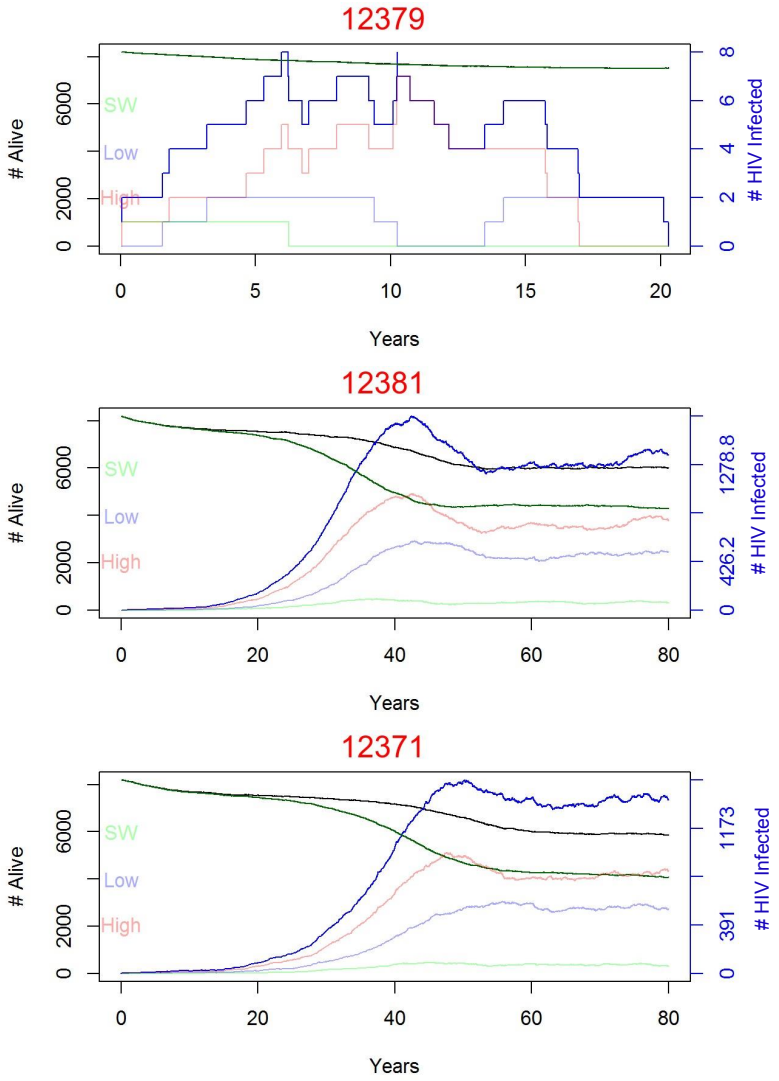

Figure 6.4: The results of what became the base configuration for the PANGEA_HIV runs. 4,000 two-person demes were created: 2,000 of the demes were high risk, with 50% of their contacts being outside of their household, and 2,000 of the demes were low risk,' with 20% of their contacts being outside of their household. One deme contained 200 females with a much higher contact rate and functioned as a 'sex-worker' deme that has contacts with high risk demes 80% of the time and low risk demes 20% of the time. The graph shows the number of people alive in black and the number of people susceptible (not infected) in green, both corresponding to the left-hand axis. The number of individuals infected with HIV is shown in dark blue, and corresponds to the right-hand axis (also in dark blue). The number of individuals infected in the high, low, and sex worker (SW) risk groups is shown in pink, light blue, and light green, respectively, and corresponds to the right-hand axis.

lected, and to the sex worker deme. The treatment parameter was set at 0.5, which corresponds to an intermediate speed of treatment roll-out. Starting treatment, as might be expected, causes the epidemic to decline, as those on treatment no longer

transmit and live longer than those not on treatment, essentially becoming 'blocks' to transmission networks that would facilitate the spread of HIV. Figure 6.5 shows how sampling during the increasing the decreasing epidemic was performed in a simulation where treatment started at year 35.

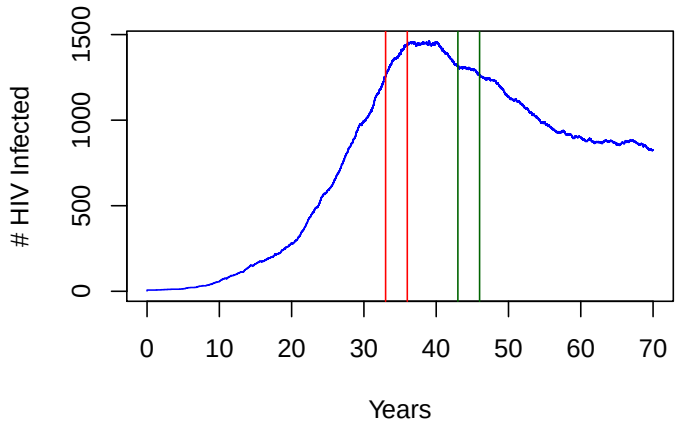

Figure 6.5: In the simulated epidemic shown here, treatment has been introduced to a random 20% of the population, plus the sex-worker deme, at year 35. The number of HIV-infected individuals is shown on the Y-axis and plotted in dark blue. The red lines indicate a 3-year time period during which sampling might be carried out to reflect a 'growing' epidemic. The green lines indicate a 3-year time period during which sampling might be done to reflect a 'declining' epidemic.

Migrant Sequences

To incorporate 'migrant' sequences, six new demes were added to the simulations. Each of these six demes were populated with one immortal male host (the birth and death parameters were set to 0) who represented a 'village' outside of the focal population. (A representation of the final deme configuration, including 'village' demes, is shown in Figure 6.6 on the next page.) The average number of contacts with the focal population was set at 12, with 30% of these being with the high risk group demes, 20% being with the low risk group demes, and 50% being with the sex worker deme. The DSPS was modified so that immediately after the initial infection of a sex worker, six infection events were generated which infect the six men in the village demes, before the simulation proceeds as normal. This allows the HIV infection to diverge from the main population within these six men. While they have a very low contact rate with

the rest of the population, they do occasionally transmit their lineage of HIV back into the focal population, where it may spread. This has the effect of generating a final phylogeny and sequence data set where some of the sequences appear more divergent, similar to what might be seen if HIV is being transmitted from outside of the focal population in a real cohort. The new village demes do little to change the overall dynamics of the epidemic, but provide a more realistic simulation of the sequences that might be obtained from cohorts (Figure 6.7 on the facing page).

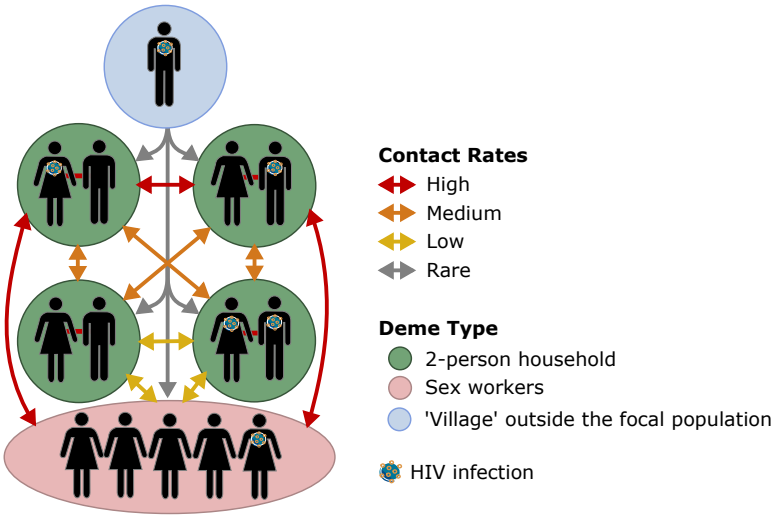

Figure 6.6: Illustration showing the final deme configuration used for the simulation of the PANGEA_HIV datasets. Demes are represented by circles containing hosts. Two-person households are shown in green, the sex worker deme is in red, and the 'village' deme is in blue. The contact rates are shown as arrows between the demes, and vary across the population.

For the PANGEA_HIV release, three different epidemics with similar dynamics were simulated, each with a growing and stable phase without treatment. When treatment was added to these simulations at year 35, all three epidemics experienced a declining phase. The simulations could be run in about 45 minutes, and demonstrated the flexibility of the DSPS in recreating realistic epidemic dynamics.

6.1.3 Creating Viral Phylogenies and Simulated Sequences

To create the sequence data sets for the first PANGEA_HIV release, an in-house R script was used, as described in Chapter 5.4, to sample three-year time periods during increasing, decreasing, and stable epidemic growth, and create 'line lists' of the infection

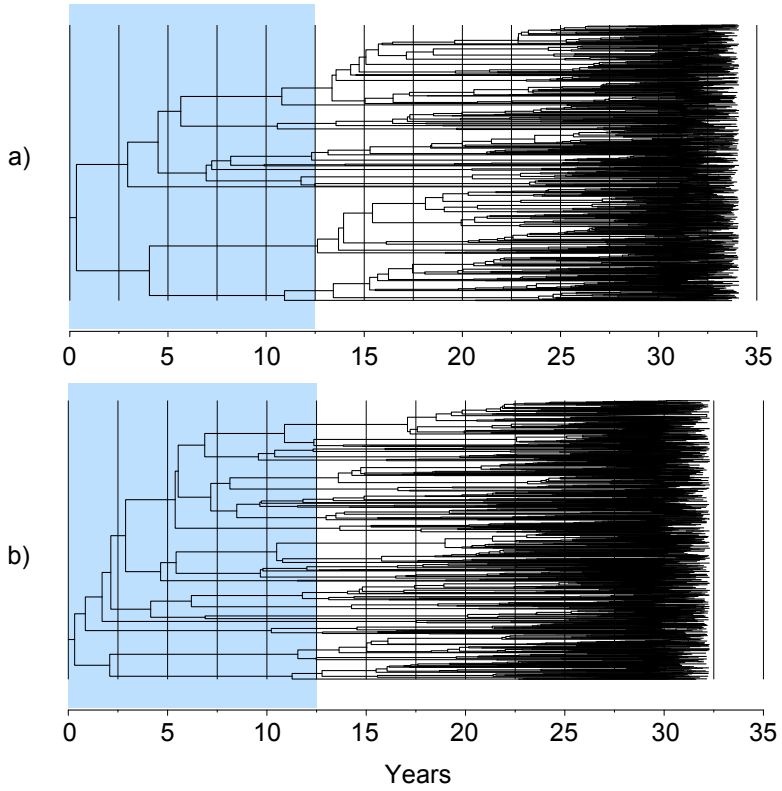

Figure 6.7: The phylogenies produced from sampling an epidemic of 4,000 two-person 'high' and 'low' risk demes and one 200-person 'sex-worker' deme for three years at roughly 30 years into the epidemic. In (a) no HIV transmission from the six 'village' demes was present; in (b) transmission from the six 'village' demes was allowed. Though the overall dynamics of the epidemic are not affected by the addition of 'village' demes, particularly after the epidemic has begun growing exponentially, the phylogenetic structure of the resulting HIV sequences is affected by divergent HIV strains being transmitted back into the population. In (a) there are only 13 branching events in the phylogeny in the first 12.5 years of the epidemic (blue area). In (b) there are 35 branching events in the phylogeny in the first 12.5 years. This is due to the transmission of divergent strains back into the population. In the (b) simulation, five transmissions from the 'village' demes occurred before the sampling time, with four of them occurring before 12.5 years.

and sampling events.

Using the VirusTreeSimulator developed by Matthew Hall (see Chapter 5.4), viral phylogenies were generated. For the PANGEA_HIV DSPS simulations, parameters were chosen with advice from MH and Oliver Ratmann, who was working on the HPTN 071 simulator. N_0 was calculated as the N_e at infection time (assumed to be 1) $\times$ the generation rate of HIV (52 hours, 0.00593 years) giving a value of 0.00593. r was calculated by OR to have an optimal value of 2.852 assuming a final N_e of 300. T_{50} was set at -2 on the recommendation of OR, to represent that the N_e was half of its

final size two years after transmission.

Using the generated viral phylogenies, piBUSS was used to simulate sequences, as described in Chapter 5.4. Gonzalo Yebra prepared ancestral *gag*, *pol*, and *env* sequences to use as starting sequences for the piBUSS simulation. GY selected 100 full genome subtype C sequences from different southern African countries (South Africa, Botswana, Malawi, and Zambia) from Los Alamos HIV Database (www.hiv.lanl.gov), with dates ranging from 1989 to 2011. The sequences were down-sampled from around 400 sequences to include only sequences equally distributed throughout the time period. The full-length sequences were separated into the constituent genes to analyse *gag*, *pol*, and *env* separately.

Each gene was run in BEAST using the Skyride demographic model with the GTR+gamma nucleotide substitution model for 200 million steps. Rather than estimating ancestral sequences for every ancestral node in the tree (ancestral state reconstruction), just the root ancestral sequence was sought, to reduce computation time. The SRD06 nucleotide substitution model was not used as it uses two partitions, one for the 1st and 2nd codon positions, and one for the 3rd codon position, which results in an ancestral sequence split across two alignments: one sequence for the 1st and 2nd codon position, and another sequence for the 3rd position. From the resulting ancestral sequences generated from the runs on *gag*, *pol*, and *env*, a consensus sequence was generated from the last 100 states.

For each sequence simulation in piBUSS, the same viral phylogeny was used to generate *gag*, *pol*, and *env*. Different mutational parameters were used for each gene, and for the 1st and 2nd codon positions and the 3rd codon position. These were provided by GY from his analyses on African subtype C sequences. For *gag* and *pol*, codon positions 1 and 2 had a molecular clock rate of 1.489×10^{-3}, while codon position 3 had a molecular clock rate of 2.764×10^{-3}, with both rates having an SD of 4.745×10^{-7}. For *env*, codon positions 1 and 2 had a molecular clock rate of 2.978×10^{-3} and codon position 3 had a molecular clock rate of 5.528×10^{-3}, with both rates having an SD of 9.490×10^{-7}.

The resulting gene sequences for each epidemic sample period were combined into one long sequence 6987bp in length, containing all three genes, and released for analysis.

6.1.4 Exponential Growth, Acute Phase, and Treatment Roll-Out

Population Growth

One of the biggest limitations of the DSPS simulations up to this point was the unrealistic population growth. As described in Chapter 5.3.2, the DSPS initially only allowed 'stable' and 'growth' population growth. While both may be acceptable in some settings, the need to maintain only two individuals in the household demes meant that the HIV DSPS simulations had to use the 'stable' growth setting (see Figure 5.4 for an illustration of the two types of growth).

The stable setting is primarily unrealistic because most populations, particularly in developing countries, are not stable. However, the stable population size also caused an unexpected problem. As can be see in Figure 6.4 on page 170, where the black line shows the number of individuals alive, the population actually declines, rather than remaining stable. Part of this decline is due to the HIV epidemic itself, and this is less problematic, as at the height of an HIV epidemic is not unrealistic to expect that populations may remain the same size rather than grow. However, even before the epidemic begins growing exponentially, it is apparent that the population is declining. Because the risk of a random death event is tied to the number of individuals in a deme, if all demes are fully populated, as they are at the start of a simulation, the risk of deaths events being generated is fairly high. This leads to a string of random deaths at the start of the simulation, until the population size drops so that death events become less frequent, creating a kind of 'equilibrium' with death (in Figure 6.4 this happens around at 10-15 years after the start of the simulation, just before the impact of the exponential growth of the epidemic can be seen).

While it's unlikely this rash of deaths at the beginning of the simulation would affect any analysis concentrating on the HIV epidemic, which grows exponentially only after the 'death-equilibrium' is reached, it is unrealistic and violates assumptions about population growth that are commonly made when analysing HIV data.

Exponential Growth

For the second set of simulations released by PANGEA_HIV, I wanted to allow for more realistic population dynamics. Thus, I implemented a way for simulated populations to grow exponentially. As discussed in Chapter 5.3.7, exponential growth is implemented by creating as many demes as will be needed at the height of the simulation, but only starting the simulation with some of them 'on.' The simulation then 'turns on' demes at the end of each year at the exponential growth rate. This implementation saves the computational overhead of the DSPS simulator having to create and then connect demes during the simulation.

One of the concerns with exponential growth is the increase in processing time that comes from having a population that will grow to be very large. As described in Chapter 5.2.3, events are generated by surveying all active demes in order to gather their 'Total Hazard' for having an event. Thus, as the number of demes increases, there are more demes to survey at every step, and the time to run the simulation increases dramatically. To prevent the runs from taking too long to complete, the initial population size was scaled back slightly. The new simulations started with 3,000 two-person household demes (1,500 high risk and 1,500 low risk) and 1 sex worker deme of 200 females. If the simulations included migration, the same six 'village' demes were also included. All of the infection, recovery, death, and migration parameters remained the same, but the birth rate was increased from 0.1 to 0.6, as this was found to approximate exponential growth during the simulation run. This is because though the simulator is easily set to 'turn on' or expand demes at an exponential rate, these empty demes must be populated by birth events for the population of hosts to actually grow. Thus, it takes a slightly higher birth rate to achieve satisfactory growth.

The simulations were run with 1% growth per year. An in-house script was used to calculate the total number of demes needed based on a starting population of 3,000 two-person demes and one 200-person sex worker deme running for 70 years with 1% growth, and to generate an XML file with these parameters. An XML file with 6,018 two-person demes (3,009 high risk, 3,009 low risk) and 1 sex worker deme was produced. 3,018 of these demes were randomly chosen to be 'turned off' at the start of the simulation, and

would be 'turned on' at the appropriate rate as the simulation progressed. With the sex worker deme also growing exponentially from 200 to 395 females over the course of the simulation, the maximum population size after 70 years would be 12,431 hosts in 6019 demes. (If migration is included, the six other hosts in the 'village' demes increase this total slightly, but are unaffected by the exponential growth.) Though the simulation is set up to hold this many hosts, it is unlikely that the maximum will be reached due to random death events and the mortality from the HIV epidemic.

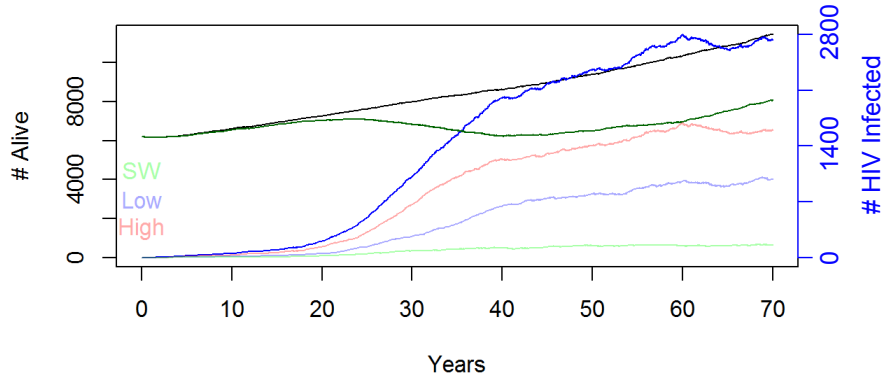

Figure 6.8: Epidemic dynamics in a DSPS simulation with exponential growth turned on. The number of hosts alive and the number of susceptible hosts are plotted in black and green, respectively, and correspond to the left-hand axis. The number of HIV-infected hosts is plotted in dark blue, and corresponds with the right-hand axis. The number of infected individuals from high risk, low risk, and sex worker demes is shown in pink, light blue, and light green, respectively, and also corresponds to the right-hand axis. The population starts with 3,000 two-person household demes (half high risk, half low risk) and one 200 women sex worker deme, and grows at a rate of 1% per year. Unlike in previous simulations where the population declined or held steady, the population here continually increases. After the exponential growth phase, the HIV epidemic stabilizes to the growth of the population (from just after year 40). See also Figure 6.9

An example of simulation run under these parameters, without treatment, can be see in Figure 6.8 and Figure 6.9 on the next page. As expected, the overall population size increases through the entire simulation. Growth does slow slightly during the period when the HIV epidemic is growing exponentially, but as the HIV epidemic stabilizes, so does population growth. Though the population size (in black) is on a different scale (axis on left-hand side) than the number of HIV infected hosts (in blue, axis on right-hand side), it is apparent that after the period of exponential growth, the HIV

epidemic stabilizes to match the population growth (the prevalence stays constant at around 25%, as shown in Figure 6.9). With the new exponential growth rate, the run time for a simulation increased from around 45 minutes to 2-3 hours.

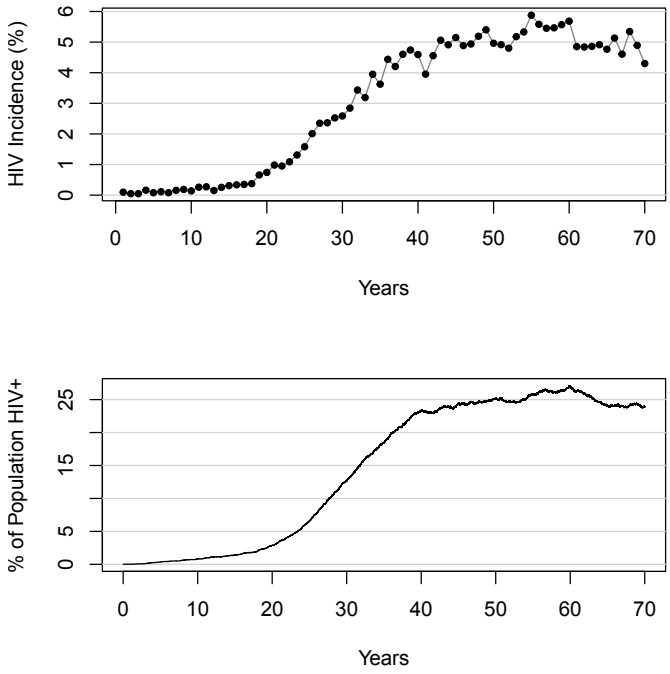

Figure 6.9: Epidemic dynamics in a DSPS simulation with exponential growth turned on. Incidence, as the number of new infections per year divided by the number of susceptible individuals at the beginning of that year, is shown at the top. Prevalence, the proportion of the population infected with HIV, is shown at the bottom. Though the absolute number of HIV infected individuals continues to grow throughout the simulation (shown in Figure 6.8), the incidence and prevalence stabilise at around year 40, as the exponential growth of the epidemic slows to match the linear population growth.

Acute Phase

As well as the previous goals of asking participants to detect and estimate changes in epidemic parameters, for PANGEA_HIV's second release of simulated datasets additional goals were developed: estimate the proportion of transmissions that occur during the acute stage of infection and detect differences in the speed of treatment roll-out.

To vary the amount of transmission that occurred during the acute stage of infection, the simulations were run once with the acute stage inactivated, as was done with the runs in the previous release, and once with the acute phase on. As described in Chapter 5.3.3, the acute stage in the DSPS lasts for 0.24 years, during which time the transmis-

sion risk per act is 0.0276 (as calculated by Fraser et al. (2007) and Hollingsworth et al. (2008)). In initial runs of turning on the acute phase, it was quickly discovered that the per-act risk of 0.0276 was producing unrealistic results. As shown in Figure 6.10 (a), the epidemic grows far too quickly – 60% of the population is HIV-infected within 20 years of the start of the simulation.

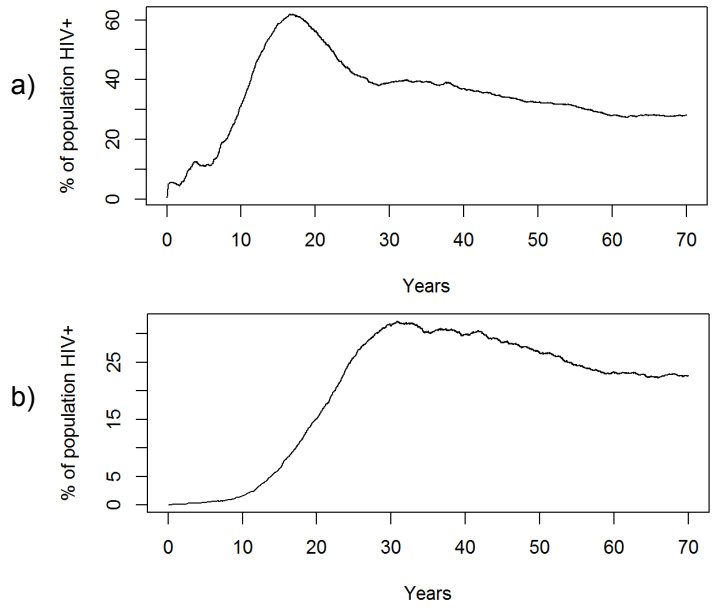

Figure 6.10: Plots showing the percentage of the simulated population infected with HIV over time. In (a), the acute phase has been turned on with the default per-act risk of 0.0276 taken from Fraser et al. (2007); in (b) the acute phase per-act risk has been divided by 2, to give a per-act risk of 0.0138. In both simulations, treatment starts at year 40. In (a), the HIV epidemic spreads far too quickly, infecting 60% of the population in the first 20 years. In (b), the epidemic spreads at a much more realistic rate, peaking at around 27% of the population infected at year 30, and levelling off at just under 25%.

Finding out that the HIV epidemic spread so effectively when the acute phase was implemented with a per-act risk of 0.0276 was promising for future contact network parametrization. The current network parameters dictate that high risk households, which compose roughly 50% of the population, have sexual contact outside of their primary relationship 50% of the time. This is almost certainly unrealistic, but was necessary in order to create an HIV epidemic that spreads through the population. As turning on the acute phase with a per-event risk of 0.0276 makes the epidemic spread significantly faster, this potentially provides a way to adjust the contact network so that more realistic parameters can be used. Unfortunately, this would require a complete

re-parametrization of the simulated population, and due to PANGEA_HIV deadlines, it was not possible to do this in the time available. Instead, the transmission risk during the acute phase was divided by two, giving a per-act transmission risk of 0.0138, which is still within the 95% confidence interval of the acute-stage transmission risk calculated by Hollingsworth et al. (2008). This allowed a noticeable change in the epidemic dynamics, without becoming completely unrealistic (compare Figure 6.10 on the preceding page (a) (per-act risk of 0.0276) and (b) (per-act risk of 0.0138)).

When the acute phase is 'turned off,' transmissions are spread evenly through the acute and asymptomatic phase, and about 5% of transmissions occur during the acute phase (first 0.24 years). When the acute phase is 'turned on' with a per-act risk of 0.0138, about 20% of transmissions occur during the acute phase.

Treatment Roll-Out

Treatment had already been successfully implemented in the first simulated data release, but for the second release the speed of treatment roll-out was to be varied. I implemented two different speeds at which treatment is made available, with 'slow ART' having a treatment parameter of 0.25, and 'fast ART' having a treatment parameter of 1.0. In both cases, treatment was started at year 40 and a random 20% of the two-person household population, plus the sex worker deme, was treated.

The results of having the acute phase 'on' and 'off' in combination with having 'slow' and 'fast' ART can be seen in Figure 6.11 on the next page and Figure 6.12 on page 182. Both figures are run in populations with exponential growth and a starting population of 3,000 two-person household demes (1,500 high risk, 1,500 low risk), one 200 person sex worker deme, and six 'village' demes to enable migration. In both, treatment was started in a randomly-selected 20% of the two-person household demes plus the sex worker deme at year 40. In Figure 6.11, where the acute phase is not implemented, the epidemic grows much more slowly than in Figure 6.12, where the acute phase is implemented with a per-act risk of 0.0138. The epidemic in Figure 6.12 grows quickly enough that it peaks and stabilizes before treatment begins, unlike in Figure 6.11, where the epidemic is still growing when treatment is started at year 40.

In (a) in both figures, the treatment that starts at year 40 has a 'slow' roll-out,

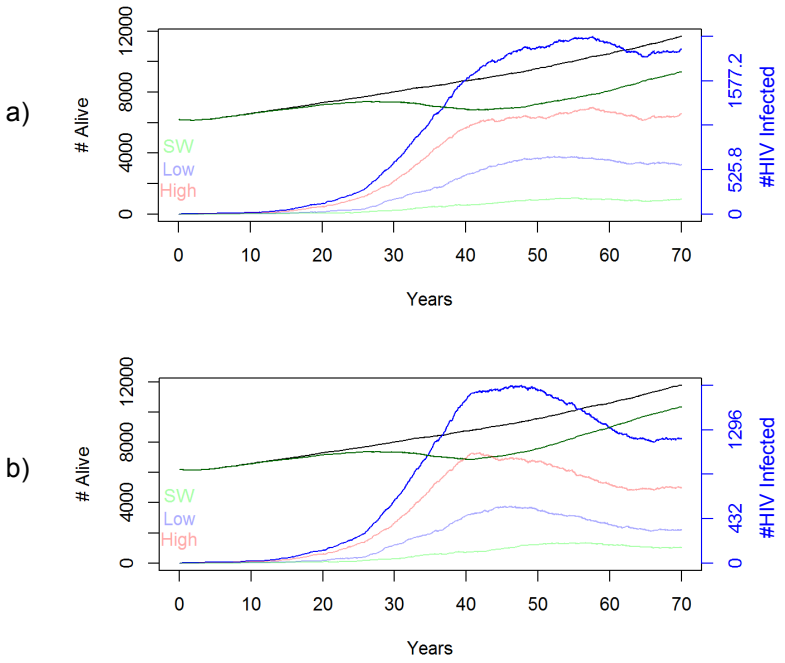

Figure 6.11: A DSPS epidemic simulated with no increased acute phase transmission (acute phase 'off') and 'fast' (a) and 'slow' (b) ART roll-out speed. The run was simulated with exponential growth, and a starting population size of 3,000 two-person household demes (1,500 high risk, 1,500 low risk) and one 200 person sex worker deme. Treatment was started in a random 20% of the population plus the sex worker deme at year 40. In these runs, the acute phase was not implemented, meaning the HIV epidemic spread more slowly than in Figure 6.12. In (a) 'slow' ART was implemented with a treatment parameter of 0.25, and in (b) 'fast' ART was implemented with a treatment parameter of 1.0. The 'slow' ART in (a) does not prevent the HIV epidemic from continuing to spread until about ten years after treatment is started. The 'fast' ART in (b) causes an immediate stabilization of the HIV epidemic when treatment is started at year 40, which leads to a decline in the number of HIV infected hosts from around year 50.

with a treatment parameter of 0.25. In both Figure 6.11 and Figure 6.12 this treatment seems to have a mild effect on the epidemic dynamics, but does little to slow epidemic growth (in Figure 6.11) or reduce the number of HIV infected individuals (in Figure 6.12). In contrast, the 'fast' ART roll-out in (b) in both figures has an immediate effect, stabilizing the HIV epidemic and then reducing the number of HIV infected hosts in Figure 6.11, and causing an immediate reduction in HIV infected hosts in Figure 6.12.

The second release of the PANGEA_HIV simulated datasets consisted of twelve different epidemics with differing combinations of acute phase, treatment roll-out speed, migration, and sampling fraction, and one epidemic where there was no treatment. As

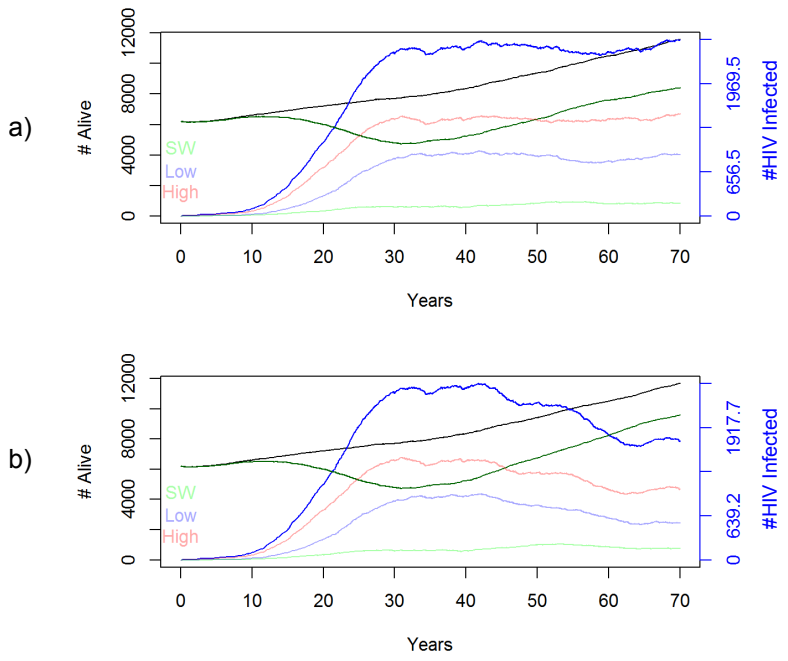

Figure 6.12: A DSPS epidemic simulated with increased acute phase transmission (acute phase 'on') and 'fast' (a) and 'slow' (b) ART roll-out speed.The run was simulated using the DSPS with exponential growth, and a starting population size of 3,000 two-person household demes (1,500 high risk, 1,500 low risk) and one 200 person sex worker deme. Treatment was started in a random 20% of the population plus the sex worker deme at year 40. In these runs, the acute phase was implemented with a per-act risk of 0.0138, meaning the HIV epidemic spread more quickly than in Figure 6.11. In (a) 'slow' ART was implemented with a treatment parameter of 0.25, and in (b) 'fast' ART was implemented with a treatment parameter of 1.0. Because of the faster spread of the epidemic due to the acute phase, the epidemic in these simulations peaks before treatment is started, unlike in Figure 6.11. The 'slow' ART in (a) seems to have little effect on the epidemic, though may be holding epidemic growth stable before it begins to increase at the same rate as the population growth in year 60. The 'fast' ART in (b) causes an immediate reduction in the number of HIV infected individuals until at least year 60, when the epidemic seems to stabilize.

detailed in Section 6.1.3, viral phylogenies were generated for all thirteen simulated epidemics. Feedback from the first data release indicated that much of the analysis time was spent attempting to get a reliable phylogeny from the sequences, which can be a very lengthy process. This would need to be done with the real sequence data, but once a phylogeny has been created it could theoretically be shared among different analysis teams. In recognition of this, and to allow participants to make optimal use of the time available for the second data release analysis, nine of the epidemics were released as time-resolved true viral transmission trees, and four as simulated sequence

datasets. This allows participants the ability to concentrate on what information their method can extract from the phylogeny under the assumption that they have obtained a reliable phylogeny from elsewhere, perhaps one run by another group with more experience and more computational resources to devote to obtaining a tree.

6.1.5 Discussion on the PANGEA Simulations

In preparing simulated epidemics under a variety of different scenarios to use as evaluation tools for phylogenetic methods, the modified DSPS was significantly improved. HIV epidemics can now be reliably started and maintained under realistic simulated conditions, including an exponentially growing population. The effect of introducing an acute phase was explored, and the flexibility and impact of different 'speeds' of treatment roll-out were demonstrated.

Though implementing exponential growth increased the simulation run-time from 45 minutes to 2-3 hours, most of this is due to the overhead associated with surveying every deme to calculate its 'Total Hazard' (see Chapter 5.2.3), a process that takes an increasing amount of time as the number of demes grows. However, random number and probability algorithms already exist that should be able to streamline this process in future, hopefully reducing the run-time to under an hour, allowing ever-larger populations to be simulated.

In the contact networks simulated here, the acute phase implemented with a per-act risk of 0.0276, as calculated by Fraser et al. (2007), caused the epidemic to spread too quickly. However, as the current contact network includes an unrealistic number of extra-household contacts (Section 6.1.4), future implementations can hopefully rely on more realistic parametrization of extra-household contacts by implementing the acute stage with a per-act risk of 0.0276 instead.

6.2 Heritability Estimates with the DSPS

The realistic simulations produced by the DSPS allow a unique opportunity to evaluate the heritability estimation pipeline described in this thesis (see Chapter 2). Allowing variability in viral load values makes the spread of an HIV epidemic considerably less

predictable, and so heritability was set at one (100%) for all the PANGEA_HIV related runs. However, setting the heritability to other values and tracking the resulting viral loads allows a simulated dataset to be generated from which heritability can be estimated using the heritability estimation pipeline.

6.2.1 Introduction

As described in Chapter 5.3.4, Alizon et al. (2010) used simulation to verify their own phylogenetic method by using an SIR model to generate twenty random phylogenies over thirteen generations with a probability of death of 0.25, and a probability of transmission of 0.75 (modelled as a branching in the tree). Viral load was then simulated down the tree at heritability values of 0.3, 0.5, 0.7, and 0.9. When transmission occurred (a branching event), one 'child' branch retained the viral load of the 'parent,' while the other 'child' branch obtained a new viral load calculated by the equation:

$$x_{a+1} = \zeta x_a + (1 - \zeta)y \tag{6.1}$$

where x_{a+1} represents the viral load of the new 'child,' x_a represents the viral load of the 'parent,' y represents a random value drawn from the empirical distribution of viral loads in the population, and ζ represents the heritability of viral load. Each of the simulated phylogenies was then sampled at random down to 128 tips, before the heritability was estimated using Pagel (1999)'s λ and Blomberg et al. (2003)'s K.

Alizon et al. (2010)'s simulation has limitations. Under a true Brownian motion model, which is assumed both in their own phylogenetic analysis and in the method presented here, both 'child' nodes would assume new viral loads, unlike what happens in Alizon et al. (2010)'s simulations, where ancestral nodes are quite likely to end up becoming tips, unless they happen to die. Alizon et al. (2010)'s equation also fails to control for the loss of variance that occurs at each step of his simulation. Starting with a 'population' of 20 viral loads selected at random from the subtype B dataset (see Chapter 3), I simulated ten 'generations' under a heritability of 50% where 20 random viral loads from the 'population' each had one 'child' generated using Alizon et al. (2010)'s equation, which were then added to the population. The variance of the initial

20 viral loads was 0.69, and after only 10 generations (and no death) the variance had decreased to 0.38. Running 20 replicates of these 10 generations caused the variance to decrease on average by 0.40. Running 20 replicates where there were 50 generations caused the variance to decrease on average by 0.58.

By using a variation of Alizon et al. (2010)'s equation (described previously in Chapter 5.3.4), the DSPS is able to avoid the problem of losing variance at each step. The new equation takes the square root of the heritability and one minus the heritability to maintain variance, and then uses viral loads standardized by the population mean to keep the simulated viral loads at reasonable values. The resulting equation is:

$$x_{a+1} = z + [\sqrt{\zeta}(x_a - z) + \sqrt{(1-\zeta)}(y - z)] \tag{6.2}$$

Where all parameters are as previously described in Equation 6.1, and z represents the mean of the population distribution of viral loads read in at the beginning of the simulation (Chapter 5.3.3). Repeating the same simulation as described above for 20 replicates of 10 generations, but using the modified equation, showed an average decrease in variance of 0.040. 20 replicates where there are 50 generations showed the variance to decrease on average by 0.023. Though this simulation is very basic, it demonstrates how the modified equation maintains variance in the simulated population of viral loads.

6.2.2 Methods

In order to avoid an over-complicated simulation that could unpredictably affect the ability to estimate the heritability of viral load, and in order to create a dataset similar to the UK HIV DRB subtype B dataset, I decided to simulate heritability using the DSPS in a fully-connected MSM population. The viral load data from my subtype B dataset (Chapter 3 on page 67) was used as the population distribution of viral loads to approximate the UK HIV subtype B epidemic.

The starting population consisted of three demes: a 'low risk' deme, a 'medium risk' deme, and a 'high risk' deme. This roughly corresponds with the patterns found in a previous analysis of subtype B sequences from MSM, where individual sequences were

linked to other closely-related sequences to create 'clusters' that represent transmissions five or fewer years before sampling (Leigh Brown et al., 2011). 29% of sequences linked to only one other sequence, 41% linked to between 2-10 sequences, and 29% linked to more than 10 sequences (Leigh Brown et al., 2011). In the DSPS simulation, low risk hosts had 20 sexual contacts a year on average, of which 40% were to other low-risk individuals, 40% to medium risk hosts, and 20% were to high-risk hosts. Medium risk hosts had 40 contacts a year on average, of which 35% were with other medium risk hosts, 35% were to low-risk hosts, and 30% were to high-risk hosts. High risk hosts had an average of 200 contacts per year, of which 30% were to other high risk hosts, 35% to medium-risk hosts, and 35% to low-risk hosts. Population growth was exponential at a rate of 1% per year, with all three demes expanding accordingly. Birth and death parameters were set at 0.6 and 0.01, as previously, to approximate exponential growth.

The initial infection was to one individual from the high risk deme, who was infected with a viral load of 4.5 $\log_{10}$ copies/mL. The acute phase was implemented, with a per-act risk of 0.0276. No treatment was included in the simulation. Heritability was specified in the XML file at four levels: 30%, 50%, 70%, and 90%.

Initially, simulations were run with a starting population of 1,000 low-risk, 1,000 medium-risk, and 200 high-risk individuals for 70 years. All hosts that had been infected with HIV for more than 3 months between years 40-50 were sampled (100% sampling), and this sampled line list was used to generate a viral phylogeny using the VirusTreeSimulator as detailed in Section 6.1.3. The simulated set-point viral loads and the viral phylogeny were put through the heritability estimation pipeline in the same was as the UK HIV DRB data was processed, as described in Chapter 2.6. Five simulation replicates were run for each heritability level, and the mean heritability estimate and 95% confidence intervals were calculated for each heritability level simulated. As was done previously (see Chapter 2.6), the heritability estimate was standardized by the average root-to-tip distance so that the estimate was directly interpretable as the variance explained over the length of the phylogeny created from the samples.

Simulating an HIV epidemic in a population of this size took 1.5–3 minutes. Post-processing (checking the simulation executed properly, finished, and producing graphs and statistics of the epidemic) took approximately 1 minute, and sampling also took

approximately 1 minute.

Simulations were then run with a much larger starting population of 4,000 low-risk, 4,000 medium-risk, and 800 high-risk individuals for a much longer time period of 140 years. Again, five replicate simulations were created for each level of heritability. Of HIV-infected individuals who had been infected for more than 3 months between years 120-140, a random 15% were sampled. Again, viral phylogenies were generated using the VirusTreeSimulator, and the resulting phylogenies and viral load data was run through the heritability pipeline. Again, the resulting heritability estimates were standardized by the average root-to-tip distance.

The increase in population size severely affected all aspects of the run-time for the simulations. The simulations themselves ran in 1.2–2 hours. Post-processing took another 2–3 hours, and sampling took approximately an hour. Though simulations, post-processing, and sampling were performed in parallel for the different heritability levels, five replicates of the four levels of heritability took a total of 100 hours of computational time.

6.2.3 Results

Initially, simulations were run for 70 years with a starting population size of 2,200 hosts and exponential growth, and sampled at 100% density between years 40-50. Five replicates were run for each heritability value of 30%, 50%, 70%, and 90%. The heritability estimates resulting from these simulations did not match up well with the actual heritability values the runs were simulated under. The mean heritability estimates produced were 75.8% (CI 66.1–85.5%), 83.1% (CI 75.9–90.4%), 88.5% (CI 84.1–92.8%), and 94.2% (CI 92.1–96.3%) for simulated heritabilities of 30%, 50%, 70%, and 90%, respectively (see Figure 6.13 on the following page).

ASReml assumes that the traits being measured evolve along the branch lengths according to Brownian motion. However, the assumption made in the DSPS simulations of HIV infection is that set-point viral load only changes at transmission, when a new set-point viral load is calculated based upon the heritability equation, and that viral load is constant along the branch lengths in between transmissions. It is also assumed that this holds true in real HIV transmission – an individual's set-point viral load is

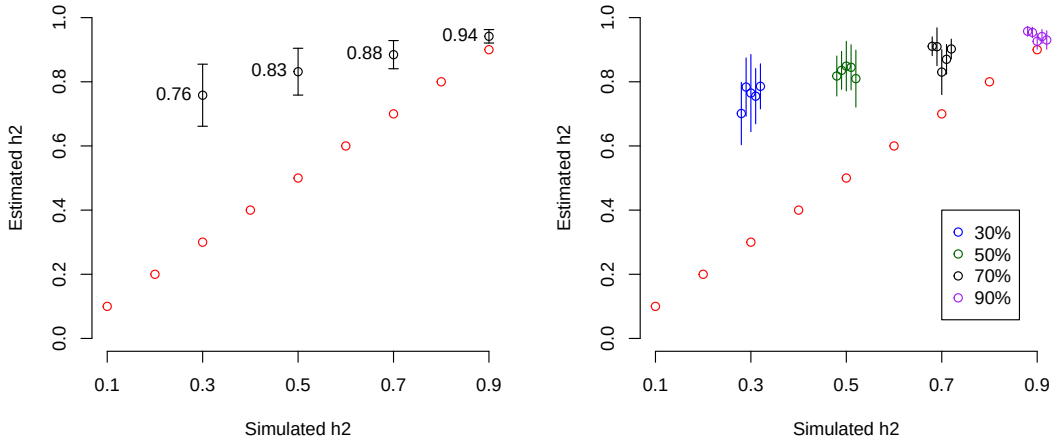

Figure 6.13: Heritability estimates from simulations in the DSPS where heritability was varied. Simulations were run in an exponentially growing population for 70 years, with a starting population of 1,000 low risk, 1,000 medium risk, and 200 high risk hosts. Simulations were sampled at 100% density between years 40-50, and viral phylogenies were generated. Five replicate simulation runs were performed for each heritability level. The phylogenies and viral load information was run in the heritability estimation pipeline. On the left, the mean heritability estimate for each heritability level is plotted, with the 95% CI. On the right, the estimate resulting from each of the replicates is plotted, with the 95% CI of each estimate shown as a line extending from the estimate. The red dots show $x = y$.

constant throughout their asymptomatic phase, but when they transmit to another person the new set-point viral load could be different.

In the real UK HIV RDB data, the time period covered by the sequences is at least 24 (subtype B; Chapter 3.4.1) to 45 (subtype C; Chapter 4.3.3) years, and encompasses thousands to tens of thousands of past transmissions. Though the coverage of the UK HIV DRB is admirable (in 2006, it was estimated to contain two-thirds of subtype B-infected individuals in the UK who were receiving treatment (Leigh Brown et al., 2011)), it is not 100%, and the use of only sequences with at least one viral load before starting ART will have further decreased the sampling fraction. The long time period until the root of the UK HIV RDB sequences and the low sampling fraction of the datasets used here combine to produce a tree where many 'hidden' transmissions will have taken place at time points along the branches. This effect produces a tree where viral load is 'evolving' along the branch, due to the 'hidden' transmissions that occurred on that branch.

In simulations covering a short time period where only a limited number of transmissions have occurred, and where there is a high sampling density, there will be fewer 'hidden' transmissions along the branch lengths. When sampling density is high, this will become more pronounced at the tips, as individuals may link directly to a transmission partner, creating a branch where there are no 'hidden' transmissions, and thus almost no 'evolution' of viral load along the branch length. This causes the variance in viral load over time to be lower than what ASReml expects under Brownian motion. This reduction in variance is attributed to the phylogeny as being due to a genetic effect, causing an over-estimation of the heritability of viral load (see Figure 6.13).

One way to address this issue is to run a larger, longer simulation with a lower sampling density so that after sampling, more transmissions are 'hidden' along the branch length, creating the effect of viral load evolving along the branch. Another way is to convert the DSPS output into the pedigree format usually accepted by ASReml, allowing complete sampling, and including viral load measures for all infected individuals. However, the heritability estimation pipeline was developed specifically to extract heritability estimates from phylogenies rather than pedigrees, so bypassing this step is not a very strong test of the pipeline's abilities.

Thus, a second round of larger simulation runs was conducted. Runs lasted 140 years, and had a starting population of 8,800 individuals and experienced exponential growth. 15% of HIV-infected individuals were randomly sampled between years 120-140. Five replicates were run each for heritability values of 30%, 50%, 70%, and 90% (only four completed successfully for the 70% simulations), and produced estimates of 10.8% (CI 3.9–17.8%), 56.1% (CI 46.0–66.2%), 83.0% (CI 78.5–87.5%), and 96.0% (CI 94.9–97.1%), respectively. The mean and individual replicate heritability estimates are shown in Figure 6.14 on the next page.

Though the estimates from the new, larger simulations suggest slightly less overestimation of the heritability than previously (see Figure 6.13), they still do not align well with the heritability level that was simulated. Because of this, I decided to test the simulations by converting the line list output of the DSPS to ASReml-format 'pedigrees,' with each infected individual having the host that infected them as their 'parent,' tracing back to the initial infection at the start of the simulation. Though this bypasses

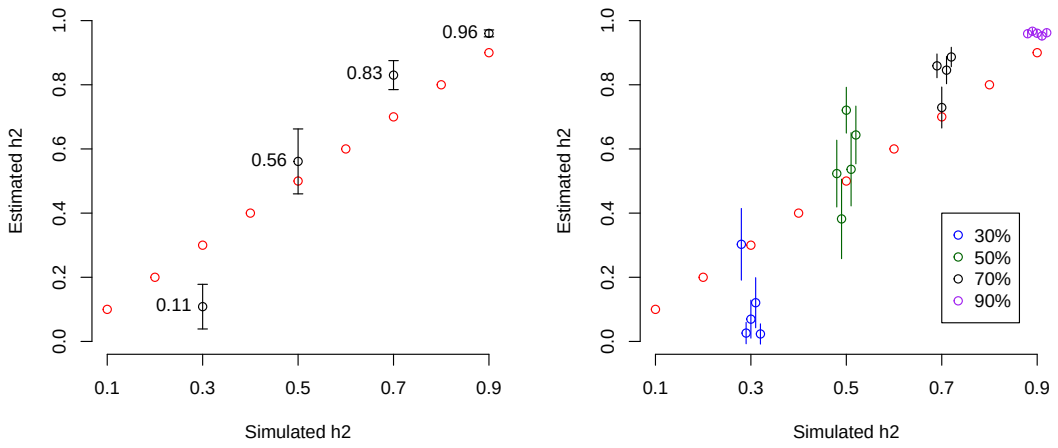

Figure 6.14: Resulting heritability estimates from simulations in the DSPS where heritability was varied. Simulations were run in an exponentially growing population for 140 years, with a starting population of 4,000 low risk, 4,000 medium risk, and 800 high risk hosts. Simulations were sampled at 15% density between years 120-140, and viral phylogenies were generated. Five replicate simulation runs were performed for each heritability level. The phylogenies and viral load information was run in the heritability estimation pipeline. On the left, the mean heritability estimate for each heritability level is plotted, with the 95% CI. On the right, the estimate resulting from each of the replicates is plotted, with the 95% CI of each estimate shown as a line extending from the estimate. The red dots show $x = y$. Only four replicates completed for the 70% heritability value.

the heritability estimation pipeline's ability to utilize phylogenies, it provides a way to check that the simulations are behaving as expected. If so, then when provided data in this format, ASReml should return very good estimates of the heritability runs were simulated under.

The results of running the simulated data in pedigree format confirmed that the simulations were not behaving as expected with regard to the viral load values. The ASReml runs failed completely, giving errors due to the residual variance shrinking to impossibly small values as the variance ratio grows exponentially (the error given refers to 'a singularity in the AI matrix'). This often occurs when there is some familial ('group') effect that is overshadowing any genetic effect (for example, a strong sire effect that is not controlled for in the analysis). In the context of a one-parent transmission chain this is difficult to interpret, but suggests that the viral loads generated by the DSPS are violating the expectations of the Brownian motion ASReml model.

The equation being used to generate the heritability standardizes the viral load values to the population mean, to prevent them from becoming unrealistically large (see Equation (6.2) on page 185). JH has since pointed out that by constraining the viral load to realistic values in this manner, the equation prevents viral load from evolving according to Brownian motion. The original equation from Alizon et al. (2010) (see Equation (6.1) on page 184) **loses** variance at every transmission, and the modified equation using standardized viral loads (see Equation (6.2) on page 185) **maintains** variance, but ASReml assumes that the variance **increases** over time according to Brownian motion. Thus, by using standardized viral loads to maintain realistic viral load values, the assumption of Brownian motion is violated, and the ASReml analysis fails.

In order to obtain simulated data sets that do not violate the assumption of Brownian motion, I changed the DSPS to use a modified version of Alizon et al. (2010)'s equation, without standardizing the viral loads (first introduced in Chapter 5.3.4):

$$x_{a+1} = \sqrt{\zeta} x_a + \sqrt{(1-\zeta)} y \tag{6.3}$$

Using the same simple simulation detailed in Section 6.2.1, I simulated 20 replicates of 50 generations with a heritability of 50%, which showed the variance increasing by 1.7 on average.

However, allowing viral loads to grow to unrealistic values would be problematic to the running of the DSPS. Viral load dictates transmission risk and also the length of the asymptomatic phase (Chapter 5.3.3). An individual with a viral load of 4.5 $\log_{10}$ copies/mL will be in the asymptomatic stage for 7.0 years on average, whereas an individual with a viral load of 8.0 $\log_{10}$ copies/mL will stay in the asymptomatic phase only 0.35 years on average (just over 4 months), according to the implemented equation from Fraser et al. (2007) (Chapter 5.3.3). This suggets that as viral load grows larger, individuals will progress to AIDS (where they have no contacts) more and more quickly, and eventually so quickly that almost no individuals have a chance to infect anyone before progressing to AIDS. At this point, the epidemic would die out.

Surprisingly, this isn't the dynamic observed in the preliminary runs. There does

seem to be a higher chance of the epidemic dying out very quickly, as early in the epidemic when viral load values rise, there is a chance that all those infected progress to AIDS before infecting anyone. However, in about 1/3–1/2 of simulations, the epidemic does continue growing until the end of the simulation time period. This happens because as more and more people become infected and very quickly pass the calculated time when they should progress to AIDS, the recovery parameter of 0.8 is not high enough to generate enough recovery events to move everyone out of the asymptomatic phase within the time period they should be progressing to AIDS. In the PANGEA_HIV simulations, where everyone had a viral load of 4.5 $\log_{10}$ copies/mL, post-simulation analysis revealed that infected individuals spent on average 7 years in the asymptomatic phase, matching the calculated length of the asymptomatic phase according to the population viral load, and implying that the recovery parameter of 0.8 was appropriate. However, as the viral load increases and everyone almost immediately becomes eligible to progress to AIDS, the recovery parameter of 0.8 is not high enough to generate enough events to move everyone to AIDS when they should be progressing. As an example, a cursory examination of DSPS output from these simulations showed that an individual with a viral load of 15.3 $\log_{10}$ copies/mL who should have been in the asymptomatic stage for only 3 hours instead remained in the asymptomatic period for 10 years.

Even in runs where the epidemic does successfully grow in the population, the dynamics are still unusual. As shown in Figure 6.15 on the next page, the number of infected individuals is prone to dips and spikes, and the proportion of HIV infected individuals in the population never rises above 4%. This, combined with the fact that the DSPS is no longer acting as expected made me hesitant to use these runs to test the heritability estimation pipeline. Instead, I decided to turn off the relationship between viral load and disease progression and transmission risk. Rather than calculating the asymptomatic stage length and transmission risk from the viral load, both were calculated as if every individual had a viral load of 4.5 $\log_{10}$ copies/mL. This allows viral load to evolve through the run to unrealistic values while the simulation continues to behave in all other aspects in a realistic fashion, with individuals spending a reasonable time in the asymptomatic period and having a reasonable transmission risk.

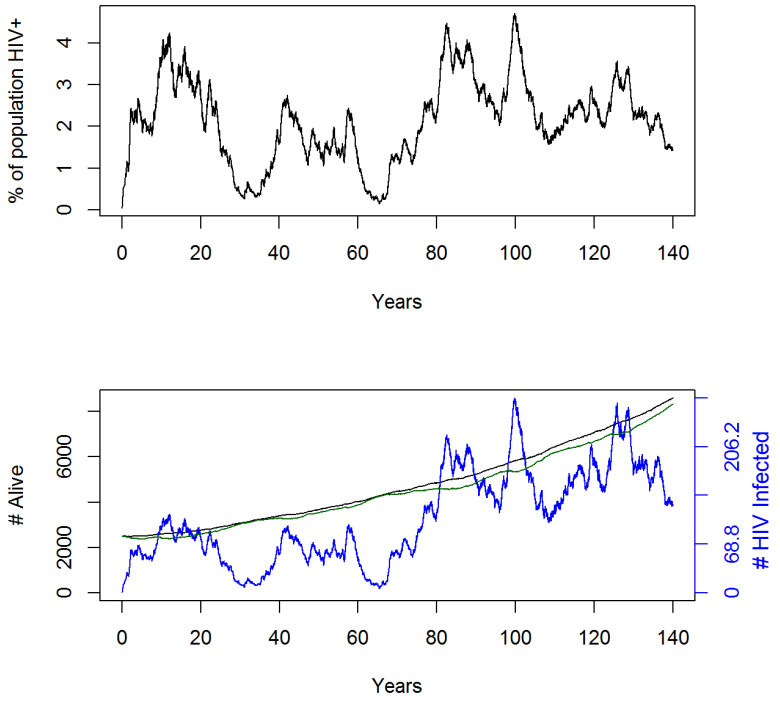

Figure 6.15: The epidemic dynamics in a preliminary DSPS simulation where equation 6.3 was used to generate new viral load values according to the simulated heritability. This equation allows viral loads to grow to unrealistic values in order to allow variance to increase over time. The simulation was run with exponential growth and a starting population of 1,000 low-risk, 1,000 medium risk, and 200 high risk individuals. The simulation was run for 140 years with a simulated heritability value of 70%. Though this epidemic does grow in the population, it exhibits unusual dynamics compared to previous epidemics without treatment (see Figure 6.8), with many dips and spikes. The proportion of the population infected with HIV also does not rise above 4%.

Due to time constraints, a smaller starting population size and runtime was chosen so that the simulations and sampling could complete more quickly. The starting population included 1,000 low-risk, 1,000 medium-risk and 200 high-risk individuals, and ran for 100 years. Ten replicate simulations were run for each level of heritability (30%, 50%, 70%, 90%), and 15% of the HIV-infected hosts who had been infected for more than 3 months between years 60-80 were randomly sampled. As before, VirusTreeSimulator was used to generate viral phylogenies, and the viral load data and viral phylogenies were run through the heritability estimation pipeline. The heritability estimates were standardized by the average root-to-tip distance. The mean heritability estimated for the simulated values of 30%, 50%, 70%, and 90% was 14.5% (CI 0.0–30.1%), 51.9% (CI 33.2–70.5%), 86.1% (CI 78.0–94.2%) and 98.9% (CI 97.5–1.0%), respectively, and

the resulting heritability estimates obtained from this run are shown in Figure 6.16. Running these simulations takes approximately 15 minutes, post-processing takes 2–8 minutes, and sampling takes 2 minutes.

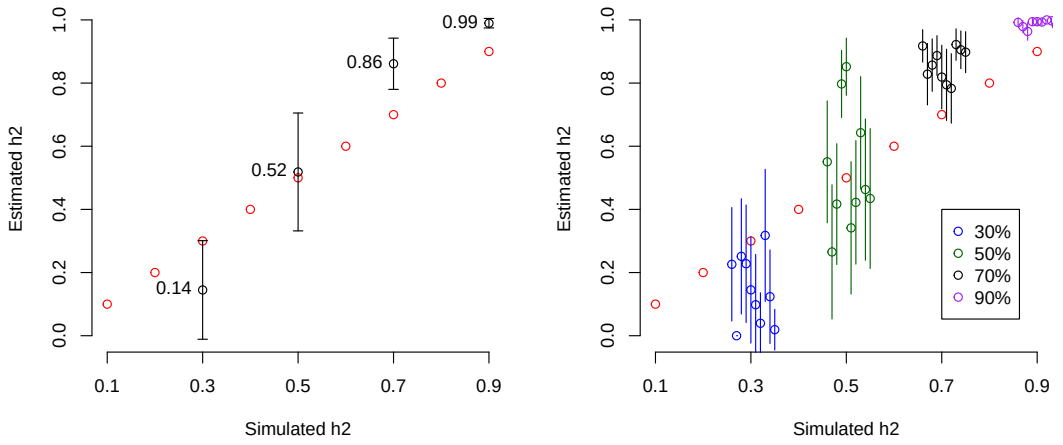

Figure 6.16: Heritability estimates from simulations in the DSPS where heritability was varied. Simulations were run in an exponentially growing population for 100 years, with a starting population of 1,000 low risk, 1,000 medium risk, and 200 high risk hosts. Simulations were sampled at 15% density between years 60-80, and viral phylogenies were generated. Ten replicate simulation runs were performed for each heritability level. The phylogenies and viral load information was run in the heritability estimation pipeline. On the left, the mean heritability estimate for each heritability level is plotted, with the 95% CI. On the right, the estimate resulting from each of the replicates is plotted, with the 95% CI of each estimate shown as a line extending from the estimate. The red dots show $x = y$.

Once again, the heritability estimates obtained do not correspond very well with the simulated values. This may well be due to the fact that though Equation 6.3 allows variance to increase over time, it also fails to generate true Brownian motion (see Chapter 5.3.4). Viral load does not increase for the entire simulation as originally expected, however – it reaches a limit and then stays constant, as shown in Figure 6.17 on page 199. In a true Brownian motion, the viral load would be expected to decrease or remain unchanged at the same probability of increasing, and would not reach a limit.

This limit is reached due to the fact that some proportion of the new viral load always comes from the population distribution of viral loads (Chapter 5.3.3), which range from 1.69 to 7.26 $\log_{10}$ copies/mL with a mean of 4.5 $\log_{10}$ copies/mL in the subtype B dataset used here. At the beginning of the simulation, the viral load rapidly

increases, as $\sqrt{\zeta}$ and $\sqrt{1-\zeta}$ sum to more than one, which always produces a larger number if the parental viral load (multiplied by $\sqrt{\zeta}$) and the viral load chosen from the population distribution (multiplied by $\sqrt{1-\zeta}$) are relatively close in magnitude, which they are at the simulation start. The parental viral loads continue to increase while the population distribution stays constant, and the value that the mean viral load can reach is limited by the heritability value being simulated. At 30% heritability, 40% of the new viral load value comes from the parental viral load and 60% of the new viral load value comes from the population distribution. As the parental viral load values grow, the comparatively very low viral loads from the population distribution counter the influence of the proportion of the viral load contributed by the parental viral load, reaching a limit. This limit is lower for low heritability values, as the (high) parental viral loads contribute less and the (low) population distribution viral loads contribute more, but high for high heritability values, as up to 75% of the new viral load is determined by the parental viral load.

It should be noted that Shirreff et al. (2013) also used a modified version of Alizon et al. (2010)'s equation to simulate heritability down phylogenies, which is very similar to Equation 6.3:

$$x_{a+1} = \zeta x_a + y\sqrt{(1-\zeta^2)} \tag{6.4}$$

However, this equation suffers from the same flaws as Equation 6.3, with viral loads increasing to unrealistic values and then reaching a limit, though the limit is lower due to the larger influence of the population distribution of viral loads when multiplied by $\sqrt{1-\zeta^2}$. Shirreff et al. (2013) do acknowledge that this equation constrains the distribution of viral loads generated, but this still violates the assumption of true Brownian motion assumed in ASReml, and so cannot be used in the DSPS simulations.

None of the simulated epidemics run produced heritability estimates that correlate well with the heritability values they were simulated under. Unfortunately, time and computational limitations prevented further DSPS simulations of heritable viral load.

6.2.4 Discussion

The simulation of heritable viral load values proved to be considerably more difficult than initially expected. The initial runs simulating viral load consistently produced overestimates of the simulated heritability values (Figure 6.13 on page 188), which led to the realisation that short runs with very high sampling frequency violate the assumption that viral load evolves continuously along branch lengths rather than discretely at branching (transmission) events. I addressed this issue by running simulations with larger starting populations for twice as long, and sampling at much lower frequency (Figure 6.14 on page 190). This may indeed generate enough 'hidden' transmissions that viral load appears to evolve along the branch length, as the heritability estimates obtained from these runs are not all overestimates. The heritability estimates obtained from this run still do not correlate well with the simulated values, and in investigating this, seriously flaws were uncovered in the underlying equations used.

Of the three equations (6.1, 6.2, 6.3) used here to generate new viral loads at transmission depending on the heritability being simulated, none approximate the evolution under Brownian motion that is assumed in ASReml. Alizon et al. (2010)'s original equation (6.1) loses variance at every transmission. When modified so that variance would increase yet viral loads would maintain realistic values (6.2), variance is not lost, but does not increase with time. Finally, allowing viral loads to grow to unrealistic values allowed equation 6.3 to generate viral loads with increasing variance through time, but as the viral loads always increases and then stabilizes, rather than randomly evolving, it is not true Brownian motion.

The results from using Equation 6.2 (which maintains variance and standardizes viral loads; Figure 6.14 on page 190) on a very large population running for 140 years and Equation 6.3 (which allows unrealistically high viral loads but allow variance to increase; Figure 6.16 on page 194) on a much smaller population for only 100 years look fairly similar. It may be that the much larger population size is not necessary to generate enough 'hidden' transmissions to overcome the problem of viral load not evolving along the branch lengths, and that simply allowing the smaller population to grow for slightly more time and then sampling at 15% is sufficient to prevent over-

estimation of the heritability. Both sets of runs also used equations that violate the assumptions of Brownian motion, though in different ways – the first prevents variance from increasing with time, and the second isn't a true Brownian walk, as all new viral loads generated will be larger than or equal to their parental viral load. It may be that both of these violations produce similar simulated viral load distributions that lead to similar inaccurate heritability estimates. Further runs separating out the effect of run time and sampling from the equation used could clarify this.

The goal of using the DSPS to simulate viral load data under different heritability values was to provide a dataset that could be used to test the accuracy of the heritability estimation pipeline. To do this effectively, the DSPS must be able to simulate realistic viral load data under the same conditions (Brownian motion) that are assumed in ASReml. Despite a variety of different simulations run, none of the equations used are appropriate, and so a reliable dataset on which I could test the DSPS was not obtained. To ensure the simulated viral loads contain the heritability signal expected, the technique of converting the DSPS line list output to ASReml pedigree format can be used as a test of whether the correct heritability value is estimated, excluding the potential complications of extracting information from a phylogeny. Simulations that lead to accurate heritability estimates from pedigree format files can then have viral phylogenies generated to put through the full heritability estimation pipeline as a test data set.

6.3 Conclusion: The DSPS as an HIV Epidemic Simulator

Here, the DSPS was successfully used to generate complex, realistic HIV simulations with differing dynamics to use as test data sets for phylogenetic methods. The DSPS was greatly improved by the challenges presented in getting an epidemic to grow in a realistic contact network and creating datasets with variations in acute phase and treatment. The flexibility demonstrated by the DSPS in creating both African-village-like and Western-MSM-like epidemics suggests that simulations produced by the DSPS could prove very useful to test a variety of phylogeny-based methods. Unfortunately,

I did not succeed in using the DSPS to simulate realistic viral loads under differing heritability values to use as a test data set for the heritability estimation pipeline.

The DSPS is a core part of my ongoing research, and the issues raised here highlight the weaknesses of the DSPS that need to addressed in the near future.

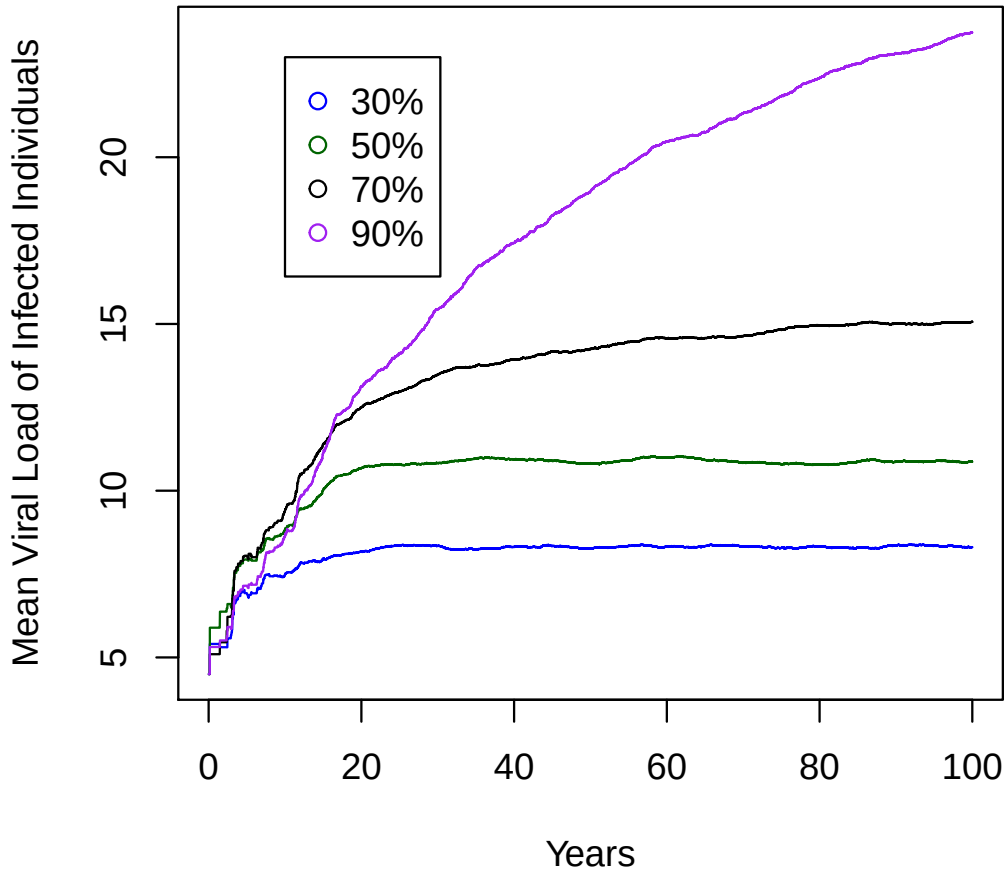

Figure 6.17: The mean set-point viral load value of all alive, infected individuals, through time. One simulation at each heritability level was run using Equation 6.3, allowing unrealistic viral loads, but not allowing viral loads to determine disease progression or transmission risk. Though viral load increases swiftly during the beginning of the run, it eventually reaches a limit and stays constant. While the mean viral load in the 90% simulation has not yet reached a limit, the shape of the curve suggests that it would if the simulation was allowed to continue running.

"All I ever wanted was to know this world. I can say now ... that I know quite a bit more of it than I knew when I arrived. Moreover, my little bit of knowledge has been added to all the other accumulated knowledge of history – added to the great library, as it were. That is no small feat, sir. Anyone who can say such a thing has lived a fortunate life."

Elizabeth Gilbert - 'The Signature of All Things' (2013)

7 Discussion

7.1 Estimating the Heritability of Viral Load

Set-point viral load is an important clinical marker in HIV, as it predicts disease progression (Mellors et al., 1996; Fraser et al., 2007; Langford et al., 2007) and transmission risk (Quinn et al., 2000; Fideli et al., 2001). Set-point viral load is highly variable between individuals, but the cause for this variation is not fully understood. Host genetic effects such as HLA type (Steel et al., 1988; Kaslow et al., 1990; O'Brien and Nelson, 2004; Tang et al., 2004; Fellay et al., 2009; Salgado et al., 2010) have been shown to influence viral load, along with demographic effects like age (O'Brien et al., 1996; Nogueras et al., 2006) and gender (Farzadegan et al., 1998; Sterling et al., 1999; Gandhi et al., 2002). However, the effect of the viral genome on viral load is less clear.

The idea that HIV could be evolving to become more virulent has lead to decades of research investigating whether prognostic markers like viral load have changed over time (Chapter 1.2.1), and hypotheses about whether the trade-off between asymptomatic phase length and transmission risk would lead to an intermediate 'optimal' viral load (Fraser et al., 2007). However, in order for set-point viral load to evolve, there must be a heritable viral genetic component influencing viral load.

Previous studies attempting to estimate the heritability of viral load in HIV have mostly relied on transmission-pair analyses (Tang et al., 2004; Hecht et al., 2010;

Hollingsworth et al., 2010; van der Kuyl et al., 2010; Yue et al., 2013; Lingappa et al., 2013), which are limited in size and potentially influenced by a number of confounding effects. Only one study has made use of a phylogenetics-based method (Alizon et al., 2010), which could potentially avoid some of these issues, though the stringent inclusion criteria also greatly reduced the potential sample size. These studies have produced heritability estimates ranging from 1–60%, making the results hard to interpret.

Here, I have described a new method to estimate the heritability of viral load in HIV. By adapting an efficient REML-based method of variance component estimation commonly used in animal breeding to incorporate information from HIV phylogenies, I have created a pipeline that allows the inclusion of more HIV samples than ever previously used, while avoiding the confounding effects associated with transmission pair analyses. In order to include as many samples as possible, a very liberal definition of set-point viral load was adopted. Generally, the first viral load available was used, but data-cleaning rules were employed to eliminate viral loads that were potentially taken during the acute phase, during AIDS, or while on unreported ART (Chapter 2.3). 8,483 subtype B and 1,821 subtype C samples were analysed using the heritability estimation pipeline, the largest datasets ever used to estimate heritability in HIV.

In subtype B, the mean heritability of set-point viral load was estimated at 5.7% (CI 2.8–8.6%), and this estimate proved robust to collapsing poorly supported nodes, using less thorough phylogenetic reconstruction methods, and removing all samples with only one viral load available (Chapter 3.4.1). A 652 sequence sub-sample of the subtype B dataset was used to construct a time-dated phylogeny, where the estimated root of the UK subtype B sequences was placed at 1985 (1983-1987) (Chapter 3.3.3). The phylogenetic effect on viral load was plotted back onto the time-resolved tree in order to visualise how the viral genetic effect on set-point viral load varies across the phylogeny, and illustrated that some lineages are associated a positive effect on viral load, while others are associated with a negative effect (Figure 3.2). Finally the time-resolved tree was used to investigate the change in viral load over time due to between- and within-host selection (Chapter 3.4.4), which I estimated to have caused a small but significant decrease over time of -0.05 $\log_{10}$ copies/mL/year in the component of viral load determined by the viral genotype.

In subtype C, the mean heritability of set-point viral load was estimated at 29.7% (CI 14.8–44.7%). Collapsing poorly supported nodes and removing all samples with only one viral load available did not affect the heritability estimate, but using less thorough phylogenetic reconstruction methods greatly reduced the ability to detect a significant heritability signal (Chapter 4.4.1). A 350 sequence sub-sample of the subtype C dataset was used to construct a time-resolved phylogeny, where the estimated root of the UK subtype C sequences was placed at 1963 (1952-1972) (Chapter 4.4.2). However, a heritability signal could not be detected from the 350 sequence sample dataset. Larger samples and alternate dating methods were used in an attempt to obtain a time-resolved tree with a significant heritability estimate, but none were successful (Chapter 4.3.3). Though it did not produce a significant heritability estimate, the 350 sequence time-resolved tree was used to illustrate the phylogenetic effect on viral load across the tree, showing that some lineages are associated with a positive effect on viral load, and others with a negative effect (Figure 4.3). Because a time-resolved tree with a significant heritability estimate was not obtained, the change in viral load over time could not be investigated.

7.2 Defining and Comparing Heritability Estimates

The heritability estimates obtained for the subtype B and subtype C datasets (5.7% and 29.7%, respectively) differed greatly, which suggests that subtype C has more viral genetic influence on viral load. Three previous studies had datasets that were primarily subtype B, and two had datasets that were primarily subtype C.

Two of the previous heritability estimates from subtype B datasets were done using transmission pair analysis, and estimated the heritability of viral load at between 25–55% (Hecht et al., 2010; van der Kuyl et al., 2010). Alizon et al. (2010)'s phylogenetic method was also used on a subtype B dataset, where in MSM a heritability of 9% and 50–60% was estimated for the 'liberal' and 'strict' definitions of set-point viral load, respectively. My estimate of 5.7% only compares favourably to Alizon et al. (2010)'s 'liberal' set-point estimate of 9%. It is possible that confounding effects in the two transmission pair studies cause some of the similarity in viral load to be attributed

to viral genetic effects when it is actually due to other factors, upwardly biasing the heritability estimates. The fact that the heritability estimate obtained here from the subtype B dataset is similar to Alizon et al. (2010)'s 'liberal' set-point estimate is interesting, as this study was the only other analysis to use a phylogeny-based method. The discrepancy observed between the 'liberal' and 'strict' estimates may indicate that the 'strict' estimate is only specific to a small population of MSM with exceptionally stable viral load measures.

Both of the two previous subtype C analyses were done using transmission pairs, and estimated the heritability of viral load at 1.3–2% and 21% (Tang et al., 2004; Yue et al., 2013). Tang et al. (2004)'s estimate of 21% corresponds well with my own, but Yue et al. (2013)'s estimate of <3% is hard to reconcile. Both the heritability estimation and the reconstruction of a time-resolved tree proved to be more difficult in the subtype C dataset than in subtype B. As discussed more extensively in Chapter 4.4.6, subtype C seems to have transmitted to the UK directly from the much larger original subtype C epidemic in Africa through immigration of infected individuals. This means the subtype C dataset is potentially a 'sub-sample' of the global subtype C epidemic, with very divergent tips, and little information deeper in the tree about the history of the subtype, making the phylogeny more difficult to reconstruct.

7.2.1 Comparing Heritability Estimates: MRCA

Comparisons between different estimates of the heritability of viral load are rarely straightforward, as estimates often come from very different populations, ethnic groups, risk groups, and subtypes. This comparison is made more complex still by the lack of standardization of heritability estimates.

If a phylogeny is dominated by neutral evolution, the variance of the phenotype at the tips is expected to increase with increasing time to the most recent common ancestor (MRCA). Thus, heritability estimates between studies where the sequences have different times to their respective MRCA are not really comparable. Studies with a more distant MRCA are likely to have higher heritability estimates, as the distance from the tips to the root is greater. One possible solution to this is to compare the variance in units of time or substitutions rather than in heritability, although as the

sequences and thus time to the MRCA are often not included in the publications for partner transmission studies, such scaling cannot be done readily.

This can be done for the subtype C and subtype B estimates presented here, however. In the original resistance-site-stripped RAxML subtype C runs, the average root-to-tip distance of the two replicates was 0.1497 and 0.2039 substitutions/site. Rather than scaling by these values, the heritability estimates were instead scaled by the average root-to-tip distance in the two subtype B replicates (0.1412 and 0.1473 substitutions/site), to estimate the heritability of subtype C over the same time period as subtype B. This brought the original subtype C estimates of 34.1% (CI 19.5–48.6%) and 25.4% (CI 10.0–40.7%) (mean 29.7% (CI 14.8–44.7%)) down slightly to 32.8% (CI 18.5–47.0%) and 19.7% (CI 6.9–32.6%) (mean 26.3% (CI 12.7–39.8%)). This suggests the difference in the heritability estimates between the subtypes B and C is not an artefact of the different time to the MRCA of the two subtypes, and is instead a difference due to differing genetic control over viral load between the two subtypes.

7.2.2 Heritability Metrics: Are they Equal?

As a further complication in comparing heritability estimates, several different statistics have been used to express the proportion of the variance in viral load that is determined by the viral genome (shown in Table 1.1). Transmission pair studies have generally used Pearson's correlation coefficient (often denoted r or ρ) (Tang et al., 2004; Hecht et al., 2010; van der Kuyl et al., 2010) or the coefficient of determination (R^2) (Hollingsworth et al., 2010; Yue et al., 2013; Lingappa et al., 2013), while Alizon et al. (2010)'s phylogenetic analysis used two different measures of phylogenetic signal, K and λ. As ASReml partitions variance and calculates the proportion due to genetics, I have attempted to measure heritability more directly (h^2), and so use none of these statistics. The many different metrics used to describe an estimate of 'heritability' raise the question as to whether all of these statistics are indeed measuring the same thing.

Correlations, Coefficients, and Regressions

This issue was first explicitly raised by Fraser and Hollingsworth (2010), where they compared estimates of heritability using Person's correlation coefficient (r) from Tang

et al. (2004), Hecht et al. (2010), and van der Kuyl et al. (2010) to the coefficient of determination of the ANOVA (R^2) estimate from Hollingsworth et al. (2010). Though they determined that ANOVA provided better estimates of the heritability due to the ability to adjust for confounding factors, they found that if the set-point viral load values in the recipient and transmitter are normally distributed with identical variance, then $R^2 = r$ (Fraser and Hollingsworth, 2010). This was again confirmed by Müller et al. (2011), who also went on to conclude that r, R^2, and the phylogenetic signal measures used by Alizon et al. (2010) all provide numerically comparable estimates of the heritability of viral load.

However, Fraser et al. (2014) argue that R^2 is not a direct measure of heritability, and that the slope of the regression of recipient on transmitter viral load (b) is the best measure of the heritability of set-point viral load, conceding that r is an acceptable, but less accurate, measure. Fraser et al. (2014) re-analysed the Hollingsworth et al. (2010) data using regression rather than ANOVA, obtaining a heritability estimate of b=36%, and re-interpreted the heritability estimates from Yue et al. (2013) and Lingappa et al. (2013), using the regression slopes rather than the R^2 to give heritability estimates of b=26% and b=44%, respectively. The transmission pair studies included in Table 1.1 on page 16 are repeated below in Table 7.1, this time with information on the method used, and the heritability estimates given as r, R^2, and b whenever possible. It is apparent from this table that estimates obtained by these different methods and using these different metrics can differ greatly, even within the same dataset.

By using a direct, commonly accepted method of estimating heritability as h^2, my analyses have avoided the complications of equating correlation and regression measures to heritability estimates. However, the question of how to interpret the many different answers obtained, even from the same data, when using different regression and correlation measures to estimate heritability makes it difficult to compare my results to those of other studies, and other studies to each other. Further work in comparing the estimates resulting from these methods, perhaps combined with simulated data, will hopefully help clarify which estimators are equivalent.

Table 7.1: Estimated Heritability of Viral Load in Previous Studies, with all Available Heritability Metrics Shown

Paper	**Country**	***N***	**Method**	**Heritability Estimate**		
				Pearson's Correlation	**Coefficient of Determination**	**Slope of the Regression**
Tang *et al.* 2004	Zambia	115 pairs	Generalized linear model	$r = 0.21^{\dagger}$	$R^2 = 0.04$	$b = 0.36$
Hecht *et al.* 2010	USA	22 pairs, 1 triplet	Pearson's correlation	r=0.55†	NA	$b = 0.45$
Hollingsworth *et al.* 2010	Uganda	29 pairs	ANOVA	NA	$R^2 = 0.27^{\dagger}$	$b = 0.36^{a}$
van der Kuyl *et al.* 2010	Netherlands	56 pairs	Linear regression	$r = 0.25^{\dagger}$	NA	NA
Yue *et al.* 2013	Zambia	195 pairs (1 VL) 143 pairs (mean VL)	Generalized linear model	$r = 0.14^{b}$ NA	$R^2 = 0.020^{\dagger}$ $R^2 = 0.013^{\dagger}$	$b = 0.28$ $b = 0.26$
Lingappa *et al.* 2013	East & South Africa	141 pairs	Multivariate linear mixed effects modelling	NA	$R^2 = 0.06^{\dagger}$	$b = 0.44$

† The estimate given as 'heritability' by the authors
^{a} Data re-analysed by Fraser et al. (2014)
^{b} Univariate analysis only (all other estimates from multivariate analysis)

Phylogenetic Signal Estimators

Though only Alizon et al. (2010) has used phylogenetic signal methods to estimate the heritability of set-point viral load in HIV, they also used two different heritability metrics – Blomberg et al. (2003)'s K and Pagel (1999)'s λ – which produced slightly different heritability estimates. As well as these two measures of phylogenetic signal, there are a number of others, including the Mantel test (Mantel, 1967), Bloomberg's PICv statistic (Blomberg et al., 2003), and the five variants of the Abouheif-Moran (AM) test (Pavoine et al., 2008).

Shirreff et al. (2013) investigated all of these methods, plus two new methods they developed: phylogenetic pairs (PP), based on inferred transmission pairs in a phylogeny, and hierarchical clustering (HC), which measures the amount of variance in viral load that clusters of closely related sequences can explain. Each method was extensively tested both on three independent real datasets from Uganda, Switzerland, and the Netherlands, and on simulated data. Shirreff et al. (2013) concluded that Pagel's λ, PP, HC, and AM were comparatively sensitive, consistently detecting an effect when heritability was simulated at values above 40%, though the methods performed more variably on the real data sets. Unfortunately, the heritability estimation pipeline outlined in this thesis was not yet available to be tested at the time Shirreff et al. (2013) were completing their analysis.

Shirreff et al. (2013) conclude that none of the methods tested seem able to detect heritability signals below 40% in the simulated data sets, which if true, would make the heritability estimation pipeline outlined in this thesis unique in its ability to consistently estimate a heritability of around 6%, as shown in the UK subtype B dataset (Chapter 3).

7.3 The DSPS: An HIV Epidemic Simulator

A HIV-specific agent-based model, the DSPS, was created in order to simulate realistic HIV epidemics. The DSPS allows flexible specification of contact networks, exponential growth, gendered hosts and appropriate contacts, and heritable viral load, which controls transmission risk and disease progression. The DSPS was used to create two

sets of simulations for PANGEA_HIV which approximate subtype C epidemics in small African villages through heterosexual contact, sex workers, and transmissions from surrounding areas (Chapter 6.1.1).

Though the DSPS was successful in generating simulations under a variety of conditions for the PANGEA_HIV release data sets, I was unable to configure the DSPS to simulate realistic viral load values that did not violate the assumption of Brownian motion (Chapter 6.2.2). However, further work and collaboration with other researchers working on the heritability of viral load and HIV modelling should lead to an equation that can be used to generate realistic viral loads.

Herbeck et al. (2014) have also developed an agent-based model with heritable viral load, which they have used to investigate changes in set-point viral load over time. They found that viral load evolves over time to an intermediate value that balances transmission risk with disease progression, as predicted by (Fraser et al., 2007), and also suggest that different estimates of change in viral load over time may be due to the different histories of the epidemics being measured, such as the viral load of the founding lineage or the age of the epidemic (Herbeck et al., 2014). Herbeck et al. (2014) allowed the heritability of viral load to change over the course of simulation, and found it to decrease over time.

The insights provided by this model highlight the potential of the DSPS to further explore similar questions, including change in viral load and heritability in simulated populations modelled after specific epidemics or demographic, as the contact structure in the DSPS is more customizable than that used by Herbeck et al. (2014). The flexibility would also allow further investigation of what effect different types of epidemic and contact structure could have on the ability of methods such as the heritability estimation pipeline to reliably estimate the heritability of set-point viral load.

7.4 Future Work and Implications

Given that my results suggest that viral genotype is influencing viral load, the question arises as to the source of this effect in the viral genome. The analyses here have been performed on the *pol* gene, where both drug resistant and naturally occurring variation

is known to affect replicative capacity (Hinkley et al., 2011). It is also possible that the between-lineage variation observed (see Figure 3.2 and 4.3) could be a distal effect that maps to one or more genes, such as *env* (Ariën et al., 2005), which I am detecting through its linkage with variants in the *pol* gene. With increasing availability of full-genome datasets it may be possible to address this question more directly in future.

The DSPS's usefulness as an HIV epidemic simulator will be greatly increased once viral loads can be simulated under differing heritability values. Finding an equation that allow simulation of realistic viral load values without violating the assumptions of Brownian motion will allow a comprehensive test of the ability and accuracy of the heritability estimation pipeline presented here, and also open the possibility of using the DSPS to investigate how viral load changes over time, and in different populations.

In a test of seven different measures of phylogenetic signal, Shirreff et al. (2013) concluded that none could reliably detect a heritability signal below 40% in simulated data. The heritability estimation pipeline described in this thesis was able to consistently estimate a heritability signal of around 6% in the subtype B dataset (Chapter 6.1.1), suggesting it may perform significantly better at low heritability values than other phylogeny-based methods. Further testing using viral load data simulated with the DSPS, and perhaps future collaboration with George Shirreff to subject the heritability estimation pipeline to the same tests as the other methods, would better clarify the strengths and limitations of the heritability estimation method.

Finally, the new heritability estimation pipeline presented here has the potential to be applied to other organisms and populations where sequence data are available and could allow estimation of heritability where it has not previously been possible.

8

Bibliography

Abecasis, A. B., Vandamme, A.-M., and Lemey, P. (2009). Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution. *Journal of Virology*, 83(24):12917–12924.

Adams, A. L., Barth-Jones, D. C., Chick, S. E., and Koopman, J. S. (1998). Simulations to Evaluate HIV Vaccine Trial Designs. *SIMULATION*, 71(4):228–241.

Aghaizu, A., Brown, A., Nardone, A., Gill, O., Delpech, V., and contributors (2013). HIV in the United Kingdom 2013 Report: data to end 2012. Technical report, Public Health England, London.

Albert, J., Wahlberg, J., Leitner, T., Escanilla, D., and Uhlén, M. (1994). Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *Journal of Virology*, 68(9):5918–5924.

Alizon, S., von Wyl, V., Stadler, T., Kouyos, R. D., Yerly, S., Hirschel, B., Böni, J.,

Shah, C., Klimkait, T., Furrer, H., Rauch, A., Vernazza, P. L., Bernasconi, E., Battegay, M., Bürgisser, P., Telenti, A., Günthard, H. F., Bonhoeffer, S., and the Swiss HIV Cohort Study (2010). Phylogenetic Approach Reveals That Virus Genotype Largely Determines HIV Set-Point Viral Load. *PLoS Pathog*, 6(9):e1001123.

Altmann, M. (1995). Susceptible-infected-removed epidemic models with dynamic partnerships. *Journal of Mathematical Biology*, 33(6):661–675.

Anderson, R. M., May, R. M., Boily, M. C., Garnett, G. P., and Rowley, J. T. (1991). The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. *Nature*, 352(6336):581–589.

Anderson, R. M., May, R. M., and McLean, A. R. (1988). Possible demographic consequences of AIDS in developing countries. *Nature*, 332(6161):228–234.

Anderson, R. M., May, R. M., Ng, T. W., and Rowley, J. T. (1992). Age-Dependent Choice of Sexual Partners and the Transmission Dynamics of HIV in Sub-Saharan Africa. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 336(1277):135–155.

Anderson, R. M., Medley, G. F., May, R. M., and Johnson, A. M. (1986). A Preliminary Study of the Transmission Dynamics of the Human Immunodeficiency Virus (HIV), the Causative Agent of AIDS. *Mathematical Medicine and Biology*, 3(4):229–263.

Anderson, R. M., Ng, T. W., Boily, M. C., and May, R. M. (1989). The influence of different sexual-contact patterns between age classes on the predicted demographic impact of AIDS in developing countries. *Annals of the New York Academy of Sciences*, 569:240–274.

Arien, K. K., Abraha, A., Quinones-Mateu, M. E., Kestens, L., Vanham, G., and Arts, E. J. (2005). The Replicative Fitness of Primary Human Immunodeficiency Virus Type 1 (HIV-1) Group M, HIV-1 Group O, and HIV-2 Isolates. *J. Virol.*, 79(14):8979–8990.

Ariën, K. K., Troyer, R. M., Gali, Y., Colebunders, R. L., Arts, E. J., and Vanham,

G. (2005). Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time. *AIDS*, 19(15):1555–1564.

Arnold, C., Barlow, K. L., Parry, J. V., and Clewley, J. P. (1995). At Least Five HIV-1 Sequence Subtypes (A, B, C, D, A/E) Occur in England. *AIDS Research and Human Retroviruses*, 11(3):427–429.

Arts, E. J. and Hazuda, D. J. (2012). HIV-1 Antiretroviral Drug Therapy. *Cold Spring Harbor Perspectives in Medicine*, 2(4):a007161.

Åsjö, B., Albert, J., Karlsson, A., Morfeldt-Månson, L., Biberfeld, G., Lidman, K., and Fenyö, E. (1986). Replicative capacity of human immunodeficiency virus from patients with varying severity of HIV infection. *The Lancet*, 328(8508):660–662.

Balfe, P., Simmonds, P., Ludlam, C. A., Bishop, J. O., and Brown, A. J. (1990). 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. *Journal of Virology*, 64(12):6221–6233.

Bansode, V., Drebert, Z. J., Travers, S. A., Banda, E., Molesworth, A., Crampin, A., Ngwira, B., French, N., Glynn, J. R., and McCormack, G. P. (2011). Drug Resistance Mutations in Drug-Naive HIV Type 1 Subtype C-Infected Individuals from Rural Malawi. *AIDS Research and Human Retroviruses*, 27(4):439–444.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science (New York, N.Y.)*, 220(4599):868–871.

Bates, T. W., Thurmond, M. C., and Carpenter, T. E. (2003). Description of an epidemic simulation model for use in evaluating strategies to control an outbreak of foot-and-mouth disease. *American Journal of Veterinary Research*, 64(2):195–204.

Bennett, D. E., Camacho, R. J., Otelea, D., Kuritzkes, D. R., Fleury, H., Kiuchi, M., Heneine, W., Kantor, R., Jordan, M. R., Schapiro, J. M., Vandamme, A.-M., Sand-

strom, P., Boucher, C. A. B., van de Vijver, D., Rhee, S.-Y., Liu, T. F., Pillay, D., and Shafer, R. W. (2009). Drug Resistance Mutations for Surveillance of Transmitted HIV-1 Drug-Resistance: 2009 Update. *PLoS ONE*, 4(3):e4724.

Bielejec, F., Lemey, P., Carvalho, L. M., Baele, G., Rambaut, A., and Suchard, M. A. (2014). πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics*, 15(1):133.

Blomberg, S. P., Garland, T., and Ives, A. R. (2003). Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits arre More Labile. *Evolution*, 57(4):717–745.

Blower, S. M. and Dowlatabadi, H. (1994). Sensitivity and Uncertainty Analysis of Complex Models of Disease Transmission: An HIV Model, as an Example. *International Statistical Review / Revue Internationale de Statistique*, 62(2):229–243.

Blower, S. M., Gershengorn, H. B., and Grant, R. M. (2000). A Tale of Two Futures: HIV and Antiretroviral Therapy in San Francisco. *Science*, 287(5453):650–654.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287.

Boyd, A., Murad, S., O'Shea, S., De Ruiter, A., Watson, C., and Easterbrook, P. (2005). Ethnic differences in stage of presentation of adults newly diagnosed with HIV-1 infection in south London. *HIV Medicine*, 6(2):59–65.

Brenner, B. G., Roger, M., Routy, J.-P., Moisi, D., Ntemgwa, M., Matte, C., Baril, J.-G., Thomas, R., Rouleau, D., Bruneau, J., Leblanc, R., Legault, M., Tremblay, C., Charest, H., and Wainberg, M. A. (2007). High Rates of Forward Transmission Events after Acute/Early HIV-1 Infection. *Journal of Infectious Diseases*, 195(7):951–959.

Brown, A. E., Malone, J. D., Zhou, S. Y. J., Lane, J. R., and Hawkes, C. A. (1997). Human Immunodeficiency Virus RNA Levels in US Adults: A Comparison Based upon Race and Ethnicity. *Journal of Infectious Diseases*, 176(3):794–797.

Burns, F. M. a., Fakoya, A. O. b., Copas, A. J. c., and French, P. D. a. (2001). Africans in London continue to present with advanced HIV disease in the era of highly active antiretroviral therapy. [Letter]. *AIDS*, 15(18):2453–2455.

Cardozo, E. F., Vargas, C. A., and Zurakowski, R. (2012). A compartment based model for the formation of 2-LTR circles after raltegravir intensification. In *51st IEEE Conference on Decision and Control*, pages 4924–4929, Maui, Hawaii, USA.

Carré, N., Prins, M., Meyer, L., Brettle, R. P., Robertson, J. R., McArdle, H., Goldberg, D. J., Zangerle, R., Coutinho, R. A., and van den Hoek, A. (1997). Has the rate of progression to AIDS changed in recent years? *AIDS*, 11(13):1611–1618.

CASCADE Collaboration (2003). Differences in CD4 cell counts at seroconversion and decline among 5739 HIV-1-infected individuals with well-estimated dates of seroconversion. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 34(1):76–83.

CDC (1981). Kaposi's Sarcoma and Pneumocystis Pneumonia Among Homosexual Men – New York City and California. *MMWR Morb Mortal Wkly Rep*, 30:305–308.

Celum, C., Wald, A., Lingappa, J., Magaret, A., Wang, R., Mugo, N., Mujugira, A., Baeten, J., Mullins, J., Hughes, J., Bukusi, E., Cohen, C., Katabira, E., Ronald, A., Kiarie, J., Farquhar, C., Stewart, G., Makhema, J., Essex, M., Were, E., Fife, K., de Bruyn, G., Gray, G., McIntyre, J., Manongi, R., Kapiga, S., Coetzee, D., Allen, S., Inambao, M., Kayitenkore, K., Karita, E., Kanweka, W., Delany, S., Rees, H., Vwalika, B., Stevens, W., Campbell, M., Thomas, K., Coombs, R., Morrow, R., Whittington, W., McElrath, M., Barnes, L., Ridzon, R., and Corey, L. (2010). Acyclovir and Transmission of HIV-1 from Persons Infected with HIV-1 and HSV-2. *New England Journal of Medicine*, 362(5):427–439.

Chen, Z., Luckay, A., Sodora, D. L., Telfer, P., Reed, P., Gettie, A., Kanu, J. M., Sadek, R. F., Yee, J., Ho, D. D., Zhang, L., and Marx, P. A. (1997). Human Immunodeficiency Virus Type 2 (HIV-2) Seroprevalence and Characterization of a Distinct HIV-2 Genetic Subtype from the Natural Range of Simian Immunodeficiency Virus-Infected Sooty Mangabeys. *Journal of Virology*, 71(5):3953–3960.

Chun, T.-W., Nickle, D. C., Justement, J. S., Meyers, J. H., Roby, G., Hallahan, C. W., Kottilil, S., Moir, S., Mican, J. M., Mullins, J. I., Ward, D. J., A, K. J., Mannon, P. J., and Fauci, A. S. (2008). Persistence of HIV in Gut-Associated Lymphoid Tissue despite Long-Term Antiretroviral Therapy. *Journal of Infectious Diseases*, 197(5):714–720.

Chun, T.-W., Stuyver, L., Mizell, S. B., Ehler, L. A., Mican, J. A. M., Baseler, M., Lloyd, A. L., Nowak, M. A., and Fauci, A. S. (1997). Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences*, 94(24):13193–13197.

Clavel, F. and Hance, A. J. (2004). HIV Drug Resistance. *New England Journal of Medicine*, 350(10):1023–1035.

Clewley, J. P., Arnold, C., Barlow, K. L., Grant, P. R., and Parry, J. V. (1996). Diverse HIV-1 genetic subtypes in UK. *The Lancet*, 347(9013):1487.

Coffin, J., Haase, A., Levy, J. A., Montagnier, L., Oroszlan, S., Teich, N., Temin, H., Toyoshima, K., Varmus, H., and Vogt, P. (1986). What to call the AIDS virus? *Nature*, 321(6065):10.

Concerted Action on SeroConversion to AIDS and Death in Europe (2000). Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. *The Lancet*, 355(9210):1131–1137.

Cori, A., Ayles, H., Beyers, N., Schaap, A., Floyd, S., Sabapathy, K., Eaton, J. W., Hauck, K., Smith, P., Griffith, S., Moore, A., Donnell, D., Vermund, S. H., Fidler, S., Hayes, R., Fraser, C., and HPTN 071 (PopART) study team (2014). HPTN 071 (PopART): A Cluster-Randomized Trial of the Population Impact of an HIV Combination Prevention Intervention Including Universal Testing and Treatment: Mathematical Model. *PLoS ONE*, 9(1):e84511.

Crum-Cianflone, N., Eberly, L., Zhang, Y., Ganesan, A., Weintrob, A., Marconi, V., Barthel, R. V., Fraser, S., Agan, B. K., and Wegner, S. (2009). Is HIV Becoming More Virulent? Initial CD4 Cell Counts among HIV Seroconverters During the Course of

the HIV Epidemic: 1985-2007. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 48(9):1285–1292.

Daar, E. S., Moudgil, T., Meyer, R. D., and Ho, D. D. (1991). Transient High Levels of Viremia in Patients with Primary Human Immunodeficiency Virus Type 1 Infection. *New England Journal of Medicine*, 324(14):961–964.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. ArticleType: research-article / Full publication date: 1977 / Copyright © 1977 Royal Statistical Society.

Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Di Marzio, P., Marmon, S., Sutton, R. E., Hill, C. M., Davis, C. B., Peiper, S. C., Schall, T. J., Littman, D. R., and Landau, N. R. (1996). Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, 381(6584):661–666.

Dietz, K. and Hadeler, K. P. (1988). Epidemiological models for sexually transmitted diseases. *Journal of Mathematical Biology*, 26(1):1–25.

Dietz, K. and Tudor, D. (1992). Triangles in Heterosexual HIV Transmission. In Jewell, N. P., Dietz, K., and Farewell, V. T., editors, *AIDS Epidemiology*, pages 143–155. Birkhäuser Boston.

Doitsh, G., Cavrois, M., Lassen, K. G., Zepeda, O., Yang, Z., Santiago, M. L., Hebbeler, A. M., and Greene, W. C. (2010). Abortive HIV Infection Mediates CD4 T Cell Depletion and Inflammation in Human Lymphoid Tissue. *Cell*, 143(5):789–801.

Dorak, M. T., Tang, J., Penman-Aguilar, A., Westfall, A. O., Zulu, I., Lobashevsky, E. S., Kancheya, N. G., Schaen, M. M., Allen, S. A., and Kaslow, R. A. (2004). Transmission of HIV-1 and HLA-B allele-sharing within serodiscordant heterosexual Zambian couples. *The Lancet*, 363(9427):2137–2139.

Dorrucci, M., Phillips, A. N., Longo, B., Rezza, G., and The Italian Seroconversion Study (2005). Changes over time in post-seroconversion CD4 cell counts in the Italian HIV-Seroconversion Study: 1985-2002. *AIDS*, 19(3):331–335.

Dorrucci, M., Rezza, G., Porter, K., and Phillips, A. (2007). Temporal Trends in Postseroconversion CD4 Cell Count and HIV Load: The Concerted Action on Seroconversion to AIDS and Death in Europe Collaboration, 1985–2002. *The Journal of Infectious Diseases*, 195(4):525–534.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(1):214.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

du Bois, R., Branthwaite, M., Mikhail, J., and Batten, J. (1981). Primary Pneumocystis carinii and cytomegalovirus infections. *The Lancet*, 318(8259):1339.

Edwards, A. and Cavalli-Sforza, L. (1964). Reconstruction of evolutionary trees. In Heywood, W. and McNeill, J., editors, *Phenetic and phylogenetic classification*, pages 67–76. Systematics Association Publication No. 6, London.

Falconer, D. S. and Mackay, T. F. (1996). *Introduction to Quantitative Genetics*. Pearson Education Limited, Essex, England, 4 edition.

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., Tatem, A. J., Sousa, J. D., Arinaminpathy, N., Pépin, J., Posada, D., Peeters, M., Pybus, O. G., and Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61.

Farzadegan, H., Hoover, D. R., Astemborski, J., Lyles, C. M., Margolick, J. B., Markham, R. B., Quinn, T. C., and Vlahov, D. (1998). Sex differences in HIV-1 viral load and progression to AIDS. *The Lancet*, 352(9139):1510–1514.

Fauci, A. and Desrosiers, R. (1997). Pathogenesis of HIV and SIV. In Coffin, J., Hughes, S., and Varmus, H., editors, *Retroviruses*, pages 587–636. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Fellay, J., Ge, D., Shianna, K. V., Colombo, S., Ledergerber, B., Cirulli, E. T., Urban, T. J., Zhang, K., Gumbs, C. E., Smith, J. P., Castagna, A., Cozzi-Lepri, A., De Luca,

A., Easterbrook, P., Günthard, H. F., Mallal, S., Mussini, C., Dalmau, J., Martinez-Picado, J., Miro, J. M., Obel, N., Wolinsky, S. M., Martinson, J. J., Detels, R., Margolick, J. B., Jacobson, L. P., Descombes, P., Antonarakis, S. E., Beckmann, J. S., O'Brien, S. J., Letvin, N. L., McMichael, A. J., Haynes, B. F., Carrington, M., Feng, S., Telenti, A., Goldstein, D. B., and for HIV/AIDS Vaccine Immunology (CHAVI), N. C. (2009). Common Genetic Variation and the Control of HIV-1 in Humans. *PLoS Genet*, 5(12):e1000791.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., and Goldstein, D. B. (2007). A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science*, 317(5840):944–947.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology*, 27(4):401–410.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125(1):1–15.

Felsenstein, J. (2008). Comparative Methods with Sampling Error and Within-Species Variation: Contrasts Revisited and Revised. *The American Naturalist*, 171(6):713–725. ArticleType: research-article / Full publication date: June 2008 / Copyright © 2008 The University of Chicago Press.

Fenner, F. and Chapple, P. J. (1965). Evolutionary changes in myxoma virus in Britain: An examination of 222 in naturally occurring strains obtained from 80 counties during the period October–November 1962. *Journal of Hygiene*, 63(2):175–185.

Fenyo, E. M., Morfeldt-Manson, L., Chiodi, F., Lind, B., von Gegerfelt, A., Albert, J., Olausson, E., and Asjo, B. (1988). Distinct replicative and cytopathic characteristics of human immunodeficiency virus isolates. *J. Virol.*, 62(11):4414–4419.

Ferguson, N. M. and Garnett, G. P. (2000). More realistic models of sexually transmit-

ted disease transmission dynamics: sexual partnership networks, pair models, and moment closure. *Sexually Transmitted Diseases*, 27(10):600–609.

Fideli, U. S., Allen, S. A., Musonda, R., Trask, S., Hahn, B. H., Weiss, H., Mulenga, J., Kasolo, F., Vermund, S. H., and Aldrovandi, G. M. (2001). Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa. *AIDS Research and Human Retroviruses*, 17(10):901–910.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D., and Siliciano, R. F. (1997). Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science*, 278(5341):1295–1300.

Fiore, J. R., Calabró, M. L., Angarano, G., De Rossi, A., Fico, C., Pastore, G., and Bianchi, L. C. (1990). HIV-1 variability and progression to AIDS: A longitudinal study. *Journal of Medical Virology*, 32(4):252–256.

FitzJohn, R. G. (2010). Quantitative Traits and Diversification. *Systematic Biology*, 59(6):619–633.

FitzJohn, R. G. (2012). Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3(6):1084–1092.

Fraser, C. and Hollingsworth, T. D. (2010). Interpretation of correlations in setpoint viral load in transmitting couples. *AIDS*, 24(16):2596–2597.

Fraser, C., Hollingsworth, T. D., Chapman, R., de Wolf, F., and Hanage, W. P. (2007). Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis. *Proceedings of the National Academy of Sciences*, 104(44):17441–17446.

Fraser, C., Lythgoe, K., Leventhal, G. E., Shirreff, G., Hollingsworth, T. D., Alizon, S., and Bonhoeffer, S. (2014). Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective. *Science*, 343(6177):1243727.

Fryer, H. R., Frater, J., Duda, A., Roberts, M. G., Phillips, R. E., McLean, A. R., and

The SPARTAC Trial Investigators (2010). Modelling the Evolution and Spread of HIV Immune Escape Mutants. *PLoS Pathog*, 6(11):e1001196.

Galai, N., Lepri, A. C., Vlahov, D., Pezzotti, P., Sinicco, A., Rezza, G., and Study, H. I. V. I. S. (1996). Temporal Trends of Initial CD4 Cell Counts Following Human Immunodeficiency Virus Seroconversion in Italy, 1985–1992. *American Journal of Epidemiology*, 143(3):278–282.

Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., Palker, T. J., Redfield, R., Oleske, J., Safai, B., and Et, A. (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 224(4648):500–503.

Gandhi, M., Bacchetti, P., Miotti, P., Quinn, T. C., Veronese, F., and Greenblatt, R. M. (2002). Does Patient Sex Affect Human Immunodeficiency Virus Levels? *Clinical Infectious Diseases*, 35(3):313–322.

Gao, F., Yue, L., Robertson, D. L., Hill, S. C., Hui, H., Biggar, R. J., Neequaye, A. E., Whelan, T. M., Ho, D. D., and Shaw, G. M. (1994). Genetic Diversity of Human Immunodeficiency Virus Type 2: Evidence for Distinct Sequence Subtypes with Differences in Virus Biology. *Journal of Virology*, 68(11):7433–7447.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M., and Hahn, B. H. (1992). Human infection by genetically diverse SIVSM-related HIV-2 in West Africa. *Nature*, 358(6386):495–499.

García, F., Vidal, C., Gatell, J. M., Miró, J. M., Soriano, A., and Pumarola, T. (1997). Viral load in asymptomatic patients with CD4+ lymphocyte counts above 500 x 106/l. *AIDS*, 11(1).

Gardner, M. (1970). Mathematical Games: The fantastic combination of John Conway's new solitaire game 'life'. *Scientific American*, 223:120–123.

Garnett, G. P. and Anderson, R. M. (1996). Sexually Transmitted Diseases And Sexual Behavior: Insights From Mathematical Models. *Journal of Infectious Diseases*, 174(Supplement 2):S150–S161.

Gifford, R. J., Oliveira, T. d., Rambaut, A., Pybus, O. G., Dunn, D., Vandamme, A.-M., Kellam, P., and Pillay, D. (2007). Phylogenetic Surveillance of Viral Genetic Diversity and the Evolving Molecular Epidemiology of Human Immunodeficiency Virus Type 1. *Journal of Virology*, 81(23):13050–13056.

Gilbert, M. T. P., Rambaut, A., Wlasiuk, G., Spira, T. J., Pitchenik, A. E., and Worobey, M. (2007). The emergence of HIV/AIDS in the Americas and beyond. *Proceedings of the National Academy of Sciences*, 104(47):18566–18570.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361.

Gilmour, A., Gogel, B., Cullis, B., and Thompson, R. (2009). ASReml User Guide Release 3.0.

Gorham, E. D., Garland, F. C., Mayers, D. L., Goforth, R. R., Brodine, S. K., Weiss, P. J., McNally, M. S., and Group, N. R. W. (1993). CD4 Lymphocyte Counts Within 24 Months of Human Immunodeficiency Virus Seroconversion: Findings in the US Navy and Marine Corps. *Arch Intern Med*, 153(7):869–876.

Gottlieb, M. S., Schroff, R., Schanker, H. M., Weisman, J. D., Fan, P. T., Wolf, R. A., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *The New England Journal of Medicine*, 305(24):1425–1431.

Goulder, P. J. R., Brander, C., Tang, Y., Tremblay, C., Colbert, R. A., Addo, M. M., Rosenberg, E. S., Nguyen, T., Allen, R., Trocha, A., Altfeld, M., He, S., Bunce, M., Funkhouser, R., Pelton, S. I., Burchett, S. K., McIntosh, K., Korber, B. T. M., and Walker, B. D. (2001). Evolution and transmission of stable CTL escape mutations in HIV infection. *Nature*, 412(6844):334–338.

Grant, R. M., Hecht, F. M., Warmerdam, M., Liu, L., Liegler, T., Petropoulos, C. J., Hellmann, N. S., Chesney, M., Busch, M. P., and Kahn, J. O. (2002). Time trends in primary HIV-1 drug resistance among recently infected persons. *JAMA*, 288(2):181–188.

Grant, R. M., Kuritzkes, D. R., Johnson, V. A., Mellors, J. W., Sullivan, J. L., Swanstrom, R., D'Aquila, R. T., Gorder, M. V., Holodniy, M., Robert M. Lloyd, J., Reid, C., Morgan, G. F., and Winslow, D. L. (2003). Accuracy of the TRUGENE HIV-1 Genotyping Kit. *Journal of Clinical Microbiology*, 41(4):1586–1593.

Gray, R. H., Wawer, M. J., Brookmeyer, R., Sewankambo, N. K., Serwadda, D., Wabwire-Mangen, F., Lutalo, T., Li, X., vanCott, T., and Quinn, T. C. (2001). Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. *The Lancet*, 357(9263):1149–1153.

Greene, W. C. (2007). A history of AIDS: Looking back to see ahead. *European Journal of Immunology*, 37(S1):S94–S102.

Gribaldo, S. and Philippe, H. (2002). Ancient Phylogenetic Relationships. *Theoretical Population Biology*, 61(4):391–408.

Gupta, S., Anderson, R. M., and May, R. M. (1989). Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS*, 3(12).

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, 33(2):1–22.

Hadfield, J. D. and Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, 23(3):494–508.

Hadfield, J. D., Wilson, A. J., Garant, D., Sheldon, B. C., and Kruuk, L. E. B. (2010). The Misuse of BLUP in Ecology and Evolution. *The American Naturalist*, 175(1):116–125.

Hadjiandreou, M., Conejeros, R., and Vassiliadis, V. S. (2007). Towards a long-term

model construction for the dynamic simulation of HIV infection. *Mathematical biosciences and engineering: MBE*, 4(3):489–504.

Health Protection Agency (2011). HIV in the United Kingdom: 2011 Report. Technical report, Health Protection Services, Colindale, London.

Health Protection Agency, SCIEH, ISD, National Public Health Service for Wales, and CDSC Northern Ireland and the UASSG (2003). Renewing the Focus: HIV and other Sexually Transmitted Infections in the United Kingdom in 2002. Technical report, Health Protection Agency, London.

Hecht, F. M., Hartogensis, W., Bragg, L., Bacchetti, P., Atchison, R., Grant, R., Barbour, J., and Deeks, S. G. (2010). HIV RNA level in early infection is predicted by viral load in the transmission source. *AIDS*, 24(7):941–945.

Hemelaar, J., Gouws, E., Ghys, P. D., and Osmanov, S. (2006). Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS (London, England)*, 20(16):W13–23.

Henderson, C. R. (1950). Estimation of Genetic Parameters. *The Annals of Mathematical Statistics*, 21(2):309.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. *Biometrics*, 9(2):226–252.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, pages 10–41.

Henderson, C. R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics*, 32(1):69–83.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2):192–218.

Herbeck, J. T., Mittler, J. E., Gottlieb, G. S., and Mullins, J. I. (2014). An HIV

Epidemic Model Based on Viral Load Dynamics: Value in Assessing Empirical Trends in HIV Virulence and Community Viral Load. *PLoS Comput Biol*, 10(6):e1003673.

Herbeck, J. T., Müller, V., Maust, B. S., Ledergerber, B., Torti, C., Di Giambenedetto, S., Gras, L., Günthard, H. F., Jacobson, L. P., Mullins, J. I., and Gottlieb, G. S. (2012). Is the virulence of HIV changing? A meta-analysis of trends in prognostic markers of HIV disease progression and transmission. *AIDS*, 26(2):193–205.

Hernandez-Vargas, E. A. and Middleton, R. H. (2013). Modeling the three stages in HIV infection. *Journal of Theoretical Biology*, 320:33–40.

Hethcote, H. W. and Yorke, J. A. (1984). Gonorrhea Transmission Dynamics and Control. *Lecture Notes in Biomathematics*, 56.

Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. (2011). A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature Genetics*, 43(5):487–489.

Hogg, R. S., Bangsberg, D. R., Lima, V. D., Alexander, C., Bonner, S., Yip, B., Wood, E., Dong, W. W. Y., Montaner, J. S. G., and Harrigan, P. R. (2006). Emergence of Drug Resistance Is Associated with an Increased Risk of Death among Patients First Starting HAART. *PLoS Med*, 3(9):e356.

Hollingsworth, T., Laeyendecker, O., Shirreff, G., Donnelly, C. A., Serwadda, D., Wawer, M. J., Kiwanuka, N., Nalugoda, F., Collinson-Streng, A., Ssempijja, V., Hanage, W. P., Quinn, T. C., Gray, R. H., and Fraser, C. (2010). HIV-1 Transmitting Couples Have Similar Viral Load Set-Points in Rakai, Uganda. *PLoS Pathog*, 6(5):e1000876.

Hollingsworth, T. D., Anderson, R. M., and Fraser, C. (2008). HIV-1 Transmission, by Stage of Infection. *Journal of Infectious Diseases*, 198(5):687–693.

Holmberg, S. D., Conley, L. J., Luby, S. P., Cohn, S., Wong, L. C., and Vlahov, D. (1995). Recent Infection with Human Immunodeficiency Virus and Possible Rapid Loss of CD4 T Lymphocytes. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 9(3):291–296.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Rogers, A. S., and Brown, A. J. L. (1993). Molecular Investigation of Human Immunodeficiency Virus (HIV) Infection in a Patient of an HIV-Infected Surgeon. *Journal of Infectious Diseases*, 167(6):1411–1414.

Housworth, E. A., Martins, E. P., and Lynch, M. (2004). The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96.

Huang, Y., Paxton, W. A., Wolinsky, S. M., Neumann, A. U., Zhang, L., He, T., Kang, S., Ceradini, D., Jin, Z., Yazdanbakhsh, K., Kunstman, K., Erickson, D., Dragon, E., Landau, N. R., Phair, J., Ho, D. D., and Koup, R. A. (1996). The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med*, 2(11):1240–1243.

Hué, S., Clewley, J. P., Cane, P. A., and Pillay, D. (2004). HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS (London, England)*, 18(5):719–728.

Hué, S., Pillay, D., Clewley, J. P., and Pybus, O. G. (2005). Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4425–4429.

Hutchinson, C. M., Wilson, C., Reichart, C. A., Marsiglia, V. C., Zenilman, J. M., and Hook, E. W. (1991). CD4 Lymphocyte Concentrations in Patients With Newly Identified HIV Infection Attending STD Clinics: Potential Impact on Publicly Funded Health Care Resources. *JAMA*, 266(2):253–256.

Im, S., Fernando, R., and Gianola, D. (1989). Likelihood inferences in animal breeding under selection: a missing-data theory view point. *Genetics, Selection, Evolution : GSE*, 21(4):399–414.

Jacquez, J. A., Simon, C. P., Koopman, J., Sattenspiel, L., and Perry, T. (1988). Modeling and analyzing HIV transmission: the effect of contact patterns. *Mathematical Biosciences*, 92(2):119–199.

Jaffe, H. W., Darrow, W. W., Echenberg, D. F., O'Malley, P. M., Getchell, J. P., Kalyanaraman, V. S., Byers, R. H., Drennan, D. P., Braff, E. H., Curran, J. W., and Francis, D. P. (1985). The Acquired Immunodeficiency Syndrome in a Cohort of Homosexual MenA Six-Year Follow-up Study. *Annals of Internal Medicine*, 103(2):210–214.

Jalvingh, A. W., Nielen, M., Maurice, H., Stegeman, A. J., Elbers, A. R. W., and Dijkhuizen, A. A. (1999). Spatial and stochastic simulation to evaluate the impact of events and control measures on the 1997–1998 classical swine fever epidemic in The Netherlands.: I. Description of simulation model. *Preventive Veterinary Medicine*, 42(3–4):271–295.

Kaleebu, P., Nankya, I. L., Yirrell, D. L., Shafer, L. A., Kyosiimire-Lugemwa, J., Lule, D. B., Morgan, D., Beddows, S., Weber, J., and Whitworth, J. A. G. (2007). Relation between chemokine receptor use, disease stage, and HIV-1 subtypes A and D: results from a rural Ugandan cohort. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 45(1):28–33.

Kaleebu, P., Ross, A., Morgan, D., Yirrell, D., Oram, J., Rutebemberwa, A., Lyagoba, F., Hamilton, L., Biryahwaho, B., and Whitworth, J. (2001). Relationship between HIV-1 Env subtypes A and D and disease progression in a rural Ugandan cohort. *AIDS (London, England)*, 15(3):293–299.

Kanki, P. J., Hamel, D. J., Sankalé, J.-L., Hsieh, C.-C., Thior, I., Barin, F., Woodcock, S. A., Guèye-Ndiaye, A., Zhang, E., Montano, M., Siby, T., Marlink, R., NDoye, I., Essex, M. E., and MBoup, S. (1999). Human Immunodeficiency Virus Type 1 Subtypes Differ in Disease Progression. *Journal of Infectious Diseases*, 179(1):68–73.

Kaslow, R. A., vanRaden, M., Friedman, H., Duquesnoy, R., Marrari, M., Kingsley, L., Rinaldo, C. R., Su, S., Saah, A., Detels, R., and Phair, J. (1990). A1, Cw7, B8, DR3 HLA antigen combination associated with rapid decline of T-helper lymphocytes in HIV-1 infection : A report from the Multicenter AIDS Cohort Study. *The Lancet*, 335(8695):927–930.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press.

Keet, I. P., Veugelers, P. J., Koot, M., de Weerd, M. H., Roos, M. T., Miedema, F., de Wolf, D. F., Goudsmit, J., and Coutinho, R. A. (1996). Temporal trends of the natural history of HIV-1 infection following seroconversion between 1984 and 1993. *AIDS*, 10(13):1601–1602.

Kermack, W. O. and McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721.

Kiwanuka, N., Laeyendecker, O., Robb, M., Kigozi, G., Arroyo, M., McCutchan, F., Eller, L. A., Eller, M., Makumbi, F., Birx, D., Wabwire-Mangen, F., Serwadda, D., Sewankambo, N. K., Quinn, T. C., Wawer, M., and Gray, R. (2008). Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. *The Journal of Infectious Diseases*, 197(5):707–713.

Klatt, N. R., Silvestri, G., and Hirsch, V. (2012). Nonpathogenic Simian Immunodeficiency Virus Infections. *Cold Spring Harbor Perspectives in Medicine*, 2(1):a007153.

Klatzmann, D., Barre-Sinoussi, F., Nugeyre, M. T., Danquet, C., Vilmer, E., Griscelli, C., Brun-Veziret, F., Rouzioux, C., Gluckman, J. C., Chermann, J. C., and Et, A. (1984). Selective tropism of lymphadenopathy associated virus (LAV) for helper-inducer T lymphocytes. *Science*, 225(4657):59–63.

Knorn, S. and Middleton, R. (2014). Lymph compartment models and HIV intra patient infection dynamics. In *2014 IEEE Conference on Control Applications (CCA)*, pages 1699–1704.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the Ancestor of the HIV-1 Pandemic Strains. *Science*, 288(5472):1789–1796.

Kouri, V., Khouri, R., Alemán?, Y., Abrahantes, Y., Vercauteren, J., Pineda-Peña, A.-C., Theys, K., Megens, S., Moutschen, M., Pfeifer, N., Van Weyenbergh, J., Pérez, A. B., Pérez, J., Pérez, L., Van Laethem, K., and Vandamme, A.-M. (2015). CRF19_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba. *EBioMedicine*.

Kozal, M. J., Hullsiek, K. H., MacArthur, R. D., Berg-Wolf, M. v. d., Peng, G., Xiang, Y., Baxter, J. D., Uy, J., Telzak, E. E., and Novak, R. M. (2007). The Incidence of HIV Drug Resistance and Its Impact on Progression of HIV Disease Among Antiretroviral-Naïve Participants Started on Three Different Antiretroviral Therapy Strategies. *HIV Clinical Trials*, 8(6):357–370.

Kretzschmar, M., Duynhoven, Y. T. H. P. v., and Severijnen, A. J. (1996). Modeling Prevention Strategies for Gonorrhea and Chlamydia Using Stochastic Network Simulations. *American Journal of Epidemiology*, 144(3):306–317.

Kuritzkes, D. R., Grant, R. M., Feorino, P., Griswold, M., Hoover, M., Young, R., Day, S., Lloyd, Jr., R. M., Reid, C., Morgan, G. F., and Winslow, D. L. (2003). Performance Characteristics of the TRUGENE HIV-1 Genotyping Kit and the Opengene DNA Sequencing System. *Journal of Clinical Microbiology*, 41(4):1594–1599.

Kuyu, C. (2008). Les Haïtiens au Congo. *Cahiers d'études africaines*, 48(192):895.

Lackner, A. A., Lederman, M. M., and Rodriguez, B. (2012). HIV pathogenesis: the host. *Cold Spring Harbor Perspectives in Medicine*, 2(9):a007005.

Langford, S. E., Ananworanich, J., and Cooper, D. A. (2007). Predictors of disease progression in HIV infection: a review. *AIDS research and therapy*, 4:11.

Leigh Brown, A. J., Lobidel, D., Wade, C. M., Rebus, S., Phillips, A. N., Brettle, R. P., France, A. J., Leen, C. S., McMenamin, J., McMillan, A., Maw, R. D., Mulcahy, F., Robertson, J. R., Sankar, K. N., Scott, G., Wyld, R., and Peutherer, J. F. (1997). The Molecular Epidemiology of Human Immunodeficiency Virus Type 1 in Six Cities in Britain and Ireland. *Virology*, 235(1):166–177.

Leigh Brown, A. J., Lycett, S. J., Weinert, L., Hughes, G. J., Fearnhill, E., and Dunn, D. T. (2011). Transmission Network Parameters Estimated From HIV Sequences for a Nationwide Epidemic. *Journal of Infectious Diseases*, 204(9):1463–1469.

Lingappa, J. R., Thomas, K. K., Hughes, J. P., Baeten, J. M., Wald, A., Farquhar, C., de Bruyn, G., Fife, K. H., Campbell, M. S., Kapiga, S., Mullins, J. I., and Celum, C. (2013). Partner Characteristics Predicting HIV-1 Set Point in Sexually Acquired HIV-1 Among African Seroconverters. *AIDS Research and Human Retroviruses*, 29(1):164–171.

Little, S. J. (2001). Is transmitted drug resistance in HIV on the rise? *BMJ : British Medical Journal*, 322(7294):1074–1075.

Liu, K., Linder, C. R., and Warnow, T. (2011). RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. *PLoS ONE*, 6(11).

Liu, T. F. and Shafer, R. W. (2006). Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 42(11):1608–1618.

Lockett, S. F., Alonso, A., Wyld, R., Martin, M. P., Robertson, J. R., Gore, S. M., Leen, C. L., Brettle, R. P., Yirrell, D. L., Carrington, M., and Brown, A. J. (1999). Effect of chemokine receptor mutations on heterosexual human immunodeficiency virus transmission. *The Journal of Infectious Diseases*, 180(3):614–621.

Lockett, S. F., Robertson, J. R., Brettle, R. P., Yap, P. L., Middleton, D., and Leigh Brown, A. J. (2001). Mismatched Human Leukocyte Antigen Alleles Protect Against Heterosexual HIV Transmission. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 27(3):277–280.

Lycett, S. J., Hodcroft, E., Leigh Brown, A. J., and Kao, R. R. (2015). Phylodynamics scenario simulation using the DiscreteSpatialPhyloSimualtor. *In Prep.*

Lyles, C. M., Dorrucci, M., Vlahov, D., Pezzotti, P., Angarano, G., Sinicco, A., Alberici, F., Alcorn, T. M., Vella, S., and Rezza, G. (1999). Longitudinal Human Immunod-

eficiency Virus Type 1 Load in the Italian Seroconversion Study: Correlates and Temporal Trends of Virus Load. *Journal of Infectious Diseases*, 180(4):1018–1024.

Lyles, R. H., Muñoz, A., Yamashita, T. E., Bazmi, H., Detels, R., Rinaldo, C. R., Margolick, J. B., Phair, J. P., and Mellors, J. W. (2000). Natural History of Human Immunodeficiency Virus Type 1 Viremia After Seroconversion and Proximal to AIDS in a Large Cohort of Homosexual Men. *Journal of Infectious Diseases*, 181(3):872–880.

Lynch, M. (1991). Methods for the Analysis of Comparative Data in Evolutionary Biology. *Evolution*, 45(5):1065–1080.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, Massachusetts.

Mangen, M. J. J., Nielen, M., and Burrell, A. M. (2002). Simulated effect of pig-population density on epidemic size and choice of control strategy for classical swine fever epidemics in The Netherlands. *Preventive Veterinary Medicine*, 56(2):141–163.

Mansky, L. M. and Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of Virology*, 69(8):5087–5094.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(2 Part 1):209–220.

Marlink, R., Kanki, P., Thior, I., Travers, K., Eisen, G., Siby, T., Traore, I., Hsieh, C. C., Dia, M. C., and Gueye, E. H. (1994). Reduced rate of disease development after HIV-2 infection as compared to HIV-1. *Science (New York, N.Y.)*, 265(5178):1587–1590.

Martinson, J. J., Chapman, N. H., Rees, D. C., Liu, Y.-T., and Clegg, J. B. (1997). Global distribution of the CCR5 gene 32-basepair deletion. *Nat Genet*, 16(1):100–103.

May, R. M. and Anderson, R. M. (1987). Transmission dynamics of HIV infection. *Nature*, 326(6109):137–142.

May, R. M. and Anderson, R. M. (1990). Parasite—host coevolution. *Parasitology*, 100(Supplement S1):S89–S101.

McLean, A. R. (1993). The balance of power between HIV and the immune system. *Trends in Microbiology*, 1(1):9–13.

McLean, A. R., Emery, V. C., Webster, A., and Griffiths, P. D. (1991). Population dynamics of HIV within an individual after treatment with zidovudine. *AIDS*, 5(5).

McLean, A. R. and Nowak, M. A. (1992). Models of interactions between HIV and other pathogens. *Journal of Theoretical Biology*, 155(1):69–86.

Mei, Y., Wang, L., and Holte, S. E. (2008). A comparison of methods for determining HIV viral set point. *Statistics in Medicine*, 27(1):121–139.

Mellors, J. W., Kingsley, L. A., Rinaldo, C. R., Todd, J. A., Hoo, B. S., Kokka, R. P., and Gupta, P. (1995). Quantitation of HIV-1 RNA in Plasma Predicts Outcome after Seroconversion. *Annals of Internal Medicine*, 122(8):573 –579.

Mellors, J. W., Margolick, J., Phair, J., Rinaldo, C. R., Detels, R., Jacobson, L. P., and Munoz, A. (2007). Prognostic Value of HIV-1 RNA, CD4 Cell Count, and CD4 Cell Count Slope for Progression to AIDS and Death in Untreated HIV-1 Infection. *The Journal of the American Medical Association*, 297(21):2346–2350.

Mellors, J. W., Rinaldo, C. R., Gupta, P., White, R. M., Todd, J. A., and Kingsley, L. A. (1996). Prognosis in HIV-1 Infection Predicted by the Quantity of Virus in Plasma. *Science*, 272(5265):1167–1170.

Miller, V., Vandamme, A. M., Loveday, C., Staszewski, S., Lundgren, J., Youle, M., Ait-Khaled, M., Boucher, C., Brun-Vezinet, F., Dedes, N., Giaquinto, C., Hertogs, K., Houyez, F., Perrin, L., Pillay, D., Schmit, J. C., Schuurman, R., Lange, J., Banhegyi, D., Biondi, G., Broekhuizen, A., Bush-Donovan, C., Camacho, R., Carlier, H., Clavel, F., Clotet, B., Clumeck, N., Colebunders, R., De Clerq, K., De Jaegher, J. J., De Schrijver, G., De Smet, K., Hall, W., Harrigan, R., Hatzakis, A., Hellmann, N., Hoetelmans, R., Holtzer, C., Katlama, C., Larder, D., Loriaux, E., McCreedy, B., Mulcahy, F., Opravil, M., Phillips, A., Ruiz, N., Shulse, E., Sonnerborg, A., Soriano,

V., Steel, H., Vella, S., Williams, A., and Resistance, E. G. H. (2001). Clinical and laboratory guidelines for the use of HIV-1 drug resistance testing as part of treatment management: recommendations for the European setting. *AIDS*, 15(3):309–320.

Morris, M. and Kretzschmar, M. (1997). Concurrent partnerships and the spread of HIV. *AIDS*, 11(5).

Müller, V., Fraser, C., and Herbeck, J. T. (2011). A Strong Case for Viral Genetic Factors in HIV Virulence. *Viruses*, 3(3):204–216.

Müller, V., Ledergerber, B., Perrin, L., Klimkait, T., Furrer, H., Telenti, A., Bernasconi, E., Vernazza, P., Günthard, H. F., Bonhoeffer, S., and the Swiss HIV Cohort Study (2006). Stable virulence levels in the HIV epidemic of Switzerland over two decades. *AIDS*, 20(6):889–894.

Müller, V., Maggiolo, F., Suter, F., Ladisa, N., De Luca, A., Antinori, A., Sighinolfi, L., Quiros-Roldan, E., Carosi, G., and Torti, C. (2009a). Increasing Clinical Virulence in Two Decades of the Italian HIV Epidemic. *PLoS Pathog*, 5(5):e1000454.

Müller, V., von Wyl, V., Yerly, S., Böni, J., Klimkait, T., Bürgisser, P., Ledergerber, B., Günthard, H. F., Bonhoeffer, S., and Swiss HIV Cohort Study (2009b). African descent is associated with slower CD4 cell count decline in treatment-naive patients of the Swiss HIV Cohort Study. *AIDS (London, England)*, 23(10):1269–1276.

Nabel, G. and Baltimore, D. (1987). An inducible transcription factor activates expression of human immunodeficiency virus in T cells. *Nature*, 326(6114):711–713.

Nagelkerke, N. J. D., Jha, P., de Vlas, S. J., Korenromp, E. L., Moses, S., Blanchard, J. F., and Plummer, F. A. (2002). Modelling HIV/AIDS epidemics in Botswana and India: impact of interventions to prevent transmission. *Bulletin of the World Health Organization*, 80(2):89–96.

Nogueras, M., Navarro, G., Antón, E., Sala, M., Cervantes, M., Amengual, M., and Segura, F. (2006). Epidemiological and clinical features, response to HAART, and survival in HIV-infected patients diagnosed at the age of 50 or more. *BMC Infectious Diseases*, 6:159.

O'Brien, S. J. and Nelson, G. W. (2004). Human genes that limit AIDS. *Nat Genet*, 36(6):565–574.

O'Brien, T. R., Blattner, W. A., Waters, D., Eyster, M., Hilgartner, M., Cohen, A., Luban, N., Hatzakis, A., Aledort, L., Rosenberg, P. S., Miley, W. J., Kroner, B. L., and Goedert, J. J. (1996). Serum HIV-1 RNA Levels and Time to Development of AIDS in the Multicenter Hemophilia Cohort Study. *Journal of the American Medical Association*, 276(2):105–110.

O'Brien, T. R., Hoover, D. R., Rosenberg, P. S., Chen, B., Detels, R., Kingsley, L. A., Phair, J., and Saah, A. J. (1995). Evalution of Secular Trends in CD4+ Lymphocyte Loss among Human Immunodeficiency Virus Type 1 (HIV-1)-infected Men with Known Dates of Seroconversion. *American Journal of Epidemiology*, 142(6):636–642.

Ou, C.-Y., Ciesielski, C. A., Myers, G., Bandea, C. I., Luo, C.-C., Korber, B. T. M., Mullins, J. I., Schochetman, G., Berkelman, R. L., Economou, A. N., Witte, J. J., Furman, L. J., Satten, G. A., Maclnnes, K. A., Curran, J. W., and Jaffe, H. W. (1992). Molecular Epidemiology of HIV Transmission in a Dental Practice. *Science*, 256(5060):1165–1171.

Owen, D. (2009). African migration to the UK. In *In the Context of Homeland Development from the Perspectives of Euro-African Relations and Latin American Experience*, University of Warwick, Coventry, UK.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Palm, A. A., Esbjörnsson, J., Månsson, F., Kvist, A., Isberg, P.-E., Biague, A., Silva, Z. J. d., Jansson, M., Norrgren, H., and Medstrand, P. (2013). Faster Progression to AIDS and AIDS-Related Death Among Seroincident Individuals Infected With Recombinant HIV-1 A3/CRF02_ag Compared With Sub-subtype A3. *Journal of Infectious Diseases*, page jit416.

Palmer, D., Frater, J., Phillips, R., McLean, A. R., and McVean, G. (2013). Integrating genealogical and dynamical modelling to infer escape and reversion rates in

HIV epitopes. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1762):20130696.

Pantazis, N., Porter, K., Costagliola, D., De Luca, A., Ghosn, J., Guiguet, M., Johnson, A. M., Kelleher, A. D., Morrison, C., Thiebaut, R., Wittkop, L., and Touloumi, G. (2014). Temporal trends in prognostic markers of HIV-1 virulence and transmissibility: an observational cohort study. *The Lancet HIV*, 1(3):e119–e126.

Pao, D., Fisher, M., Hué, S., Dean, G., Murphy, G., Cane, P. A., Sabin, C. A., and Pillay, D. (2005). Transmission of HIV-1 during primary infection: relationship to sexual risk and sexually transmitted infections. *AIDS (London, England)*, 19(1):85–90.

Parry, J. V., Murphy, G., Barlow, K. L., Lewis, K., Rogers, P. A., Belda, F. J., Nicoll, A., McGarrigle, C., Cliffe, S., Mortimer, P. P., and Clewley, J. P. (2001). National surveillance of HIV-1 subtypes for England and Wales: design, methods, and initial findings. *Journal of Acquired Immune Deficiency Syndromes (1999)*, 26(4):381–388.

Parsons, V., Phillips, A., Gilson, R., Fidler, S., Fisher, M., Johnson, A., Hawkins, D., McLean, K., Johnson, M., Porter, K., and UK Reigster of HIV Seroconverters (2014). Increase in HIV Plasma Viral Load Set-Point Among UK MSM. In *Conference on Retroviruses and Opportunistic Infections*, Boston, Massachusetts. Session P-W7, Poster 1015. http://www.croiconference.org/sessions/increase-hiv-plasma-viral-load-set-point-among-uk-msm.

Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information When Block Sizes Are Unequal. *Biometrika*, 58(3):545–554.

Pavoine, S., Ollier, S., Pontier, D., and Chessel, D. (2008). Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities. *Theoretical Population Biology*, 73(1):79–91.

Payne, R., Muenchhoff, M., Mann, J., Roberts, H. E., Matthews, P., Adland, E., Hempenstall, A., Huang, K.-H., Brockman, M., Brumme, Z., Sinclair, M., Miura, T., Frater, J., Essex, M., Shapiro, R., Walker, B. D., Ndung'u, T., McLean, A. R.,

Carlson, J. M., and Goulder, P. J. R. (2014). Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proceedings of the National Academy of Sciences*, 111(50):E5393–E5400.

Pepin, J., Morgan, G., Gevao, D. D. S., Mendy, M., Gaye, I., Scollen, N., Tedder, R., and Whittle, H. (1991). HIV-2-induced immunosuppression among asymptomatic West African prostitutes: evidence that HIV-2 is pathogenic, but less so than HIV-1. *AIDS*, 5(10).

Peutherer, J. F., Edmonds, E., Simonds, P., Dickson, J., and Bath, G. (1985). HTVL-III antibody in Edinburgh drug addicts. *Lancet*, 2(8464):1129–1130.

Phillips, A. N., Pillay, D., Miners, A. H., Bennett, D. E., Gilks, C. F., and Lundgren, J. D. (2008). Outcomes from monitoring of patients on antiretroviral therapy in resource-limited settings with viral load, CD4 cell count, or clinical observation alone: a computer simulation model. *Lancet*, 371(9622):1443–1451.

Phillips, A. N., Sabin, C., Pillay, D., and Lundgren, J. D. (2007). HIV in the UK 1980-2006: reconstruction using a model of HIV infection and the effect of antiretroviral therapy. *HIV medicine*, 8(8):536–546.

Plantier, J.-C., Leoz, M., Dickerson, J. E., De Oliveira, F., Cordonnier, F., Lemée, V., Damond, F., Robertson, D. L., and Simon, F. (2009). A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, 15(8):871–872.

Potter, S. J., Lemey, P., Achaz, G., Chew, C. B., Vandamme, A.-M., Dwyer, D. E., and Saksena, N. K. (2004). HIV-1 compartmentalization in diverse leukocyte populations during antiretroviral therapy. *Journal of Leukocyte Biology*, 76(3):562–570.

Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988). Fidelity of HIV-1 reverse transcriptase. *Science*, 242(4882):1168–1171.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7):1641 –1650.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490.

Quinn, T. C., Wawer, M. J., Sewankambo, N., Serwadda, D., Li, C., Wabwire-Mangen, F., Meehan, M. O., Lutalo, T., and Gray, R. H. (2000). Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1. *New England Journal of Medicine*, 342(13):921–929.

R Development Core Team (2011). R: A language and environment for statistical computing.

Raboud, J. M., Montaner, J. S. G., Conway, B., Haley, L., Sherlock, C., O'Shaughnessy, M. V., and Schechter, M. T. (1996). Variation in Plasma RNA Levels, CD4 Cell Counts, and p24 Antigen Levels in Clinically Stable Men with Human Immunodeficiency Virus Infection. *Journal of Infectious Diseases*, 174(1):191 –194.

Ragonnet-Cronin, M. (2014). Personal Communications.

Ragonnet-Cronin, M., Lycett, S. J., Hodcroft, E., Hué, S., Fearnhill, E., Dunn, D., Delpech, V., Leigh Brown, A. J., and on HIV Drug Resistance, U. C. G. (2013). Dynamics of Non-B HIV Transmission in the UK. In *Session 490*, Atlanta, GA.

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31(1):298–303.

Richardson, B. A., Mbori-Ngacha, D., Lavreys, L., John-Stewart, G. C., Nduati, R., Panteleeff, D. D., Emery, S., Kreiss, J. K., and Overbaugh, J. (2003). Comparison of Human Immunodeficiency Virus Type 1 Viral Loads in Kenyan Women, Men, and Infants During Primary and Early Infection. *Journal of Virology*, 77(12):7120–7123.

Robertson, A. (1978). The time of detection of recessive visible genes in small populations. *Genetics Research*, 31(03):255–264.

Robertson, J. R., Bucknall, A. B., Welsby, P. D., Roberts, J. J., Inglis, J. M., Peutherer, J. F., and Brettle, R. P. (1986). Epidemic of AIDS related virus (HTLV-III/LAV) infection among intravenous drug abusers. *BMJ*, 292(6519):527–529.

Rousseau, C. M., Learn, G. H., Bhattacharya, T., Nickle, D. C., Heckerman, D., Chetty, S., Brander, C., Goulder, P. J. R., Walker, B. D., Kiepiela, P., Korber, B. T., and Mullins, J. I. (2007). Extensive Intrasubtype Recombination in South African Human Immunodeficiency Virus Type 1 Subtype C Infections. *Journal of Virology*, 81(9):4492–4500.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Sabin, C. A., Devereux, H., Phillips, A. N., Hill, A., Janossy, G., Lee, C. A., and Loveday, C. (2000). Course of Viral Load Throughout HIV-1 Infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 23(2).

Salgado, M., Brennan, T., O'Connell, K., Bailey, J., Ray, S., Siliciano, R., and Blankson, J. (2010). Evolution of the HIV-1 nef gene in HLA-B*57 Positive Elite Suppressors. *Retrovirology*, 7:94.

Saul, J., Erwin, J., Bruce, J., and Peters, B. (2000). Ethnic and demographic variations in HIV/AIDS presentation at two London referral centres 1995–9. *Sexually Transmitted Infections*, 76(3):215.

Shafer, R. W. (2006). Rationale and Uses of a Public HIV Drug-Resistance Database. *The Journal of Infectious Diseases*, 194(s1):S51–S58.

Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., and Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1410):867 –876.

Shirreff, G., Alizon, S., Cori, A., Günthard, H. F., Laeyendecker, O., van Sighem, A., Bezemer, D., and Fraser, C. (2013). How effectively can HIV phylogenies be used to measure heritability? *Evolution, Medicine, and Public Health*, 2013(1):209–224.

Shirreff, G., Pellis, L., Laeyendecker, O., and Fraser, C. (2011). Transmission Selects for HIV-1 Strains of Intermediate Virulence: A Modelling Approach. *PLoS Comput Biol*, 7(10).

Siekevitz, M., Josephs, S. F., Dukovich, M., Peffer, N., Wong-Staal, F., and Greene, W. C. (1987). Activation of the HIV-1 LTR by T cell mitogens and the trans-activator protein of HTLV-I. *Science (New York, N.Y.)*, 238(4833):1575–1578.

Simmonds, P., Balfe, P., Ludlam, C. A., Bishop, J. O., and Brown, A. J. (1990). Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *Journal of Virology*, 64(12):5840–5850.

Sinicco, A., Fora, R., Raiteri, R., Sciandra, M., Bechis, G., Calvo, M. M., and Gioannini, P. (1997). Is the clinical course of HIV-1 changing? Cohort study. *BMJ*, 314(7089):1232–1237.

Sinka, K., Mortimer, J., Evans, B., and Morgan, D. (2003). Impact of the HIV epidemic in sub-Saharan Africa on the pattern of HIV in the UK. *AIDS (London, England)*, 17(11):1683–1690.

Smith, M. W., Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Lomb, D. A., Goedert, J. J., O'Brien, T. R., Jacobson, L. P., Kaslow, R., Buchbinder, S., Vittinghoff, E., Vlahov, D., Hoots, K., Hilgartner, M. W., (HGDS), H. G. a. D. S., Study, M. A. C. S. M., and O'Brien, S. J. (1997). Contrasting Genetic Influence of CCR2 and CCR5 Variants on HIV-1 Infection and Disease Progression. *Science*, 277(5328):959–965.

Sorensen, D. A., Wang, C. S., Jensen, J., and Gianola, D. (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics Selection Evolution*, 26(4):1–28.

Spira, S., Wainberg, M. A., Loemba, H., Turner, D., and Brenner, B. G. (2003). Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. *Journal of Antimicrobial Chemotherapy*, 51(2):229–240.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics*, 22(21):2688–2690.

Stamatakis, A., Blagojevic, F., Nikolopoulos, D., and Antonopoulos, C. (2007). Exploring New Search Algorithms and Hardware for Phylogenetics: RAxML Meets the IBM Cell. *The Journal of VLSI Signal Processing*, 48(3):271–286.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology*, 57(5):758–771.

Steel, C. M., Beatson, D., Cuthbert, R. J. G., Morrison, H., Ludlam, C. A., Peutherer, J. F., Simmonds, P., and Jones, M. (1988). HLA Haplotype A1 B8 DR3 as a Risk Factor for HIV-Related Disease. *The Lancet*, 331(8596):1185–1188.

Steinhauer, D. A., Domingo, E., and Holland, J. J. (1992). Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene*, 122(2):281–288.

Sterling, T. R., Lyles, C. M., Vlahov, D., Astemborski, J., Margolick, J. B., and Quinn, T. C. (1999). Sex Differences in Longitudinal Human Immunodeficiency Virus Type 1 RNA Levels among Seroconverters. *Journal of Infectious Diseases*, 180(3):666–672.

Stevens, C. E., Taylor, P. E., Zang, E. A., Morrison, J. M., Harley, E. J., Rodriguez de Cordoba, S., Bacino, C., Ting, R. C., Bodner, A. J., and Sarngadharan, M. G. (1986). Human T-cell lymphotropic virus type III infection in a cohort of homosexual men in New York City. *JAMA*, 255(16):2167–2172.

Stürmer, M., Preiser, W., Gute, P. a., Nisius, G. b., and Doerr, H. W. (2004). Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. [Editorial]. *AIDS*, 18(16):2109–2113.

Sundquist, W. I. and Kräusslich, H.-G. (2012). HIV-1 assembly, budding, and maturation. *Cold Spring Harbor Perspectives in Medicine*, 2(7):a006924.

Swindells, S., Cobos, D. G., Lee, N., Lien, E. A., Fitzgerald, A. P., Pauls, J. S., and Anderson, J. R. (2002). Racial/ethnic differences in CD4 T cell count and viral load at presentation for medical care and in follow-up after HIV-1 infection. *AIDS (London, England)*, 16(13):1832–1834.

Tang, J., Tang, S., Lobashevsky, E., Zulu, I., Aldrovandi, G., Allen, S., and Kaslow, R. A. (2004). HLA allele sharing and HIV type 1 viremia in seroconverting Zambians with known transmitting partners. *AIDS Research and Human Retroviruses*, 20(1):19–25.

Tatt, I. D., Barlow, K. L., Clewley, J. P., Gill, O. N. F., and Parry, J. V. (2004). Surveillance of HIV-1 Subtypes Among Heterosexuals in England and Wales, 1997-2000. [Miscellaneous Article]. *Journal of Acquired Immune Deficiency Syndromes August 15*, 36(5):1092–1099.

Tee, K. K., Pybus, O. G., Li, X.-J., Han, X., Shang, H., Kamarulzaman, A., and Takebe, Y. (2008). Temporal and Spatial Dynamics of Human Immunodeficiency Virus Type 1 Circulating Recombinant Forms 08_bc and 07_bc in Asia. *Journal of Virology*, 82(18):9206–9215.

Telenti, A. and Johnson, W. E. (2012). Host genes important to HIV replication and evolution. *Cold Spring Harbor Perspectives in Medicine*, 2(4):a007203.

The Swiss HIV Cohort Study (2010). Cohort Profile: The Swiss HIV Cohort Study. *International Journal of Epidemiology*, 39(5):1179–1189.

The UK Collaborative Group for HIV and STI Surveillance (2004). Focus on Prevention: HIV and other Sexually Transmitted Infections in the United Kingdom in 2003. Technical report, Health Protection Agency Centre for Infections, London.

The UK Collaborative HIV Cohort Steering Committee (2004). The creation of a large UK-based multicentre cohort of HIV-infected individuals: The UK Collaborative HIV Cohort (UK CHIC) Study. *HIV Medicine*, 5(2):115–124.

Thompson, M. A., Aberg, J. A., Hoy, J. F., Telenti, A., Benson, C., Cahn, P., Eron, J. J., Günthard, H. F., Hammer, S. M., Reiss, P., Richman, D. D., Rizzardini, G., Thomas, D. L., Jacobsen, D. M., and Volberding, P. A. (2012). Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society-USA panel. *JAMA*, 308(4):387–402.

Thompson, R., Brotherstone, S., White, I. M. S., Thompson, R., Brotherstone, S., and White, I. M. S. (2005). Estimation of Quantitative Genetic Parameters. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1469–1477.

To, T. H., Jung, M., Lycett, S. J., and Gascuel, O. (2015). Fast dating using least-squares criteria and algorithms. *In Press*.

Trask, S. A., Derdeyn, C. A., Fideli, U., Chen, Y., Meleth, S., Kasolo, F., Musonda, R., Hunter, E., Gao, F., Allen, S., and Hahn, B. H. (2002). Molecular Epidemiology of Human Immunodeficiency Virus Type 1 Transmission in a Heterosexual Cohort of Discordant Couples in Zambia. *Journal of Virology*, 76(1):397–405.

Travers, S. A. A., Clewley, J. P., Glynn, J. R., Fine, P. E. M., Crampin, A. C., Sibande, F., Mulawa, D., McInerney, J. O., and McCormack, G. P. (2004). Timing and Reconstruction of the Most Recent Common Ancestor of the Subtype C Clade of Human Immunodeficiency Virus Type 1. *Journal of Virology*, 78(19):10501–10506.

UK Collaborative Group on HIV Drug Resistance (2012). Time trends in drug resistant HIV-1 infections in the United Kingdom up to 2009: multicentre observational study. *BMJ*, 345.

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

Unlinked Anonymous Surveys Steering Group (2002). Prevalence of HIV and hepatitis infections in the United Kingdom 2001. Technical report, Department of Health, London.

Vallari, A., Holzmayer, V., Harris, B., Yamaguchi, J., Ngansop, C., Makamche, F., Mbanya, D., Kaptué, L., Ndembi, N., Gürtler, L., Devare, S., and Brennan, C. A. (2011). Confirmation of Putative HIV-1 Group P in Cameroon. *Journal of Virology*, 85(3):1403–1407.

Van de Peer, Y. (2003). Phylogeny inference based on distance methods. In Salemi, M. and Vandamme, A.-M., editors, *The Phylogenetic Handbook: A Practical Ap-*

proach to DNA and Protein Phylogeny, pages 101–136. Cambridge University Press, Cambridge, UK.

van der Kuyl, A. C. (2012). Personal Communications.

van der Kuyl, A. C., Jurriaans, S., Pollakis, G., Bakker, M., and Cornelissen, M. (2010). HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients. *AIDS (London, England)*, 24(10):1607–1608.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B. F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E. M., Sharp, P. M., Shaw, G. M., Delaporte, E., Hahn, B. H., and Peeters, M. (2006). Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature*, 444(7116):164–164.

van Vliet, C., Meester, E. I., Korenromp, E. L., Singer, B., Bakker, R., and Habbema, J. D. (2001). Focusing strategies of condom use against HIV in different behavioural settings: an evaluation based on a simulation model. *Bulletin of the World Health Organization*, 79(5):442–454.

Vanhems, P., Lambert, J., Guerra, M., Hirschel, B., and Allard, R. (1999). Association between the rate of CD4+ T cell decrease and the year of human immunodeficiency virus (HIV) type 1 seroconversion among persons enrolled in the Swiss HIV cohort study. *The Journal of Infectious Diseases*, 180(6):1803–1808.

Veugelers, P. J., Page, K. A., Tindall, B., Schechter, M. T., Moss, A. R., Winkelstein, W. W., Cooper, D. A., Craib, K. J. P., Charlebois, E., Coutinho, R. A., and van Griensven, G. J. P. (1994). Determinants of HIV Disease Progression among Homosexual Men Registered in the Tricontinental Seroconverter Study. *American Journal of Epidemiology*, 140(8):747–758.

Vidal, C., García, F., Romeu, J., Ruiz, L., Miró, J. M., Cruceta, A., Soriano, A., Pumarola, T., Clotet, B., and Gatell, J. M. (1998). Lack of evidence of a stable viral load set-point in early stage asymptomatic patients with chronic HIV-1 infection. *AIDS*, 12(11).

von Neumann, J. (1963). The general and logical theory of automata. In Aspray, W. and Burks, A., editors, *Papers of John von Neumann on Computing and Computer Theory (Charles Babbage Institute Reprint Series for the History of Computing vol 12)*. MIT Press, Cambridge, MA.

Waldstätter, R. (1989). Pair Formation in Sexually-Transmitted Diseases. In Castillo-Chavez, C., editor, *Mathematical and Statistical Approaches to AIDS Epidemiology*, volume 83 of *Lecture Notes in Biomathematics*, pages 260–274. Springer Berlin Heidelberg.

Walsh, B. and Lynch, M. (2012). *Evolution and Selection of Quantitative Traits: I. Foundations*, volume I. Sinauer, Sunderland, MA. pp. 179-208.

Watts, C. H. and May, R. M. (1992). The influence of concurrent partnerships on the dynamics of HIV/AIDS. *Mathematical Biosciences*, 108(1):89–104.

Wawer, M. J., Gray, R. H., Sewankambo, N. K., Serwadda, D., Li, X., Laeyendecker, O., Kiwanuka, N., Kigozi, G., Kiddugavu, M., Lutalo, T., Nalugoda, F., Wabwire-Mangen, F., Meehan, M. P., and Quinn, T. C. (2005). Rates of HIV-1 Transmission Per Coital Act, by Stage of HIV-1 Infection, in Rakai, Uganda. *Journal of Infectious Diseases*, 191(9):1403–1409.

Weiss, P. J., Brodine, S. K., Goforth, R. R., Kennedy, C. A., Wallace, M. R., Olson, P. E., Garland, F. C., Hall, F. W., Ito, S. I., and III, E. C. O. (1992). Initial Low CD4 Lymphocyte Counts in Recent Human Immunodeficiency Virus Infection and Lack of Association with Identified Coinfections. *The Journal of Infectious Diseases*, 166(5):1149–1153.

Wertheim, J. O., Brown, A. J. L., Hepler, N. L., Mehta, S. R., Richman, D. D., Smith, D. M., and Pond, S. L. K. (2014). The Global Transmission Network of HIV-1. *Journal of Infectious Diseases*, 209(2):304–313.

Whittle, H., Morris, J., Todd, J., Corrah, T., Sabally, S., Bangali, J., Ngom, P. T., Rolfe, M., and Wilkins, A. (1994). HIV-2-infected patients survive longer than HIV-1 -infected patients. *AIDS*, 8(11).

Williams, I., Churchill, D., Anderson, J., Boffito, M., Bower, M., Cairns, G., Cwynarski, K., Edwards, S., Fidler, S., Fisher, M., Freedman, A., Geretti, A. M., Gilleece, Y., Horne, R., Johnson, M., Khoo, S., Leen, C., Marshall, N., Nelson, M., Orkin, C., Paton, N., Phillips, A., Post, F., Pozniak, A., Sabin, C., Trevelion, R., Ustianowski, A., Walsh, J., Waters, L., Wilkins, E., Winston, A., and Youle, M. (2014). British HIV Association guidelines for the treatment of HIV-1-positive adults with antiretroviral therapy 2012. *HIV Medicine*, 15:1–6.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Marck, E. V., Gilbert, M. T. P., and Wolinsky, S. M. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661–664.

Yin, Z., Brown, A., Hughes, G., Nardone, A., Gill, O., Delpech, V., and contributors (2014). HIV in the United Kingdom 2014 Report: data to end 2013. Technical report, Public Health England, London.

Yirrell, D. L., Robertson, P., Goldberg, D. J., McMenamin, J., Cameron, S., and Brown, A. J. L. (1997). Molecular investigation into outbreak of HIV in a Scottish prison. *BMJ*, 314(7092):1446.

Yue, L., Prentice, H. A., Farmer, P., Song, W., He, D., Lakhi, S., Goepfert, P., Gilmour, J., Allen, S., Tang, J., Kaslow, R. A., and Hunter, E. (2013). Cumulative Impact of Host and Viral Factors on HIV-1 Viral-Load Control during Early Infection. *Journal of Virology*, 87(2):708–715.

Zhang, L., Chung, C., Hu, B.-S., He, T., Guo, Y., Kim, A. J., Skulsky, E., Jin, X., Hurley, A., Ramratnam, B., Markowitz, M., and Ho, D. D. (2000). Genetic characterization of rebounding HIV-1 after cessation of highly active antiretroviral therapy. *Journal of Clinical Investigation*, 106(7):839–845.