How does transmission in HIV influence the evolution of virulence?

A thesis submitted for a Philosophical Doctorate at the Department of Infectious Disease Epidemiology, Imperial College, London.

George Shirreff

October 2012

Contents

0.1 List of Abbreviations 6
0.2 Declaration of own work 7
0.3 Abstract 8
0.4 Collaborators 9
0.5 Acknowledgements 10
0.6 Introduction 11
0.6.1 Why study HIV? 11
0.6.2 Measuring HIV virulence 12
0.6.3 Natural selection 13
0.6.4 Is HIV virulence evolving? 14
0.6.5 Conflicting within- and between-host selection 15
0.6.6 Maintaining virulence at the population level 17
0.6.7 Mechanics of SPVL 18
0.7 Brief overview of thesis 23

1 Transmission selects for HIV-1 strains of intermediate virulence: a modelling approach 24
1.1 Abstract 25
1.2 Introduction 25
1.3 Methods 27
1.3.1 Overview 27
1.3.2 Variance decomposition 28
1.3.3 Model 29
1.4 Results 41
1.4.1 Model fitting and parameter estimation 41
1.4.2 Convergence of SPVL distribution 41
1.5 Discussion 48
1.6 Conclusion 54

2 How virulent was the first strain of HIV? 56
2.1 Abstract 57
2.2 Introduction 57
2.3 Methods 59
2.3.1 Basic reproduction number, R_0 68
2.4 Results 70

2.5 Discussion . . . 79
2.6 Conclusion . . . 83

3 How effectively can phylogenies be used to measure heritability? 84
3.1 Abstract . . . 84
3.2 Introduction . . . 86
3.3 Methods . . . 88
3.3.1 Significance tests . . . 93
3.3.2 Quantifying heritability . . . 94
3.3.3 Alternative method of simulation . . . 94
3.3.4 Estimating true heritability . . . 95
3.3.5 Sensitivity testing . . . 96
3.3.6 Data . . . 97
3.4 Results . . . 98
3.4.1 Testing significance in the true data. . . . 98
3.4.2 Estimates of heritability . . . 105
3.4.3 Testing the sensitivity of each statistic . . . 112
3.5 Discussion . . . 112
3.5.1 Results from the data . . . 121
3.5.2 Comparison of statistics . . . 121
3.5.3 Interpreting heritability . . . 122
3.6 Conclusion . . . 125

4 Detection and measurement of heritability in HIV-1 virulence in Amsterdam. 127
4.1 Abstract . . . 127
4.2 Introduction . . . 129
4.3 Methods . . . 130
4.3.1 Netherlands data . . . 130
4.3.2 Methods for measuring heritability . . . 132
4.4 Results . . . 132
4.5 Discussion . . . 139
4.6 Conclusion . . . 140

5 Final Discussion 141
5.1 Conclusion . . . 145
Bibliography . . . 162

List of Figures

1.1 Disease progression with survival curve. 31
1.2 Likelihood surface for the environmental and mutational standard deviation (σ_E and σ_M). 42
1.3 Details of the likelihood surface. 43
1.4 Set-point viral load data for pairs of cohabiting individuals. . . . 44
1.5 The evolving distribution of SPVL. 46
1.6 Mean log10 SPVL, proportion infected and heritability over time. 47
1.7 Heritability of SPVL measured at equilibrium for each combination of parameters σ_M and σ_E. 48
1.8 Population variance of SPVL measured at equilibrium for each combination of parameters σ_M and σ_E. 49
1.9 The evolution of the mean and the epidemic growth pattern. . . 50
1.10 Evolution of SPVL from various scenarios of founding genotype. 51

2.1 Parameter estimation. 72
2.2 Distribution of key parameters. 73
2.3 Univariate analysis of the effect of each parameter on probability of emergence, measured by 100 year survival of the epidemic. . . 74
2.4 Univariate analysis of the effect of each parameter on probability of emergence, measured by $> 2\%$ prevalence. 75
2.5 Probability of epidemic as number of founders increases. 76
2.6 The interaction of each parameter with the effect of virulence on emergence, measured by 100 year survival of the epidemic. 77
2.7 The interaction of each parameter with the effect of virulence on emergence, measured by $> 2\%$ prevalence. 78
2.8 Probability of emergence and the link with estimated basic reproductive number. 80
2.9 Evolution of SPVL from different levels of founding virulence. . . 81

3.1 Phylogeny inferred for all the Rakai data. 99
3.2 Phylogeny inferred for the Rakai data by subtype. 100
3.3 Phylogeny inferred for all the Swiss data. 101
3.4 Estimated heritability in the entire Rakai dataset. 106
3.5 Estimated heritability in the subtype A infected individuals in the Rakai dataset. 107

3.6 Estimated heritability in the subtype D infected individuals in the Rakai dataset. 108
3.7 Estimated heritability in the entire Swiss dataset. 109
3.8 Estimated heritability in the MSM individuals with the Strict definition of viral load in the Swiss dataset. 110
3.9 Comparison of sensitivity of various statistics to measure heritability on the Rakai phylogeny. 113
3.10 Comparison of sensitivity of various statistics to measure heritability on the Swiss phylogeny. 114
3.11 Comparison of sensitivity of various statistics to measure heritability on the Swiss MSM Strict phylogeny. 115
3.12 Comparison of sensitivity of various statistics to measure heritability on the Rakai subtype A phylogeny. 116
3.13 Comparison of sensitivity of various statistics to measure heritability on the Rakai subtype D phylogeny. 117
3.14 Comparison of sensitivity of various statistics to measure heritability on the Rakai phylogeny, using an alternative method of simulation. 118
3.15 Comparison of sensitivity of various alternative Hierarchical Clustering statistics to measure heritability on the Rakai phylogeny. . 119
3.16 Comparison of sensitivity of various alternative Abouheif-Moran statistics to measure heritability on the Rakai phylogeny. 120

4.1 The best tree inferred from the entire Netherlands *pol* sequence dataset. 134
4.2 Estimating heritability using the Phylogenetic Pairs method. . . 135
4.3 Estimating heritability using the Pagel's λ method. 136
4.4 Estimating heritability, using an alternative method of simulation. 137

List of Tables

0.1 Methods for prevention of HIV. 16

1.1 Parameter values. 33
1.2 The range of values used to construct the latin hypercube sample. 39
1.3 Maximum likelihood estimates of σ_M and σ_E in bootstrap sample. 39
1.4 The combination of parameters with the highest value of σ_M in bootstrap sample 40

2.1 Parameter values taken from the literature. 67
2.2 Parameters under investigation. 68

3.1 Results from previous studies in heritability of set-point viral load. 87
3.2 Characteristics of the Rakai data. 89
3.3 Characteristics of the Swiss data. 90
3.4 Signficance tests of Rakai data. 102
3.5 Signficance tests of Swiss data. 103
3.6 The effect of sample size reduction on significance. 104
3.7 Median and 95% confidence intervals for ρ as estimated from the Rakai data. 111
3.8 Median and 95% confidence intervals for ρ as estimated from the Swiss data. 111

4.1 Characteristics of the Netherlands dataset. 131
4.2 Significance tests and estimated heritability in each of the subdivisions 138
4.3 Heritability estimates for the MSM subdivision of the data, using both the original and alternative method of simulation. 139

0.1 List of Abbreviations

HIV	Human Immunodeficiency Virus
AIDS	Acquired Immune Deficiency Syndrome
SPVL	Set-Point Viral Load
HLA	Human Leukocyte Antigen
SIV	Simian Immunodeficiency Virus
VVG	Viral Virulence Genotype
AOI	Age Of Infection
ANOVA	ANalysis Of VAriance
PP	Phylogenetic Pairs
HC	Hierarchical Clustering
PICv	Phylogenetic Independent Contrasts, variance
AM	Abouheif-Moran
MSM	Men who have Sex with Men

0.2 Declaration of own work

I declare that the thesis presented here is my own work, completed under the supervision of Professor Christophe Fraser and Dr William Hanage. All sources used in this thesis have been cited appropriately. Collaborations and publication are specified on the following page.

Signed:

George Shirreff
September 2012

0.3 Abstract

Human Immunodeficiency Virus (HIV) has a high mutation rate which allows it to evolve rapidly in response to selective pressures. We explore how transmission may naturally select HIV virulence. Set-point viral load (SPVL), the level of stable viraemia in early asymptomatic infection, is linked with the severity of infection, and is used as a proxy for virulence.

The first part investigates the virulence-transmission trade-off, which has been identified in several host-pathogen systems. Previous work hypothesised that this trade-off favours HIV-1 viruses of intermediate SPVL which are optimal for transmission. We analysed a deterministic model incorporating host and viral contributions to SPVL, transmission rates and duration of infection. We estimate key parameters governing SPVL evolution and examine the rate of evolution towards intermediate virulence. This model was extended to incorporate stochasticity to investigate how the founding virulence affected the probability of emergence into a new population. Strains of intermediate virulence were found to be more likely to emerge.

The second part explored heritability of virulence. Several recent studies have estimated heritability of SPVL, mostly by comparing between individuals identified as transmission pairs, but one used phylogenetic comparative analysis instead, which is applicable to more datasets. Many methods are available for this approach, and we evaluate these using simulated and previously analysed data, and develop new methods. None were able to identify heritability below a substantial threshold, which was higher than many of the results in previous studies. The two most successful methods were applied to a new dataset from the Netherlands, with an ambiguous result suggesting borderline detectable heritability.

This work explores the estimation of heritability using phylogenies and provides some further evidence for genuine heritability of SPVL. The models provide strong support for a previous hypothesis and a framework for investigating how HIV virulence evolves in populations.

0.4 Collaborators

Chapter 1. Transmission selects for HIV-1 strains of intermediate virulence: a modelling approach.
George Shirreff[1], Lorenzo Pellis[1], Oliver Laeyendecker[2,3], Christophe Fraser[1]

Published as: Shirreff G, Pellis L, Laeyendecker O, Fraser C (2011) Transmission Selects for HIV-1 Strains of Intermediate Virulence: A Modelling Approach. PLoS Comput Biol 7(10): e1002185. doi:10.1371/journal.pcbi.1002185

Chapter 2. How virulent was the first strain of HIV?
George Shirreff[1], Christophe Fraser[1]

Chapter 3. How effectively can phylogenies be used to measure heritability of virulence in HIV?
George Shirreff[1], Samuel Alizon[4], Huldrych Günthard[5], Oliver Laeyendecker[2,3], Christophe Fraser[1]

Chapter 4. Detection and measurement of heritability in HIV-1 virulence in Amsterdam.
George Shirreff[1], Ard van Sighem[6], Frank de Wolf[1,6], Christophe Fraser[1]

1. Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom.
2. National Institute of Allergy and Infectious Diseases, National Institutes of Health, Baltimore, Maryland, United States of America.
3. Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America.
4. Lab MIVEGEC UMR CNRS 5290, IRD 224, UM1, UM2, F-34394 Montpellier 5, France.
5. Division of Infectious Diseases and Hospital Epidemiology, University Hospital of Zŕich, University of Zŕich, Zŕich, Switzerland.
6. Stichting HIV Monitoring, Amsterdam, Netherlands.

0.5 Acknowledgements

I would like to thank my supervisor, Christophe Fraser, with whom I have worked closely on all aspects of this thesis. I have also received very helpful advice from my second supervisor Bill Hanage on several occasions. I thank both of my supervisors as well as Anne Cori, Lulla Opatowski, Henry Shirreff and Hannah Slater for proofreading. Andrew Rambaut and Neil Ferguson gave extensive helpful comments and advice as my external and internal PhD examiners, which have been incorporated into the work.

I have also received advice from and had very progressive discussions with Lorenzo Pellis, Samuel Alizon, Thibaut Jombart, Federico Calboli, Oliver Rattman, Roland Regoes, Deirdre Hollingsworth, Christo Albor and Jeff Eaton.

Data were managed and provided by Oliver Laeyendecker on behalf of the Rakai Health Sciences Program, Johns Hopkins; Huldrych Günthard on behalf of the Swiss HIV Seroconverters Cohort; and Ard van Sighem and Frank de Wolf at Stichting HIV Monitoring.

Funding was provided by Imperial College, with some additional travel money from the Medical Research Centre for Outbreak Analysis and Modelling. Thanks also to Ruth Tipples and Andrea Johnson for making life easier.

I would like to thank my parents, Lizzie and David Shirreff for their support, in particular during the writing up of this thesis.

0.6 Introduction

The virulence of human infectious agents ranges from the barely detectable (e.g. GB virus C [23]) to the highly pathogenic and frequently fatal (e.g. Ebola [119]). Appropriately, the level of virulence is normally considered in terms of its impact on the host, but the virulence also has a large effect on the fitness of the pathogen.

The fitness of an organism is defined as the ability to survive and produce offspring carrying ones genes into the next generation [167]. Much work examines pathogen fitness during the infection of a single host, in which context fitness may be considered its ability to productively infect host cells or tissues with a high replication rate. At the level of the population, however, the ability to transmit to new hosts is a crucial aspect of fitness. All else being equal, a pathogen maximises its fitness by keeping its host alive and preventing clearance by the immune system for as long as there are opportunities to transmit itself. This led to the widespread belief that all pathogens will evolve towards a state of low virulence [151]. However, if virulence is linked to the ability to transmit, then the pathogen is faced with a trade-off between the two. A more virulent strain pays a cost in terms of duration of infection, which is cut short due to host death or immune clearance, but it trades this off against the benefit of higher transmissibility during this shorter infection, and it may therefore outcompete a less virulent strain (reviewed in [5]). This work explores the evolution of virulence in the context of HIV.

0.6.1 Why study HIV?

An estimated 34 million people were living with HIV in 2010, of whom 68% were in Sub-Saharan Africa. There were 2.7 million new cases in that year and 1.8 million deaths from Acquired Immune Deficiency Syndrome (AIDS). The prevalence is still increasing but the annual incidence has been decreasing since its peak in the late 90s. The rate of AIDS-related mortality has started to decline in many areas including Sub-Saharan Africa, but is still high [221]. The severity of infection varies widely, with some individuals progressing rapidly to AIDS and death, and some surviving for decades [162], so any variation in HIV virulence has a very large human impact.

In addition to the public health benefit, studying the evolution of HIV is relevant in evolutionary biology to the study of the virulence-transmission trade-off [73], and to the study of the conflict between selection pressures at the host and the population level [4]. The trade-off hypothesis for the evolution of virulence has become well-established both in theory, and more recently with empirical data (reviewed in [5]). However the authors of that review discuss the limitations of the trade-off hypothesis, in particular the simplicity of the model which is both a benefit and a hindrance. Every host-pathogen system is complex, and

virulence may not respond to selection pressure in a predictable way [60].

The high mutation rate of HIV equips it well to adapt to a change in environment by rapidly producing new genotypes. Unlike DNA polymerase [85], which copies the genome in cellular organisms and DNA viruses, the two enzymes involved in replicating the HIV genome, RNA polymerase II and reverse transcriptase, have no proofreading function and they introduce a high rate of error (reviewed in [116]). Frequent coinfection of cells with multiple provirus [113] allows a high rate of recombination [139], which further increases the genomic diversity and allows a greater rate of adaptation to selective pressure [117, 166].

In such a long-lasting and rapidly diverging pathogen, the within-host evolutionary dynamics occur on a comparable timescale with transmission dynamics, whereas most epidemiological studies assume that the within-host dynamics occur instantaneously [3]. HIV as a rapidly evolving and data rich organism provides opportunities to explore the complexities of evolving host parasite systems.

0.6.2 Measuring HIV virulence

Virulence is defined verbally as the tendency of an infectious agent to reduce the fitness of the host. More specifically it is often defined as the rate at which the pathogen induces host mortality, in excess of the background mortality rate [8]. Since the mortality rate from HIV may be dependent on external factors such as the availability of treatment, the rate of progression to AIDS is a more appropriate measure of virulence. Progression to AIDS is currently defined clinically by the CD4 count or by one or more AIDS defining illnesses [32]. I use the set-point viral load (SPVL) as a proxy for the rate of progression, and therefore the virulence, of HIV. In brief, SPVL refers to the relatively stable density of virions in the blood which characterises asymptomatic infection. There is considerable population level variation in SPVL, in spite of its relative stability within the individual [95]. SPVL is widely used as a prognostic indicator for AIDS, as individuals with a higher SPVL have a higher rate of CD4 cell decline, and they tend to progress more rapidly to AIDS [230, 124] and die sooner as a consequence [152]. As a result of its relative constancy during asymptomatic infection, SPVL can be measured at a wide range of time points in an individuals infection [82].

However, other proxies are available. A study comparing various different measures as proxies for virulence (recent viral load, peak viral load, set-point viral load and copy years viraemia) found that the set-point did not perform as well as the other measures in predicting the hazard of AIDS or death [41]. Another study developed a single measure using principal component analysis which incorporated the trajectory of viral load, and had the advantage that it relaxed the specific timing needed to measure the prognostic marker for an individual. However, it requires that the single measure for an individual must be calcu-

lated in the context of an entire cohort, leaving questions as to which cohort should be used [105]. SPVL must be measured in a specific time window, and although this causes problems in measuring prognosis when time to seroconversion is unknown, it is also as a result a more stringent measure, and perhaps most importantly it is widely used. SPVL is also a good proxy with respect to the virulence-transmission trade-off because it correlates positively with the rate of transmission. Studies of heterosexual couples with discordant serostatus demonstrated that individuals with higher SPVL were more likely to infect their partners [69, 187].

Both rate of progression and rate of transmission determine the number of new infections an individual will generate from the time of infection to death. The transmission potential is the product of the rate of transmission and the life expectancy. Low SPVL results in a long life expectancy with many opportunities to transmit, but at a low rate of transmission, while high SPVL is associated with a high rate of transmission but for a shorter infectious period. A higher SPVL gives a higher rate of transmission, but with limiting returns, and as a result the transmission potential is maximised at intermediate SPVL [73]. Fraser et al. observed that the mean of the current distribution of SPVL is relatively similar to the value which optimises the transmission potential, and they proposed that the between-host environment has naturally selected for virus which produces infections with intermediate SPVL.

0.6.3 Natural selection

It is a fundamental principle that a trait must have heritable variation which conveys differences in reproductive success for evolution by natural selection to occur [47]. For the purposes of this work, the infection of a single host is considered the unit of reproduction, rather than an individual virus or provirus. In this thesis I explore how virulence is inherited and selected in the population, to investigate the hypothesis that the distribution of SPVL evolves towards intermediate values, and that the current distribution of SPVL in the population is a result of this mechanism [73].

Understanding the source of variation in SPVL is important in determining how the viral population will respond to selection. Host factors which influence SPVL have been identified [31, 53, 64], while less work has been done in identifying its viral correlates. However, virulence has been shown to differ between HIV subtypes [58] which suggests that viral control is important. This is further supported by longitudinal studies which show changes in SPVL or the rate of CD4 decline over time [96]. Several studies have attempted to quantify the heritability of SPVL in HIV infection, under various routes of transmission and different settings [4, 12, 94, 104, 215, 130, 223] and collectively they suggest that the viral genotype exerts substantial control over SPVL variation [158]. The level of heritability determines whether or not the viral population will evolve on an observable timescale.

0.6.4 Is HIV virulence evolving?

A recent meta-analysis examined changing virulence of HIV between 1984 and 2010 in 12 studies, 8 of which measured SPVL [96]. The study found a significant decrease in baseline CD4 count over time ($p<0.0001$), which is an indicator of increasing virulence. Most studies found a significantly increasing SPVL over time, but the overall trend was not significant ($p=0.07$). One study [160] found that the trends were risk-group specific. Changing virulence over time implies population level evolution of virulence which suggests directional selection. If the evolutionary dynamics are indeed risk group specific this may be because these groups have been founded by separate strains which differ in their virulence, or the different routes of transmission may result in different optimal levels of virulence, or may cause evolution to occur at different rates. A growing epidemic may also favour strains with high virulence which reproduce rapidly, even if they produce fewer lifetime transmissions than a strain with virulence which is optimal for an endemic situation [73, 25]. A rise in virulence may therefore be a consequence of entry into a new population.

It is likely that there is considerable random variation in the evolution as well as the measurement of virulence markers, making it hard to identify trends where they exist. Systematic bias may also be introduced by an artefact if the sensitivity of assays used to measure SPVL decrease over time, or if the samples themselves show lower viral load as they age. However, a CASCADE study, in Europe, Australia and Canada, which found an increase in SPVL, noted that controlling for the type of assay did not change the results [55].

A genuine trend in virulence may result either from evolution of virulence within a comparatively isolated epidemic, or by the replacement of the viral population by another which differs in virulence. A phylogeographic analysis demonstrated that the epidemics in various European countries and Israel were generally interspersed [175], suggesting that strains from abroad could readily be introduced. There is also evidence for changes in subtype proportions over time in two Sub-Saharan African epidemics [42, 188] which may be explained by differences in their virulence. However, a recent study in the Netherlands found an increase in SPVL over time within transmission clusters, which suggests that this is due to internal evolution of virulence rather than introduction of more virulent genotypes from elsewhere [21].

Evolution of virulence is also important in the context of artificial changes in transmission parameters. Several new methods for preventing infection are under development (reviewed in [168], see Table 0.1). It is hoped that they will reduce incidence as widely as possible, but they are unlikely to have a perfectly uniform effect across the range of viral virulence. Therefore there is the potential for a widespread intervention to change the optimal virulence and drive selection in the population. For example, antiretroviral treatment could preferentially target individuals with higher SPVL as they present with clinical AIDS earlier, and therefore select for lower virulence in the population, as

transmission from individuals on treatment is minimal [40]. Alternatively, an intervention which preferentially reduced transmission potential among lower viral loads could select for higher virulence. A vaccination which reduces pathogen toxicity, and therefore virulence, directly may result in an overall decrease in virulence because the benefit of producing toxins is decreased while the cost remains, as in the case of diphtheria and pertussis [212]. However, modelling work suggests that a vaccine which reduces the within-host growth rate, or reduces transmission from an infected, vaccinated individual may select in the population for more virulent strains, which will display more virulence when they infect non-vaccinated individuals [78]. Whether or not such selection provides opportunities for positive virulence management or causes perverse outcomes deserves further scrutiny when data on transmission in the context of heterogeneity of virulence becomes available. The extent to which such outcomes can be predicted in advance is unknown, and this indicates the need for surveillance to detect such an effect.

0.6.5 Conflicting within- and between-host selection

The mechanism proposed by Fraser et al. [73] for the population evolution of SPVL requires that viruses are selected for their potential to transmit. For this to be possible requires that virulence is heritable from one infection to the next, and the level of heritability determines how fast this trait will evolve. Subtypes exhibit differences in disease severity which imply viral control of virulence [58]. Several studies have quantified the heritability in SPVL [4, 94, 104, 215, 130]. This heritability may be mediated by qualitative differences between viral genomes in the population. For example, some viruses carry deletions in the nef gene which result in low viral loads and long term non-progression of disease [48]. However, in this work I am concerned with quantitative and unspecified traits which vary across all scales of the viral population. I use the term viral virulence genotype (VVG) to refer to all aspects of the viral genotype which contribute to variation in SPVL.

Such quantitative factors may experience conflicting levels of selection from the within-host and between-host environments. Viral mutations which increase the replication rate within one host may reduce its replication rate in the next, or may decrease the transmissibility between the two. Such conflict was suggested by Levin and Bull [138] who hypothesised that a rapidly evolving organism like HIV undergoes short-sighted evolution which increases the replication rate and therefore the virulence within the host at the expense of transmissibility. This hypothesis is supported by competition assays which demonstrate that replicative fitness is greater in virus isolated later in infection [219], suggesting that there are grounds for the existence of such a conflict between the levels of selection. Caution is required when assuming that viruses which evolve to be more productive in vitro will result in higher SPVL, since there is evidence that the set-point is established during primary infection [214], but this does not rule out

Table 0.1: Methods for prevention of HIV.

Treatment as prevention Treatment with anti-retroviral drugs reduces an individuals viral load, which improves the prognosis in the later stages of disease, but it can also prevent transmission at any stage. In the HPTN study 052, early treatment reduced 96% of infections between HIV-serodiscordant couples, compared with a control group who were treated when their CD4 count dropped below 250 cells/mm^3 [40].
Voluntary medical male circumcision Removal of the foreskin reduced the female-to-male transmission in three studies, and the meta-analysis risk ratio was 0.44 (95% confidence interval 0.330.60) [153]. A reduction in male-to-female transmission may occur but the effect was non-significant in one study [17] and viral-load dependent in the other [91].
Vaginal microbicide In the CAPRISA 004 study, use of 1% tenofovir gel within 12 hours both before and after vaginal intercourse significantly reduced the risk of infection by 39%, or higher (54%) in women with greater than 80% adherence [115].
Oral pre-exposure prophylaxis The iPrEx study found that daily doses of oral tenofovir combined with emtricitabine reduced HIV acquisition by 44% in men who have sex with men, and as with the vaginal microbicide, this was higher in individuals with evidence of greater adherence [89]. However a separate trial found that there was no corresponding effect in women, which may reflect a lack of drug efficacy, or poor adherence in the study [106].
Prevention of mother-to-child transmission Anti-retroviral drugs can effectively reduce transmission of HIV from mother to child during pregnancy, birth and breast-feeding to very low levels [208].
Vaccination The recent RV144 study in Thailand of the combined ALVAC-AIDSVAX vaccine showed a 31% reduction in HIV acquisition, although there was no reduction in viral load in those who became infected [192]. A vaccination leading to protective antibodies in the genital tract, passive immunisation by direct administration of neutralising antibodies, or gene transfer technology are all potential technologies under development (reviewed in [168]).
Cure? A potential approach to cure is to purge the latent viral reservoir from which the virus is normally able to make a rebound if treatment is interrupted. It is hoped that inducing these latent proviruses to replicate, under combination therapy, will cause them to be recognised by the immune system and destroyed, and this method is under clinical trial [148].

that these viruses with higher replication capacity may be transmitted.

Fraser et al. proposed intermediate virulence is favoured in HIV because of diminishing returns on higher virulence [73]. These diminishing returns may also occur at the within-host level, and various studies have modelled these dynamics. Depending on the function between cell mortality rate and replication rate, viruses with intermediate replication rates may maximise their burst size and outcompete those which are more virulent [43, 84]. Spatial structure may also produce a reduction in optimal virulence [190]. These studies suggest that within-host selection will not cause virulence to spiral upwards, but it is likely that within- and between-host selection pressures act in conflict, since the optimum replication rate is unlikely to be identical for these two levels of selection.

0.6.6 Maintaining virulence at the population level

Since transmission is a comparatively rare event compared with within-host replication, the within-host selective pressures act more consistently on the viral genome. In spite of this, how does the virus maintain a level of virulence which is sufficient for transmission to continue? For HIV to remain a successful pathogen requires that sufficient fidelity in the phenotype is maintained in spite of rapid diversification. The limit at which fidelity is too low for the identity to be maintained is known as the error threshold [62]. Whether or not HIV would be subject to an error catastrophe if more highly diverged viruses were always transmitted is unclear, but there is phylogenetic evidence that divergence at the population level is much slower than expected by the rate of within-host divergence [135].

Several mechanisms have been proposed for the isolation of the within-host selection pressure from the population level evolution, which results in this rate mismatch (reviewed in [144]). This review argues that the most plausible mechanism is one in which ancestral virus continues to circulate in spite of general diversification, and is preferentially transmitted.

There is phylogenetic evidence that ancestral genotypes are favoured for transmission over the current major strains in a donor [200]. HIV normally experiences a genetic bottleneck during transmission [93], which suggests either that only a few virions are transmitted, or that the virus is immediately subject to strong selection. Such selective pressure may favour ancestral virus, which has not adapted to the different environment of the previous host.

An example of a conflicting adaptation is coreceptor usage. Transmitted strains tend to favour the CCR5 coreceptor for entry into CD4 cells, while the CXCR4 receptor tropism tends to emerge, if at all, later in infection. However recent work has suggested that transmission of CXCR4-tropic virus is more common than previously concieved, depending on the route of transmission and the subtype [34]. Nevertheless, transmitted strains often have a different cell tropism to

major strains in the infector individual [201], which suggests that host-specific or late-stage adaptations may detract from transmissibility.

Adaptation to the host immune response may reduce its fitness in the next host. Virus isolated from recently infected individuals is much more susceptible in vitro to neutralising antibody than virus from a chronic stage infector [49], suggesting that the latter virus may have adapted to the immune pressure in the within-host environment, while transmission selected for a strain without these adaptations, either at the point of transmission or by selecting for reversion. The same may be true in perinatal transmission: env sequences isolated from infected infants conferred greater in vitro replicative capacity to standardised viruses than env from their mothers in chronic infection [123].

A putatively unadapted and therefore more transmissible strain may be present in low numbers in the same compartments as other strains, as minor strains in the donor are frequently transmitted [237], in a mechanism described as store and retrieve [144].

Alternatively, it may be that the transmissible strain remains in the mucosa, where it is specifically adapted for that environment and is therefore capable of being transmitted during sexual contact. There is frequent compartmentalisation between the mucosal and systemic virus [9, 180]. Langerhans cells, which are found in the mucosa, frequently migrate to the lymph nodes, are comparatively long-lived, and are therefore candidates for this mucosal reservoir of the infection (reviewed in [112]).

It is possible that correlation in SPVL from one infection to the next is due to factors other than viral genetics. Heritability implies that similarity in SPVL between an infectee and infector is determined by their shared viral genotype. It is possible that higher SPVL in the donor induces higher SPVL in the recipient by inoculation with a larger dose of virus. In rabbit haemorrhagic disease virus there is evidence for two modes of transmission, an avirulent and a virulent form, which are indistinguishable genetically and therefore may result from differences in infecting dose [228]. However, this is highly unlikely in HIV since most infections can normally be traced back to a single infecting virus, and it is rare that the inoculum is larger [93].

In this work I will evaluate and apply methods for measuring heritability in order to gain insight into the mechanisms which maintain viral characteristics between infections.

0.6.7 Mechanics of SPVL

After the initial infection which results in a peak in viraemia, viral load decreases to the relatively constant set-point, and remains at this point for some time before viral load increases again with the onset of AIDS. I will discuss the characteristics of the primary and disease stages before discussing what causes

the level of the SPVL.

Stage 1: primary infection.

This is characterised by influenza-like illness and a very high viral load in both plasma and peripheral blood mononuclear cells [46]. The viraemia reaches as high as 10 million virions/mm3 of blood [181] but this peaks within days of the onset of symptoms, falling away again rapidly [38]. The virus preferentially infects cells in the gastrointestinal tract or other lymphoid tissue, where the greatest concentration of target cells are found [29]. The spike in viraemia is attributable to the rapid exploitation of this resource, a dense network of CD4+CCR5+ target cells, and the rapid and largely irreversible depletion of gut-associated lymphatic tissue may be partly responsible for the reduction in viraemia at the end of primary infection (reviewed in [33] and [68]).

However, this fall has been largely attributed to the cellular immune response, for which various evidence is available. There is a temporal link between the fall in viraemia and the rise of the cellular immunity, while a neutralising antibody response does not appear until much later [125]; there is a link between the efficiency of the cellular response and the fall in viraemia [28]; the cellular response has an inhibitory effect in vitro [234]; and strong selection on the HIV genotype occurs in epitopes recognised by the cellular response [27]. Furthermore, a study which fitted models to viral and immune dynamics in Simian Immunodeficiency Virus (SIV)-infected macaques found that a model with only viruses and target cells explained the observed data poorly compared to one which also included cellular immunity [191].

The infectivity of an individual is strongly correlated with his/her viral load [187] and the rate of transmission during primary infection is therefore very high, and in fact higher than would be expected purely on the basis of higher viral load [103, 145], suggesting that there is more to the infectivity than simply quantity of virus in the body.

Stage 3: disease.

Aside from a brief and often undiagnosed period in primary infection, HIV-infected individuals first show symptoms following a gradual decline in immune system function. Both the absolute density of CD4 cells and the proportion they form of the total lymphocyte count fall over the course of infection, and AIDS is characterised by CD4 count less than 200 cells/μl, less than 14% of total lymphocytes, or the presence of an AIDS defining illness [32]. Without treatment, the immune damage continues, with the subsequent risk of opportunistic infections.

Immune damage begins in primary infection with depletion of gut lymphocytes, but continues with chronic immune activation associated with massive production of cytokines (reviewed in [68]). The reduction in CD4 cells is associated

with high viraemia and has two potential causes: directly by cytolysis caused by infection with the virus [150], and indirectly by the action of cytolytic immune cells [35]. There is recent evidence that the lymphocyte depletion in the gut increases intestinal permeability [185], allowing highly immunogenic material to cross the membrane and stimulate the immune system to produce the damaging chronic activation observed in HIV infection [30].

Many immunodeficiency viruses do not cause mortality in their hosts. SIV derived from the Sooty Managabey monkey (SIVsmm) is broadly non-pathogenic in its natural host, which has coexisted with SIV, but results in progression to AIDS in the rhesus macaque, an unnatural host from Asia where SIV has not been found in wild primates [122]. Rhesus macaques given the same virus showed lower viral load but a sustained depletion of CD4 cells and most progressed to AIDS within a few years [211]. This difference has been partly attributed to less immune activation in the natural host and reduced expression of the CCR5 co-receptor which is crucial for viral entry to CD4 cells [170]. The significantly reduced expression of this co-receptor in 5 natural hosts of SIV in comparison with 5 species of more recent hosts suggests that lower pathogenicity in natural hosts is a consequence of attenuation by host evolution [171]. Humans show the same pattern of disease progression as the other unnatural hosts, as infection is not cleared, and immune damage occurs over time, which results in immunosuppression.

What causes the level of SPVL?

The asymptomatic stage of infection is characterised by relatively constant viraemia. Although the viral load fluctuates around this point during this stage of infection, there is much greater variation between individuals than within an individuals infection [95].

There are a number of factors which contribute to this variation which are generally identified with host control. The sex of an individual has an effect on the viral load which interacts with their CD4 count [53, 199], and the age at seroconversion also affects viral load, the latter also substantially affecting the risk of disease progression (reviewed in [132]). Opportunistic infections and vaccination may cause transient increases in viraemia, known as blips, through immune activation [111]. Host genetic factors also play a major role. Human Leukocyte Antigen (HLA) type B alleles have been shown to have strong effects on SPVL, infectivity and risk of disease progression [79]. A genome wide association study found host genetic factors (not including sex) which account for 13% of the variation in SPVL [64]. An additional 9% of variation was explained by sex, age and population structure, although some of the latter could be based on unidentified host genetics. The study could only detect loci which explained greater than 1.3% of the variation in SPVL. Interestingly, however, a genome wide association study failed to find human genetic factors associated with HIV-1 acquisition [142]. Human genetic factors which correlate with low SPVL have been shown to have a reduced impact over time, which suggests that

viral evolution may be counteracting human protective genes, and suggests a mechanism by which viral virulence may increase over time [146].

A substantial body of experimental and observational work has examined the mechanisms which govern the balance between the virus and the immune system. Experiments with macaques showed that the rate of viral replication early in infection correlated with the subsequent stable viraemia, but this variation was attributed to host factors since the outcomes in four animals were very different despite their infection with identical virus [140]. The cellular immune response also plays a substantial role. In humans, SPVL correlates with qualitative aspects of the cellular immune response during primary infection, and interestingly these immune factors show greater correlation with subsequent SPVL than they do with the viral load at the time they are expressed [172]. During asymptomatic infection the magnitude and breadth of cellular immune response to Gag proteins correlated with low SPVL [61]. The functionality of Natural Killer cells has also been implicated as a determinant of SPVL [128]. Modelling studies have suggested that the balance between viral replication and immune response may have multiple steady states, so the initial conditions of the infection may be important [6, 229].

Fitting within-host models to clinical data has suggested that large variations in SPVL could result from small variations in multiple parameters contributing to processes such as target cell production, CD8 competition and infection rate [159]. An alternative modelling study proposed that the three parameters which determine SPVL are the death rate of infected cells, the fraction of infected cells which go on to produce virus, and the net rate of target cell production. Using clinical data, they suggested that the former two parameters do not vary enough to explain the variation in SPVL over several orders of magnitude, leaving only the net rate of target cell production as the potential cause of the large variation [26].

Viral characteristics have also been identified which influence the viral load, such as the number of N-linked glycosylation sites [37], and whether or not the nef gene is defective [86]. Mutations which prevent viral recognition by the cellular immune system occur in many viral genes and are associated with lower viral loads due to the fitness cost of escape [198]. Kouyos et al. [126] combine data from in vitro experimental fitness assays with viral load data using a machine learning tool [98], and identify 4.4% of variation in SPVL as determined by replicative fitness in the protease and reverse transcriptase genes. The replicative capacity was also found to be increasing over time.

While experimental and theoretical correlates of SPVL appear to be in the domain of either the host or the virus, it is difficult to use these data to quantify the relative effect of the host and the virus on variation in SPVL. For example, a viral characteristic may be correlated with high SPVL, but the underlying cause of variance in SPVL may be host characteristics which favour infection with virus carrying that characteristic.

What does SPVL mean in a time of treatment?

The widespread availability of anti-retroviral drugs which extend lifespans and reduce infectivity alters the relevance of SPVL. A study in Uganda found that twenty year olds on combination anti-retroviral therapy (cART) were expected to live an additional 26.7 years [154]. For an individual in Europe or North America, where cART has been more widely available and for longer, twenty year olds are expected to live an additional 40-50 years on cART. This depends on their sex, their risk group (injecting drug users have lower life expectancy) and their starting CD4 count [102], but is a very large improvement on the 1-2 years which were expected before such treatment was available [143].

This drastically changes the relationship between initial SPVL and duration of infection. Treatment also changes the relationship between initial SPVL and transmission rate. In the HPTN 052 trial which tested treatment for prevention, as discussed previously, infected individuals in serodiscordant partnerships began treatment earlier than normal. There were significantly fewer transmissions in the early treatment group compared with the control group, who began treatment, as normal, when their CD4 count dropped below 250 cells/?l, or at their first AIDS-defining illness [40].

Currently the data is not available to be able to predict how the virulence-transmission trade-off would respond to a population in which treatment was available, and so in this work I have focused on the evolution of virulence in the pre-treatment era.

0.7 Brief overview of thesis

The broad aim of this thesis is to investigate the evolution of virulence in human populations. It is divided into two parts, the first (chapters one and two) concerned with the natural selection of virulence in HIV, and the second (chapters three and four) concerned with the heritability of that virulence. Chapter one describes a deterministic model which simulates the transmission and evolution of virus in a large population. This is fit to data to estimate key parameters which govern evolution of SPVL, and we examine the rate of its evolution over time. In chapter two this is extended to include stochastic effects to examine how virulence affects whether a virus will emerge in the population or go extinct. Chapter three explores different methods for measuring the heritability of virulence between infections using only a phylogeny to establish relatedness, and evaluates their ability to measure heritability in simulated and real previously analysed data. In chapter four the two most successful methods are applied to a dataset from the Netherlands which has not been previously analysed.

Plots were produced using R [217], and the text, equations and tables were created in LaTeX [131, 155] with accompanying bibliography in BibTeX [177].

Chapter 1

Transmission selects for HIV-1 strains of intermediate virulence: a modelling approach

1.1 Abstract

Background
Recent data shows that HIV-1 is characterised by variation in viral virulence factors that is heritable between infections, which suggests that viral virulence can be naturally selected at the population level. A trade-off between transmissibility and duration of infection appears to favour viruses of intermediate virulence.
Methods
We developed a mathematical model to simulate the dynamics of putative viral genotypes that differ in their virulence. As a proxy for virulence, we use set-point viral load (SPVL), which is the steady density of viral particles in blood during asymptomatic infection. Mutation, the dependency of survival and transmissibility on SPVL, and host effects were incorporated into the model.
Results
The model was fitted to data to estimate unknown parameters, and was found to fit existing data well. The maximum likelihood estimates of the parameters produced a model in which SPVL converged from any initial conditions to observed values within 100-150 years of first emergence of HIV-1. We estimated 1) the host effect and 2) the extent to which the viral virulence genotype mutates from one infection to the next, and found a trade-off between these two parameters in explaining the variation in SPVL.
Conclusion
The model confirms that evolution of virulence towards intermediate levels is sufficiently rapid for it to have happened in the early stages of the HIV epidemic, and confirms that existing viral loads are nearly optimal given the assumed constraints on evolution. The model provides a useful framework under which to examine the future evolution of HIV-1 virulence.

1.2 Introduction

The median time between HIV-1 seroconversion and progression to symptomatic AIDS is approximately 10 years [15]. However, there is considerable variation in this rate of progression, with substantial proportions of patients progressing to AIDS in less than 5 years, or remaining AIDS-free after 20 years. Explaining this variability is an important goal of HIV pathogenesis research. Many cofactors which influence time to AIDS have been identified e.g. host genetics [79], host age [15], and recently viral factors have been implicated [4, 94, 215, 130, 16, 104, 120, 163].

This chapter explores the extent to which the VVG, changing from one infected individual to the next, may have evolved under natural selection in the early phase of HIV-1s history. Between-host selection, leading to changes in the virulence of HIV-1, has potential major implications for the number of human life

years affected. SPVL is used as a measure of virulence, partly for its appropriateness as a proxy which was discussed earlier, and partly because this work arises from a study which also used SPVL [73].

A simple conceptual model of how SPVL may evolve by between-host natural selection (i.e. selection for the more transmissible genotypes) requires consideration of the transmission potential of individual hosts of different SPVL. The transmission potential, defined as the product of duration of infection and transmission rate, increases with either component of this product. A positive correlation between SPVL and transmission rate has been convincingly demonstrated within heterosexual couples with initially discordant serostatus [69, 187, 141]. Since there is also a negative correlation between SPVL and duration of asymptomatic infection [230], there is therefore a trade-off between duration of and transmission rate during asymptomatic infection. Previous work has quantified this trade-off to suggest that SPVL most commonly observed in infections are optimal for transmission, suggesting that the distribution of SPVL was shaped by natural selection [73].

Natural selection requires that a trait has heritability from one generation to the next, in addition to variation and differential reproductive success. A number of recent studies have identified and quantified this heritable component of SPVL variation which is maintained from one infection to the next [4, 94, 215, 130, 104].

Recent studies from the Netherlands [90] and Italy [160] have found that the mean log10 SPVL has increased over the recorded history of an HIV-infected cohort, and the rate of CD4 cell decline has increased. However different transmission groups have demonstrated different patterns of evolution of SPVL. In the initial stages of the epidemic (mid 1980s) injecting drug users showed slower CD4 declines than heterosexuals or men having sex with men, but this difference decreased over the subsequent decade [160]. A study with similar methodology in Switzerland found stable virulence over the same time period [161]. This suggests that such trends may be area- and risk-group specific. In two studies showing an increase, the levels of SPVL in the earlier time points are lower [90, 160] than those which are optimal for transmission [73]. Various studies of the rate of CD4 decline also suggest an increasing virulence [44, 54]. A study of the in vitro replicative fitness of viruses sampled at different time points reported a decrease in replicative fitness over the course of the epidemic in Amsterdam [11] although a subsequent study of the same city which controlled for time of seroconversion found an increase [76]. A meta-analysis of several studies of evolving virulence markers over time found a significant decrease in baseline CD4 count over time, and a general but not significant increase in SPVL [96]. Overall, observational results on changing virulence are inconclusive, though they suggest either an equilibrium or a slow increase in that direction. If a genuine trend is indeed absent, this suggests that a) the global distribution of SPVL has stabilised at an equilibrium level; b) the rate of evolution is very slow or c) the distribution of SPVL is determined by factors which do not evolve.

However, c) is unlikely, first because of the observations on the heritability of SPVL described above, and second because there is evidence for evolution of SPVL occurring in individual areas or risk groups [96, 160, 90].

To address the expected dynamics of SPVL evolution, we developed and analysed a deterministic mathematical model of between-host transmission and evolution of the VVG incorporating known parameters linking SPVL to the duration of infection and the rate of transmission. The broad aim was to investigate the hypothesis that viral genotypes of intermediate virulence are naturally selected by transmission [73].

The primary of aim of this study was to use the observed distribution of SPVL to estimate the quantities of unknown host and viral factors which affect the process of between-host evolution. Comparing the model to data allowed the likelihood of the unknown parameters to be calculated.

The secondary aim was to assess whether the model, under these parameter estimates, allows convergence of the SPVL distribution towards an intermediate level, or at least to slowly changing levels consistent with observational studies, regardless of the VVG of the founding strain, and whether this can occur within a plausible timescale. The estimated time of origin of HIV-1 is before the most recent common ancestor, which has been dated to 1908 with 95% confidence interval 1884-1924 [233]. If evolution has occurred between the founding strain and current infections then it has occurred over a period of 100 years.

1.3 Methods

1.3.1 Overview

The model described in this chapter simulated the dynamics of putative VVGs of HIV-1 which differ from one another in their mean log10 SPVL. SPVL was assumed to vary as a result of both host and virus factors. These genotypes differ in their reproductive success as a result of the dependency of duration of asymptomatic infection and transmission rate on SPVL. Their prevalences change over time through competition for susceptible individuals in a constant population.

The model is formulated as a standard HIV epidemic model in which different VVGs compete. Virulence is considered as a one-dimensional trait, with each genotype represented by a point on a one-dimensional spectrum of increasing virulence. When a person is infected by a virus of a given VVG, the infection is characterised by a SPVL which reflects the virulence, but also other non-viral factors. When transmitted, the viral genotype can also mutate to higher or lower levels of virulence.

The model encodes the natural history of infection. After infection, individuals experience a brief period of highly infectious acute stage, after which they progress to chronic asymptomatic infection. Their SPVL determines both the duration and infectiousness of this asymptomatic stage, after which their viral load and infectiousness increases again as they progress to AIDS and death. Individuals are assumed to engage in serially monogamous partnerships; a realistic description of the sexual network was not an aim of this study.

For the sake of parsimony, relatively simple mathematical models with minimal sets of parameters were used, and thus left some important questions open for further study. In particular, the effect of population structure, differences between subtypes, and superinfection were ignored, although analysis of a model with more realistic sexual mixing and stochastic fluctuations is described in chapter two. Only natural, untreated infection has been considered, which is appropriate to describing the evolution of the virus prior to the widespread adoption of antiretroviral therapy. We also did not address the question of conflicting directions of selection at the within and between host level which may be a fruitful area for future work. Intead within-host changes in virulence are modelled as random drift.

1.3.2 Variance decomposition

A useful practical and conceptual approach to interpreting various influences acting on SPVL is to decompose the total observed variance, ${\sigma_P}^2$, into its components, genotypic, mutational and environmental variance (${\sigma_G}^2$, ${\sigma_M}^2$, and ${\sigma_E}^2$, equation 1.1) [158].

$$\sigma_P^2 = \sigma_G^2 + \sigma_M^2 + \sigma_E^2 \tag{1.1}$$

The variance in the VVG is described by the genotypic variance ${\sigma_G}^2$. This refers to differences in SPVL between infected individuals caused by viral factors which are preserved from one infection to the next. Environmental variance, ${\sigma_E}^2$, refers to any source of SPVL variance external to the virus. Host factors e.g. age [196], sex [53] and other aspects of host genotype [64], in particular HLA type [79] contribute significantly to variation in SPVL between individuals, and there may be other human and non-human covariates of SPVL e.g. antigenic stimulation [72]. All of these factors, extrinsic to the virus, contribute to ${\sigma_E}^2$ in our terminology.

Mutational variance, ${\sigma_M}^2$, accounts for changes in the viral virulence genotype which result from mutation of the virus between one generation and the next (i.e. one infected host and the next) as a result of within-host replication and selection of the virus. Since the viral determinants of SPVL are not currently known, this cannot be related to the nucleotide substitution rate.The mutational

standard deviation, σ_M, is simply the expected difference in the viral component of SPVL between an index and a secondary infection.

Heritability, h^2, which has been quantified in previous studies, was defined as the fraction of variance explained by shared viral factors within a transmitting couple [104, 71]. We define h^2 as the proportion of variance in SPVL explained by heritable viral genetic factors (equation 1.2).

$$h^2 = \frac{\sigma_G^2}{\sigma_P^2} \tag{1.2}$$

1.3.3 Model

Viral genotypes and SPVL phenotypes

In order to simplify simulations, we modelled a discrete finite set of viral strains (genotype), each capable of producing a finite range of possible SPVL (phenotype).

Each infected host in the model carries a viral genotype, i, and has a phenotype, j. Hosts were not explicitly described in the model, rather the model specified the dynamics of relative prevalences of hosts infected with a virus of genotype i and phenotype j. In other words, a compartmental multi-strain epidemic model.

Each genotype is defined by a predisposition to give rise to higher or lower SPVL. Following the decomposition given by equation 1.1, viral loads can be given by equation 1.3, where e_j is the environmental component (with mean zero and variance ${\sigma_E}^2$) and μ_i is the component attributed to viral factors. For a population of individuals infected with viral genotype i, the mean log10 SPVL will be given by μ_i, which is therefore a natural measure of the virulence of genotype i. For two viral genotypes i and k such that i is more virulent than k, i.e. $\mu_i > \mu_k$, not all individuals infected with genotype i will have higher SPVL than individuals infected with genotype k, but on average they will.

$$V_j = \mu_i + e_j \tag{1.3}$$

The means log10 SPVL for the viral genotypes, μ_i, are in the range 2.0-7.0, and SPVL phenotypes, V_j, are in the range 0.0-9.0, discretised with step 0.05 and 0.025 respectively. An individual carrying genotype i, will have a phenotype j with a probability denoted by f_{ij} which is taken from a normal distribution with mean μ_i and variance ${\sigma_E}^2$ (equation 1.4), discretised over the interval 0.025 and normalised to sum to one for each genotype i.

$$f_{ij} = \Phi(V_j|\mu_i, {\sigma_E}^2) \tag{1.4}$$

Prevalence

The prevalence of infections with viral genotype i, SPVL phenotype j, and age of infection a is represented by $Y_{ij,a}(t)$ at time point t. The age of infection is the time since the individual was infected. During the course of an infection each host passes through three stages, primary, asymptomatic and disease (AIDS) (P, A and D) as the age of infection a increases.

Duration of infection

Primary and disease stages have equal duration (D_P and D_D) and rate of transmission (β_P and β_D), regardless of SPVL. Duration of and rate of transmission during asymptomatic infection are dependent on SPVL and the relationships were modelled as Hill functions as fitted in Fraser et al. [73], from which the parameter values relating to these functions were also taken (Table 1.1. The mean duration of the asymptomatic stage of infection for a given SPVL j is given by $D_A(V_j)$ (equation 1.5).

$$D_A(V_j) = \frac{D_{max} {D_{50}}^{D_k}}{(10^{V_j})^{D_k} + {D_{50}}^{D_k}} \tag{1.5}$$

The progression from asymptomatic to disease stage is governed by a survival function in equation 1.6, in which $SP_{j,a}$ is the probability of an individual with SPVL V_j remaining AIDS free at age of infection a.

$$SP_j, a = \begin{cases} 1 & 0 \leq x < D_P \\ e^{-\left(\frac{(a \quad D_P)\Gamma\left(1+\frac{1}{\rho}\right)}{D(V_j)}\right)^{\rho}} & D_P \leq a \end{cases} \tag{1.6}$$

Rate of transmission

The unadjusted rate of transmission during this stage is given by $\hat{\beta}_A(V_j)$ in equation 1.7.

$$\hat{\beta}_A(V_j) = \frac{\beta_{max}(10^{V_j})^{\beta_k}}{(10^{V_j})^{\beta_k} + {\beta_{50}}^{\beta_k}} \tag{1.7}$$

Each of the rates of transmission in all three stages, S, are adjusted in equation 1.8 to approximate a serial monogamy model with a partner change rate c and the duration of the stage, D_S, according to the method in Hollingsworth et al. [103].

$$\beta_{Sj} = \frac{\hat{\beta}_{Sj} c}{\hat{\beta}_{Sj} + c + \frac{1}{D_{Sj}}} \tag{1.8}$$

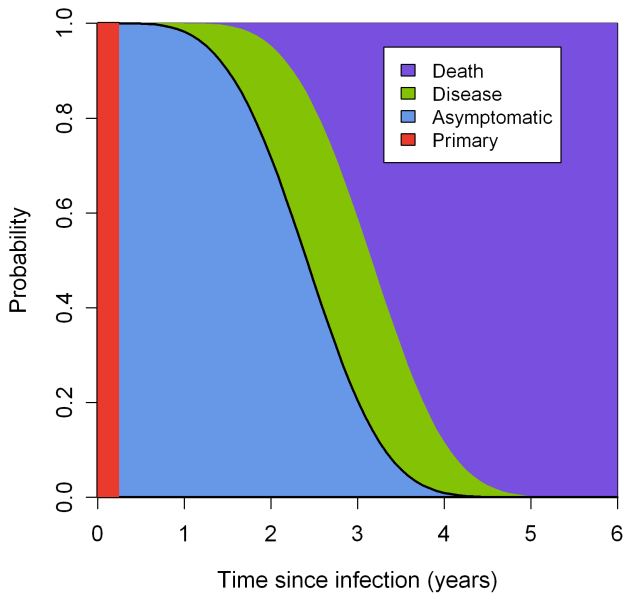


Figure 1.1: Possible disease progression outcomes for an infection with log10 SPVL of 6.0. All individuals have the same length of primary and disease stage infection, regardless of SPVL. The survival function is the border between asymptomatic and disease stage infection (survival here refers to survival from progression to AIDS, not death). A similar pattern is seen at other SPVL, but with a different survival function.

The rate of transmission for a given viral load j and age of infection a, β_j, a, is given in 1.9 which incorporates the different stages of infection and the curve for survival during asymptomatic infection. It is equivalent to β_P for the duration of primary infection D_P but thereafter a proportion of individuals of SPVL category j is in asymptomatic infection according to the survival probability $SP_{j,a}$. The proportion that enter the disease stage remain in it for a fixed period of time (D_D) before death, and the rate of transmission at that age of infection reflects the proportions of individuals in asymptomatic stage, disease stage, and death.

$$\beta_{j,a} = \begin{cases} \beta_P & 0 \leq a < D_P \\ \beta_A(V_j)SP_{j,a} + \beta_D(SP_{j,a-D_D} - SP_{j,a}) & D_P \leq a \end{cases} \tag{1.9}$$

This is illustrated for an example SPVL category j in Figure 1.1.

Mutation

Between generations a between-host mutation step occurs, so the force of infection for genotype k seeds a distribution of genotypes. The probability mik of an infection with genotype mean μ_k mutating so as to seed a new infection with genotype mean μ_i is taken from a normal distribution with mean μ_k and variance ${\sigma_M}^2$ (equation 1.10), discretised over intervals of 0.05, and normalised to sum to one for each genotype k.

$$m_{ik} = \Phi(\mu_i|\mu_k, {\sigma_M}^2) \tag{1.10}$$

Note that although this mutation refers to changes in the viral genotype, it does not relate directly to any particular mutation, but is rather a measure of the change in the distribution of viral genotypes that occurs over the course of infection within the host.

This model for the change that occurs from one infection to the next, defined by equation 1.10, represents the simplest possible model of the effect of within-host evolution on the distribution of transmitted viruses. More complex models, with directional and host-dependent selection, could feasibly be encoded in more complex mutational matrices.

Numbers of susceptibles

The change in the numbers of infected individuals cannot be described by differential equations. Although the rate of entering the infected class is governed by a constant rate, the rate of death is determined by a non-constant survival function from equation 1.6, so the net change is not at a constant rate, so it is better described in next generation equations, as follows.

The number of susceptibles, X, changes in each time step according to equation 1.11, where $Y_{ij,a}$ is the number of infected individuals of genotype i, viral load j and age of infection a.

The total number of infections, Y, for a given genotype in the next time step is calculated in equation 1.12 by the sum of the elementwise product of each SPVL j and age of infection a for a given genotype k, which is then multiplied by the probability that it will mutate into genotype i, m_{ik}. This is scaled according to $X(t)$, the proportion of susceptibles in the population at time t, meaning that the genotypes are competing for the available pool of susceptibles. To give the prevalence for each genotype and its SPVL category in the next set of new infections (where $a = 0$), this value is multiplied by the probability of genotype i producing SPVL category j, f_{ij}.

The prevalent infections are updated as in equation 1.13. The term $SP_{j,a}$ is the function of survival from progression to AIDS, given in equation 1.6. Since AIDS is a stage of determined length, D_D, the function of survival from death at

Table 1.1: Parameter values. Where possible these values have been taken from the literature, and a broad range of plausible values are applied to unknown parameters.

Symbol	Description	Value	Reference
β_P	Unadjusted rate of transmission during primary infection.	2.76 per year	[103]
β_D	Unadjusted rate of transmission during high transmission disease stage.	0.76 per year	[103]
D_P	Duration of primary infection	0.25 years	[103]
D_D	Duration of high transmission disease stage	0.75 years	[103]
β_{max}	Maximum rate of transmission during asymptomatic stage	0.317 per year	[73]
β_{50}	SPVL at which infectiousness is half maximum	13,938 copies per ml	[73]
β_k	Hill coefficient: steepness of increase in infectiousness as a function of SPVL	1.02	[73]
D_{max}	Maximum duration of asymptomatic stage	25.4 years	[73]
D_{50}	SPVL at which duration of asymptomatic infection is half maximum	3,058 copies per ml	[73]
D_k	Hill coefficient: steepness of decrease in duration as a function of SPVL	0.41	[73]
ρ	Shape coefficient for the Weibull survival distribution	3.46	[73]
c	Partner change rate	1.25	[226]
σ_E	Environmental standard deviation of log10 SPVL	0-1.2	Unknown
σ_M	Mutational standard deviation of log10 SPVL	0-1.0	Unknown
μ_0	Mean log10 SPVL of founding genotype	2.0-7.0	Unknown
Y_0	Proportion infected with founding genotype	0.001	Unknown

age of infection a is given by $SP_{j,a-D_D}$, the probability of surviving progression to AIDS at a time D_D years previously.

$$X(t+\Delta t)$$
$$= X(t) + \sum_{ij,a}(1 - SP_{j,a-D_D})Y_{ij,a}(t) - X(t)\sum_{ij} f_i j \sum_{a} \beta_{j,a} Y_{ij,a}(t) \quad (1.11)$$

$$Y_{ij,0}(t+\Delta t) = X(t) f_{ij} \sum_{k} m_{ik} \sum_{j,a} Y_{kj,a}(t)\beta_{ja}\Delta t \quad (1.12)$$

$$Y_{ij,a+\Delta t}(t+\Delta t) = Y_{ij,a}(t) SP_{j,a-D_D} \quad (1.13)$$

Solution to equilibrium using next-generation formalism

The basic reproductive rate, R_0, can be calculated for each genotype, and this can be combined with the pattern of mutation to calculate the genotype distribution at equilibrium using the next-generation formalism. The R_0 of each genotype is calculated in two steps. Firstly the transmission potential, TP_j is calculated for an infection with SPVL category j by multiplying the rate of transmission in each of the three stages of infection by the length of that stage (equation 1.14). The duration of asymptomatic infection $D_A(V_j)$ is the mean of the survival curve.

Secondly, the basic reproductive rate, R_{0i}, for each genotype i, is then calculated by taking the weighted average transmission potential, TP_j, weighted by the probability that infection with genotype i results in infection with SPVL category j (equation 1.15).

$$TP_j = \beta_P D_P + \beta_A(V_j) D_A(V_j) + \beta_D D_D \quad (1.14)$$

$$R_{0i} = \sum_{j} TP_j f_{ij} \quad (1.15)$$

The R_0 for each genotype k and the probability that genotype k mutates into genotype i (equations 1.10 and 1.15 respectively) can be used to calculate the next-generation matrix, $\mathbf{K}$, of which the individual elements are labelled K_{ik} (equation 1.16).

$$K_{ik} = m_{ik} R_{0k} \quad (1.16)$$

The distribution of genotypes at equilibrium is the eigenvector $\boldsymbol{\epsilon}$ corresponding to the dominant eigenvalue, λ, of $\mathbf{K}$ (equation 1.17).

$$\mathbf{K}\boldsymbol{\epsilon} = \lambda\boldsymbol{\epsilon} \tag{1.17}$$

The prevalence of SPVL category j, $p(V_j)$, at equilibrium in the population is then calculated according the host effect (equation 1.18).

$$p(V_j) = \sum_i \epsilon f_{ij} \tag{1.18}$$

This value can then be directly compared with the observed distribution of SPVL in the recipient partner in each transmission pair. By contrast, the probability of observing a given SPVL in an index case is conditional on the index partners having transmitted, so the probability is weighted according to the index partners rate of transmission during asymptomatic infection, wherein most infections occur (equation 1.19).

$$p(V_d|T) = \frac{\beta(V_d)p(V_d)}{\sum_j \beta(V_j)p(V_j)} \tag{1.19}$$

The likelihood at equilibrium of each run of the model is calculated by comparison with data from a previous study reporting the SPVL of phylogenetically confirmed transmission pairs [104] selected from a cohort in Rakai, Uganda [224, 225]. Putative transmission pairs were originally identified as cohabiting sexual partners by self-report. Phylogenetic analysis was performed on their env and gag genes and if sequences from each of the pair clustered together on the tree with greater than 80% bootstrap support they were considered to be a strongly linked couple, of which there were 29. In 23 of these pairs the direction of transmission was known due to the order of seroconversion. These individuals are assumed to be representative of the entire population.

The likelihood of observing the data is given by the probability of observing the index SPVL, V_d, and the recipients SPVL, V_r, (rounded to the nearest discrete viral load interval of 0.025), given in equation 1.20. This is calculated using conditional probabilities and is given as follows. The mean log10 SPVL of the genotypes infecting the recipient and index case are given by μ_x and μ_y. As these are unknown, all possible combinations of genotypes are considered. The value T represents the total number of transmissions expected to occur from a distribution of viral loads for a given genotype y (equation 1.21).

$$p(V_r \cap V_d) = \sum_x p(V_r|\mu_x) \sum_y p(\mu_x|\mu_y) \frac{\beta(V_d)p(V_d|\mu_y)p(\mu_y)}{T_y} \tag{1.20}$$

$$T_y = \sum_j \beta(V_j)p(V_j|\mu_y)p(\mu_y) \tag{1.21}$$

In equations 1.20 and 1.21 as well as in 1.19, the probability of the index case being a transmitter is proportional to β_j, which is the rate of transmission during the asymptomatic stage for viral load category j. This is used instead of the overall transmission potential since most transmissions occurred during the asymptomatic stage of infection [104].

The following constituents (equations 1.22, 1.23, and 1.24, in the latter of which ϵ_i refers to an individual element of the eigenvector $\boldsymbol{\epsilon}$) have been previously defined respectively in equations 1.4, 1.10 and 1.17.

$$p(V_r|\mu_x) = f_{xr} \tag{1.22}$$

$$p(\mu_x|\mu_y) = m_{xy} \tag{1.23}$$

$$p(\mu_i) = \epsilon_i \tag{1.24}$$

Equation 1.25 indicates the method for calculating the total log-likelihood of the data. The first term calculates and sums the total log-likelihood for each couple, z, for whom the direction of transmission is known, and this is done according to equation 1.20 according to the SPVL values of the recipient ($V_r(z)$) and the infector ($V_d(z)$). In the second term this is calculated for each couple, $\hat{z}$, for whom the direction of transmission is unknown. In each such couple the log-likelihood is calculated for each direction (male-to-female and female-to-male) and the mean is taken (in this case, $V_m(\hat{z})$ and $V_f(\hat{z})$ refer to the SPVL of males and females, respectively), assuming that either direction of transmission within a male-female couple is equally likely.

The confidence intervals in the likelihood plot were identified using a likelihood ratio test 1.26. Here is the point on the chi-square distribution with δ degrees of freedom which encompasses a proportion $(1-\alpha)$ of the area under the curve (1.39, 5.99 and 13.82 when α=0.50, 0.05 and 0.001 respectively). ℓ_{max} is the maximum log likelihood value, and $\ell_{\sigma_M \sigma_E}$ is the log likelihood for combination of parameters σ_M and σ_E. If the inequality is satisfied, then the values σ_M and σ_E produce a log likelihood not significantly different from the maximum likelihood at a $100(1-\alpha)\%$ confidence.

$$\begin{aligned}\ell = \sum_z \log(p(V_r(z) \cap V_d(z))) \\ + \sum_{\hat{z}} \frac{1}{2}\Big[\log(p(V_m(\hat{z}) \cap V_f(\hat{z})|\text{male donor})) \\ + \log(p(V_f(\hat{z}) \cap V_m(\hat{z})|\text{female donor}))\Big]\end{aligned} \tag{1.25}$$

$$-2(\ell_{\sigma_M \sigma_E} - \ell_{max}) \leq {\chi_2}^2(\alpha) \tag{1.26}$$

The likelihood was calculated for each combination of values of σ_E (range 0-1.2, step 0.005) and σ_M (range 0-1.0, step 0.005). Outside of these ranges the likelihood of observing the data is very low, as the variance of the equilibrium distribution becomes vastly higher than is observed. These values were used instead of their squares, ${\sigma_E}^2$ and ${\sigma_M}^2$, because they are on the same scale as log10 SPVL and are therefore directly related to the size of the host effect and of between-host mutation. Furthermore, using σ_E and σ_M gives greater resolution at lower values in the range of interest.

The values of Y_0 and $\mu_{\hat{i}}$ (the virulence of the founding value), were not included in the analysis of the equilibrium since they serve only as starting points in the model. All other parameter values were taken from the literature (Table 1.1).

Discretisation

The mean log10 SPVL of the viral genotypes and the categories of SPVL they produce were discretised in steps of 0.1 log10. Furthermore time was discretised into steps of one month. These step sizes were chosen for their convenience and were believed to give an appropriate level of resolution while keeping the computation time manageable. Reducing the discretisation step size had a negligible effect on the results (data not shown).

Starting prevalence

Finally, we also explored the effect of the proportion infected at the start, Y_0, and found that this had a negligible effect on the rate of mean convergence when the founding virulence was high, and an insubstantial effect when founding virulence was low (data not shown). The effect of discretisation was tested by reducing the SPVL phenotype and viral genotype step size to 0.01, and the time-step was reduced to 0.01 years. Upon repeating the analysis, there was a negligible difference in the results (data not shown).

Calculate heritability

Heritability is the proportion of total variation which is determined by genetic variation in the viral population. It was measured previously by calculating the proportion of the total variance (${\sigma_P}^2$) which was explained by carrying genetically similar virus (equation 1.2). This can be measured for the modelled distribution in a similar fashion. The total variance is calculated from the distribution of SPVL V_j (equation 1.27), and the heritability h^2 is the proportion of this variance which does not derive from non-heritable components, the environmental and mutational variance (${\sigma_E}^2$ and ${\sigma_M}^2$), and is calculated in equation 1.28.

$$\sigma_P{}^2 = \sum_r (V_r)^2 p(V_r) - \left(\sum_r V_r p(V_r) \right)^2 \tag{1.27}$$

$$h^2 = 1 - \frac{\sigma_E{}^2 + \sigma_M{}^2}{\sigma_P{}^2} \tag{1.28}$$

Bootstrap

The bootstrap was performed to refit the relationship between the data and the parameters, as reported in Fraser et al. [73]. The data had its source in papers published on studies from Netherlands [230] and Zambia [69]. The relationship between SPVL and duration of infection, and between SPVL and rate of transmission, were each fitted to a Hill function as this is sufficiently general and biologically relevant.

To measure the fit of the model for the duration of infection, the likelihood was calculated for each individual, s, and summed as in equation 1.29. Weibull(x, α, β) is the probability density function at x for the Weibull distribution with parameters α and β. $\Gamma(x)$ is the gamma function at x, and $\Gamma_c(x, \alpha, \beta)$ is the cumulative at x of the gamma distribution with parameters α and β. The value D_s is the time that the individual is followed up, regardless of outcome, and $D(V_s)$ is the expected time to progression given the individuals SPVL V_s which is calculated from equation 1.5.

$$\ell_{\text{progression}} = \sum_s \begin{cases} \ln\Big(\text{Weibull}(D_s, 1, D(V_s)) & \\ \quad + (\rho - 1)\ln\left(\rho \frac{D_s}{D(V_s)}\right) - \log(\Gamma(\rho))\Big) & \text{if } s \text{ progresses} \\ \ln\left(1 - \Gamma_c\left(D_s, \rho, \frac{D(V_s)}{\rho})\right)\right) & \text{if } s \text{ does not progress} \end{cases} \tag{1.29}$$

The likelihood for the transmission data was calculated similarly from the probability of an individual transmitting over the length of the observed period (equation 1.30). The rate of transmission $\beta(V_s)$ was calculated for each individual, s, from their SPVL as in asymptomatic infection (equation 1.7). The probability of transmission was given as follows, where F=1.833, the length of the study period.

$$\ell_{\text{transmission}} = \sum_s \begin{cases} \ln(1 - e^{-\beta(V_s)F}) & \text{if } s \text{ transmits} \\ \ln(e^{-\beta(V_s)F}) & \text{if } s \text{ does not transmit} \end{cases} \tag{1.30}$$

The bootstrap was performed by sampling with replacement the relevant dataset, and then fitting the parameters to this data using the optim command in R [217] which adjusts the parameters to maximise the likelihood. The starting point

Table 1.2: The range of values used to construct the latin hypercube sample. The values for each point in the hypercube were sampled from a uniform distribution within that range.

Parameter	**Minimum**	**Maximum**
D_{max}	1	50
D_{50}	1000	10,000
D_k	0.01	1
ρ	1	5
β_{max}	0.0001	2
β_{50}	1	100,000
β_k	0.0001	2

Table 1.3: The maximum likelihood estimates of σ_M and σ_E in 1000 bootstraps. The figures are the proportion of each combination of values of σ_M and σ_E which were the maximum likelihood estimate when a low resolution likelihood surface was calculated with 1000 sets of bootstrapped parameters. These exclude 19 bootstraps in which the optimised parameter values gave a next-generation matrix with mixed signs, rendering the result incalculable.

		σ_E												
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
σ_M	1	0	0.005	0.052	0.034	0.01	0.004	0.001	0	0	0	0	0	0
	0.9	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.7	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.5	0	0	0	0	0	0	0.007	0.003	0	0	0	0	0
	0.4	0	0	0	0	0	0	0.02	0.047	0	0	0	0	0
	0.3	0	0	0	0	0	0	0	0.282	0.007	0	0	0	0
	0.2	0	0	0	0	0	0	0	0.462	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0.065	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0

for the fit was chosen by Latin Hypercube Sampling in the *tgp* package [87, 88] to cover the available parameter space. Initial values were selected at random uniformly between the values in Table 1.2.

The bootstrap was repeated 1000 times for each relationship, and then a set of duration parameters was randomly assigned to a set of transmission parameters. These were then used as the parameters in the model. For each of these sets of parameters the maximum likelihood estimates for σ_M and σ_E were calculated (Table 1.3). The step size of σ_M and σ_E was 0.1 instead of 0.005 to reduce computation time. The point with the highest value of σ_M which was still of significant likelihood was also identified (Table 1.4).

Table 1.4: The combination of parameters with the highest value of σ_M in 1000 bootstraps. The figures are the proportion of each combination of values of σ_M and σ_E which formed the highest value of σ_M which was still consistent with the 95% confidence region of the maximum likelihood estimate. Where several values of σ_E were available, the one with the highest likelihood was chosen.

		σ_E												
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
	1	0	0.2	0.244	0.06	0.01	0.004	0.001	0	0	0	0	0	0
	0.9	0	0	0	0.036	0	0	0	0	0	0	0	0	0
	0.8	0	0	0	0	0.053	0	0	0	0	0	0	0	0
	0.7	0	0	0	0	0	0.112	0	0	0	0	0	0	0
	0.6	0	0	0	0	0	0.162	0	0	0	0	0	0	0
σ_M	0.5	0	0	0	0	0	0	0.082	0	0	0	0	0	0
	0.4	0	0	0	0	0	0	0.035	0	0	0	0	0	0
	0.3	0	0	0	0	0	0	0.002	0	0	0	0	0	0
	0.2	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Convergence of SPVL distribution

The next-generation formalism solution described above is sufficient for analysing the equilibrium distribution of SPVL as the end results are identical. However, the model must be run in full to determine the rate at which SPVL evolves in real time.

To run the model in continuous time, the infection is initialised at time $t = 0$ for the starting genotype $\hat{i}$ with mean $mu_{\hat{i}}$ and a proportion Y_0 of the population are infected (equation 1.31). The total number of infected individuals at the start of the epidemic all enter genotype category $\hat{i}$, and are divided up between all the SPVL categories according to $f_{\hat{i}j}$. All other genotype categories begin at zero (equation 1.32) as do all ages of infection greater than zero (1.5.3) (equation 1.33). The model was run for 500 years in discrete time-steps corresponding to one month for each set of the parameter values.

$$\forall_k \quad Y_{\hat{i}j,0}(0) = Y_0 f_{\hat{i}j} \tag{1.31}$$

$$\forall_{k \neq \hat{i}} \forall_j \quad Y_{kj,0}(0) = 0 \tag{1.32}$$

$$\forall_i \forall_j \forall_{a>0} \quad Y_{ij,a}(0) = 0 \tag{1.33}$$

1.4 Results

1.4.1 Model fitting and parameter estimation

The primary aim of the analysis was to quantify separately the effects of host and virus on variation in SPVL by estimating ${\sigma_M}^2$ and ${\sigma_E}^2$ simultaneously, to gain a better estimate of the extent of viral factors in individual infections, and the parameters needed to predict evolution. The values of the environmental and mutational standard deviations (σ_M and σ_E) were estimated using a maximum likelihood approach. Since the model predicts not just the distribution of SPVL, but how they change from one infection to the next, the model could predict the observed SPVL in both index and recipient partners in transmitting couples.

Figure 1.2 shows the likelihood surface for the environmental and mutational standard deviations (σ_M and σ_E), and the bivariate confidence bounds. The maximum likelihood estimates are σ_M=0.12 (95% confidence interval 0.00 to 0.39) and σ_E=0.66 (95% confidence interval 0.47-0.94). The estimates with highest mutational standard deviation within the 95% confidence bounds are $\sigma_M = 0.39$ and $\sigma_E = 0.55$ referred to later as the most mutable plausible scenario. Figure 1.3 shows the likelihood surface in more detail, and the diagonal nature of the region of high likelihood indicates a trade-off between the two parameters in terms of the quality of model fit.

The top row of Figure 1.4 shows the raw data consisting of the SPVL values of 29 couples, for whom 23 the direction of transmission is known and for 6 it is unknown (a and b, respectively). The bottom row shows the quality of fit of the model to the distribution of SPVL in index partners and recipients in transmitting couples (c and d). The model appears to describe the data well, capturing the higher peak in donors of intermediate virulence to some extent, although with so few datapoints it is difficult to assess the fit visually. The estimated heritability was 26%, which is very close to the value of 27% which emerged from a previous statistical analysis of these couples [104].

1.4.2 Convergence of SPVL distribution

Having derived maximum likelihood estimates of parameters from an equilibrium solution to the model, the dynamics of genotype competition were then simulated numerically in order to assess whether or not convergence would occur under those parameter values, and on what timescale the convergence would occur.

The evolution of the SPVL distribution is shown in Figure 1.5. Regardless of whether the virulence of the founding genotype was high or low, the SPVL evolved towards an intermediate level with a mean log10 SPVL of 4.5. This can also be seen in Figure 1.6a.

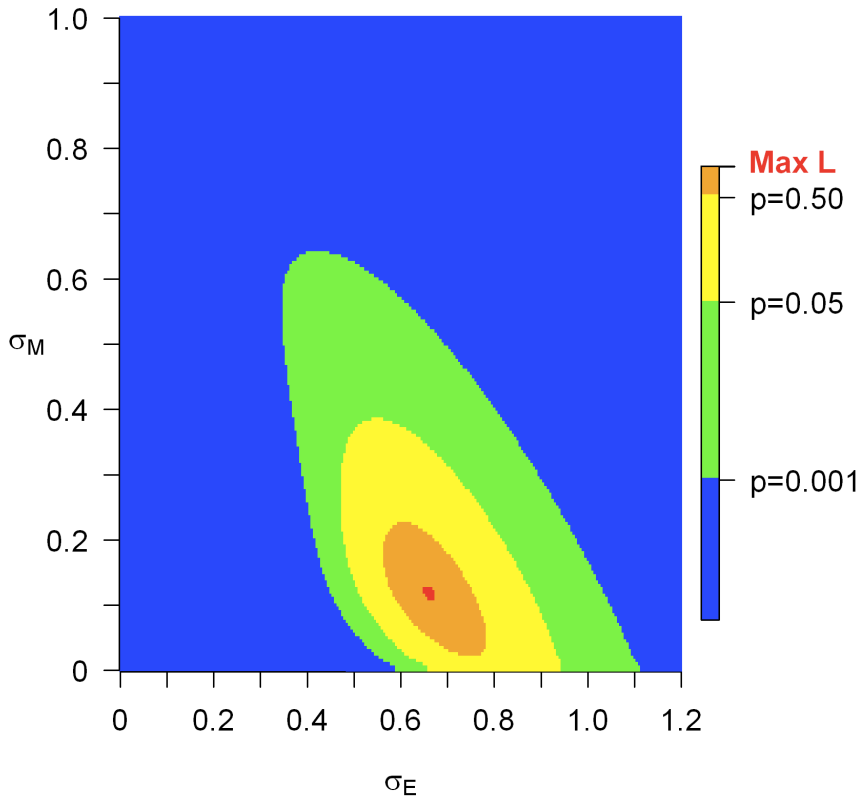


Figure 1.2: Likelihood surface for the environmental and mutational standard deviation (σ_E and σ_M). The maximum likelihood estimate is represented by the red point, and the regions of 50%, 95% and 99.9% confidence in orange, yellow and green respectively. The method for calculating confidence intervals is given in 1.26.

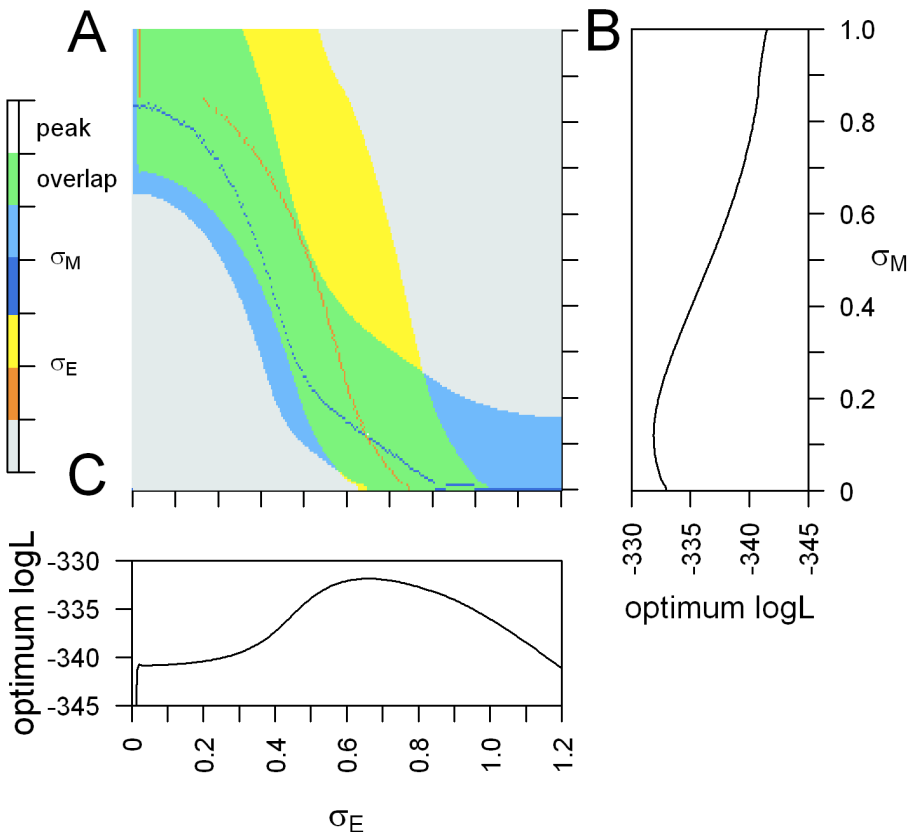


Figure 1.3: Details of the likelihood surface. (a) For each value of σ_M, the value of σ_E which gives the highest likelihood is marked in orange on the figure, while the yellow region gives the 95% confidence bounds. Similarly, for each value of σ_E, the optimum σ_M value is marked in dark blue, with 95% confidence bounds in light blue. Where the maximum likelihood regions for the two parameters overlap this is marked in green, and the point of maximum likelihood is white. (b) Likelihood at the optimum value of σ_E for each value of σ_M i.e. it tracks the likelihood of the orange line. (c) Likelihood at the optimum value of σ_M for each value of σ_E i.e. it tracks the likelihood of the dark blue line.

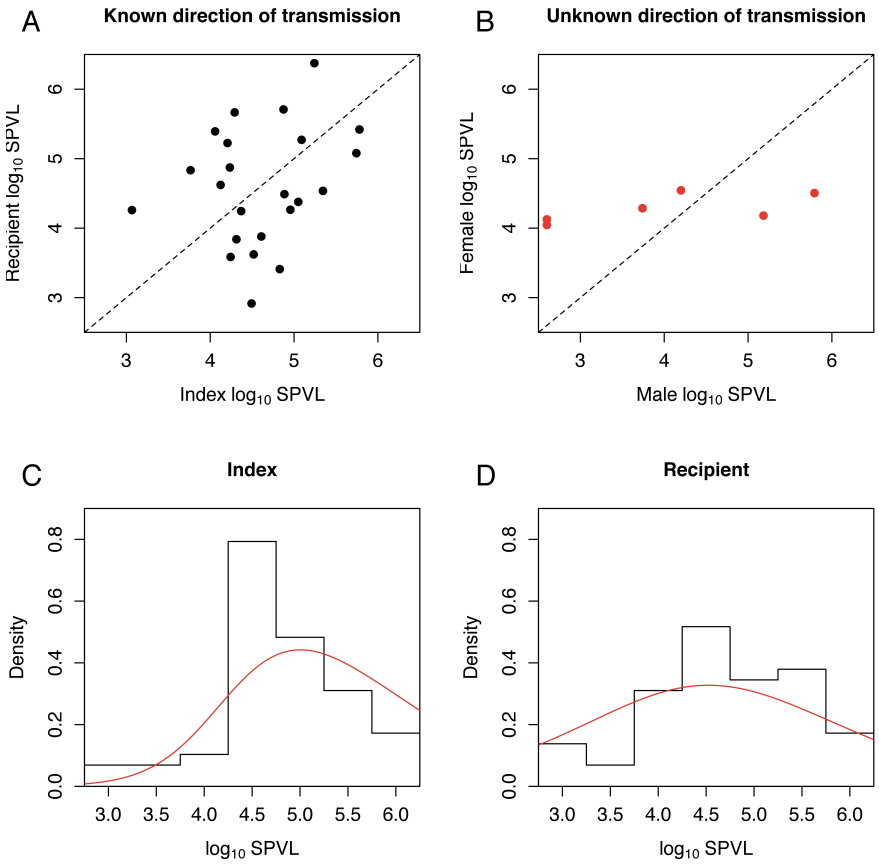


Figure 1.4: The log10 SPVL values of the strongly linked couples data, for those with a) known and b) unknown direction of transmission. The dotted line indicates where the couple would have equal values. Fit of the model (red line) to data (black line) in c) the index partner, and d) the recipient. The data in the histogram have been discretised to their nearest 0.5 log10 SPVL for purposes of display (where in the analysis they are discretised by 0.025 log10 SPVL). Where the role in transmission is unknown, each individual in the pair represents half a recipient and half an index case. The modelled distributions were calculated from equations 1.18 and 1.19 for the recipient and index partner, respectively.

This convergence on intermediate SPVL values also occurred when other combinations of parameter values in the region of high likelihood (Figure 1.2) were used instead. The rate of convergence was positively related to σ_M, as shown in Figure 1.6a, where the maximum likelihood prediction is compared to the most mutable plausible scenario. Convergence towards intermediate virulence occurred in approximately 150 years under the maximum likelihood values. There was still change in the mean after this time but runs beginning with high or low virulence converge around this time point. The same point was reached in 50 years under the most mutable plausible scenario.

The heritability was also calculated over time (Figure 1.6c) and under maximum likelihood values of σ_E and σ_M this reached equilibrium at 26%, which is consistent with previous studies [4, 94, 215, 130, 104]. Further details of the heritability and variance at equilibrium are given in Figure 1.7 and 1.8.

In order to examine how changes in virulence are related to the stage of the epidemic, the proportion infected was compared with mean log10 SPVL. The effect was most evident when the founding virulence closely matched the equilibrium virulence (Figure 1.9). During the epidemic growth phase the mean virulence increased to levels above the optimum, and then returned to the optimum as the proportion infected reached equilibrium.

We varied the founding virulence to investigate its effect on rate of convergence (Figure 1.10). This had a marked effect on how quickly the mean log10 SPVL reached equilibrium (4.52 log10 SPVL). When the founding genotype had mean 4.5 log10 SPVL, equilibrium with regard to the mean was reached very quickly, and the more different the SPVL of the founding genotype, the longer the time to convergence. A similarly rapid convergence is seen if all genotypes had equal prevalence at the start of the run. The mean underwent little change (Figure 1.10a, red line) but the variance rapidly decreased as the most successful genotype, already present in the population, began to dominate (Figure 1.10b).

Finally, bootstrapping techniques were used to investigate the sensitivity of these findings to the choice of parameter values determining the dependencies of infectiousness and duration of asymptomatic infection on SPVL. These parameters were previously estimated from datasets from Amsterdam and Zambia [73]. The original datasets were resampled with replacement to match the original sample size, and for each of 1000 resamples the corresponding maximum likelihood estimates of σ_E and σ_M were calculated. The resulting maximum likelihood estimates (Figure 1.3) are similar to those from the principal analysis (Figure 1.2).

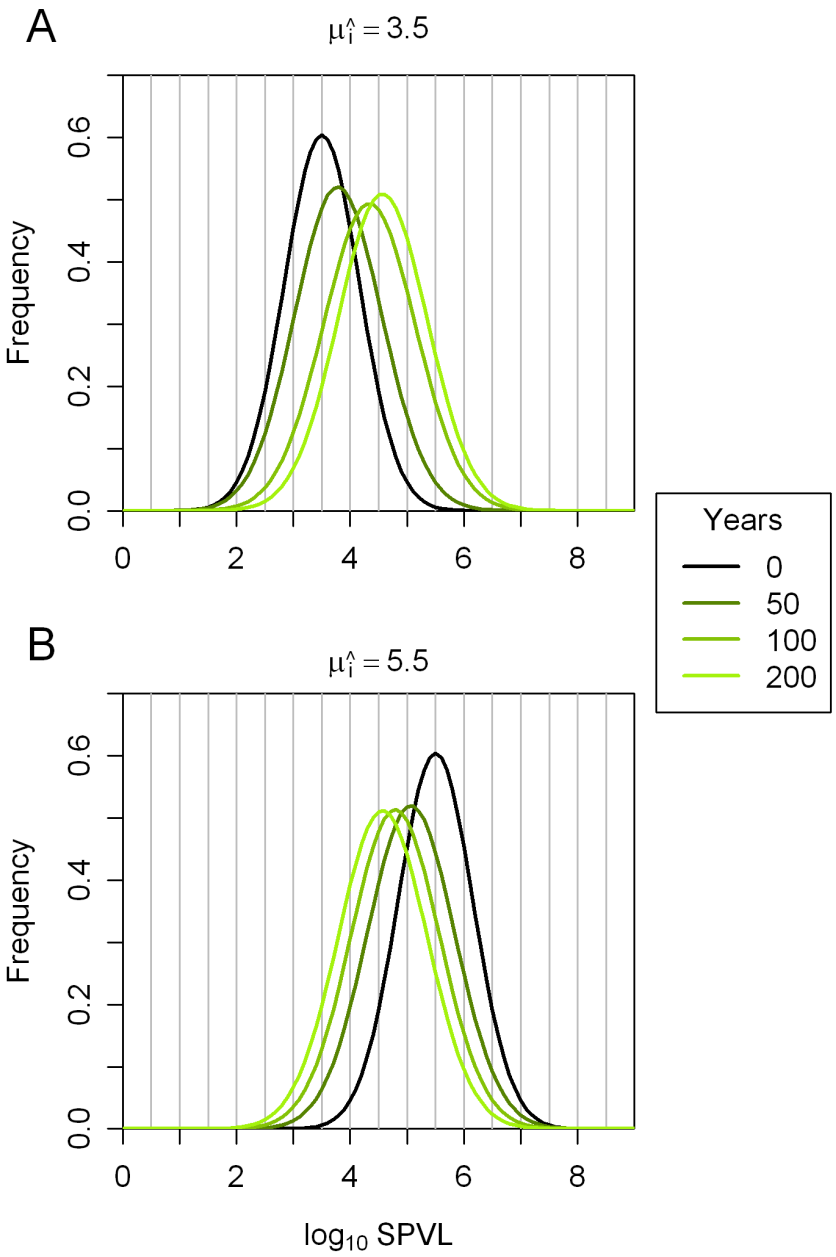


Figure 1.5: The evolving distribution of SPVL. The SPVL distribution evolves in the population over the years since introduction of the founding genotype. Maximum likelihood values from Figure 1.2 were used (σ_M=0.12, σ_E=0.66). The mean log10 SPVL of the founding genotype was a) 3.5 and b) 5.5.

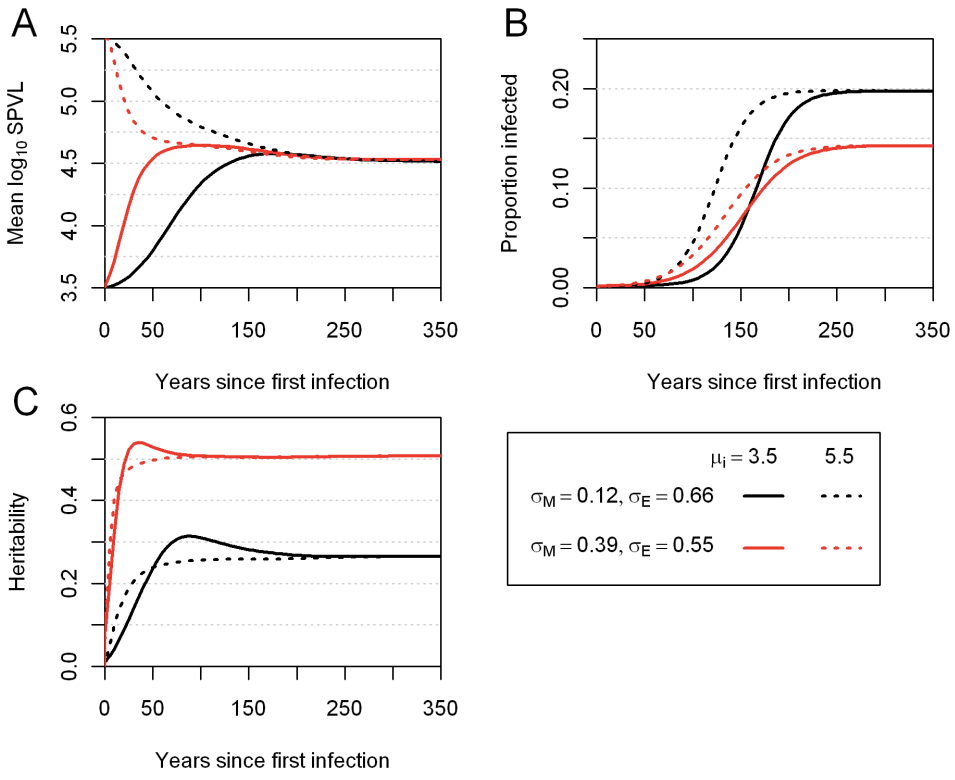


Figure 1.6: Mean log10 SPVL, proportion infected and heritability over time. a) Mean log10 SPVL, b) proportion infected, and c) heritability. The epidemic was run under maximum likelihood parameter (σ_M=0.12 and σ_E=0.66, black), or the combination of parameters with maximum σ_M, consistent with high likelihood (σ_M=0.39 and σ_E=0.55, red). The solid lines show runs in which the founding genotype had μ=5.5, while the dashed lines show runs with a founding genotype with μ=3.5.

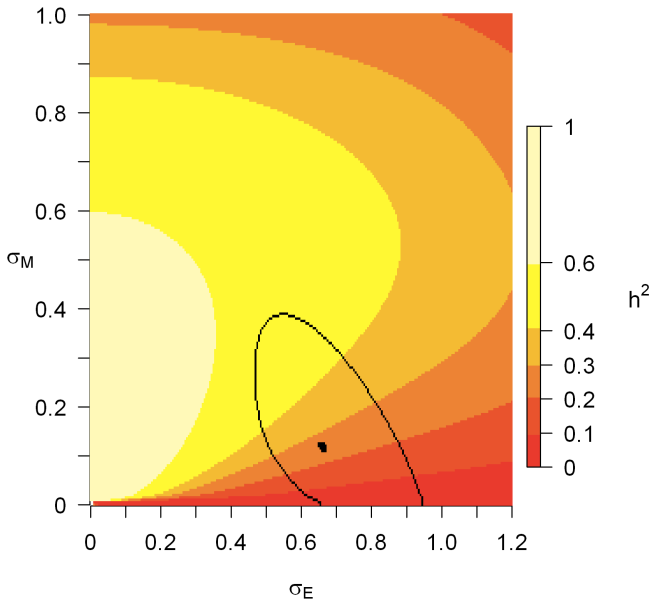


Figure 1.7: Heritability of SPVL measured at equilibrium for each combination of parameters σ_M and σ_E. The black line represents the border of the 95% confidence interval on the maximum likelihood plot, Figure 1.2.

1.5 Discussion

In this chapter, we analysed a multi-strain evolutionary epidemiological model of HIV-1 virulence, and showed that it could adequately reproduce observations on the distribution of viral load and its heritability in transmitting couples (Figure 1.4. This enabled estimation of the proportion of variance in SPVL explained by viral genetic factors (26%, from Equation 1.28), and separately how much these factors change (mutate) from one infection to the next. The best estimate is that virulence changes slowly towards an evolutionary optimum over decades, but faster changes cannot be ruled out (Figures 1.6 and 1.10).

The aim was to develop a simple, parsimonious broad-brush model to understand the principles of HIV-1 virulence evolution in a generalised epidemic using data currently available. Most of the parameters were derived from Sub-Saharan African studies (Table 1.1), suggesting that the model has most direct relevance for this context. This was intended, as this is where most of the adaptation of HIV-1 to the human population has occurred. The parameters determining the curve of survival from disease progression were derived from European data, and since these data predate antiretroviral therapy they are not expected to differ

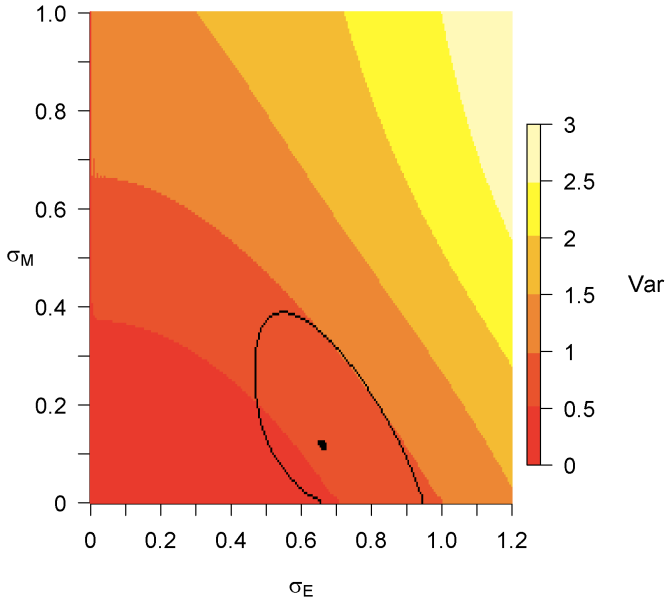


Figure 1.8: Population variance of SPVL measured at equilibrium for each combination of parameters σ_M and σ_E. The black line represents the border of the 95% confidence interval on the maximum likelihood plot, Figure 1.2.

substantially from parameters derived from Sub-Saharan Africa.

There is evidence for evolution of SPVL, but the trend differs between populations [96]. Further work is required to apply the model rigorously to the European context with a view to explaining these trends. More realistic predictions will require more detailed models, and by necessity more data.

Describing the differences between subtypes of HIV-1 is one of the biggest challenges to the model presented here. In this work the VVG is considered on only a single dimension of low-to-high, with single functions describing the relationship between SPVL, infectiousness and duration of asymptomatic infection. HIV-1 subtypes in fact differ in their transmission parameters independently of their differences in SPVL [16, 120, 163]. Subtype A shows a slower disease progression when compared to other subtypes [114]. More specifically, data from the Rakai study showed that subtype A infection results in slower disease progression than subtype D even though the distribution of SPVL is the same [16, 120]. From the same cohort it was shown that subtype A is also more transmissible than subtype D even when viral load and other confounding variables are controlled for in a regression [121]. Subtype A is therefore fitter than D in both duration and transmissibility, and the evolutionary hypothesis would predict the gradual

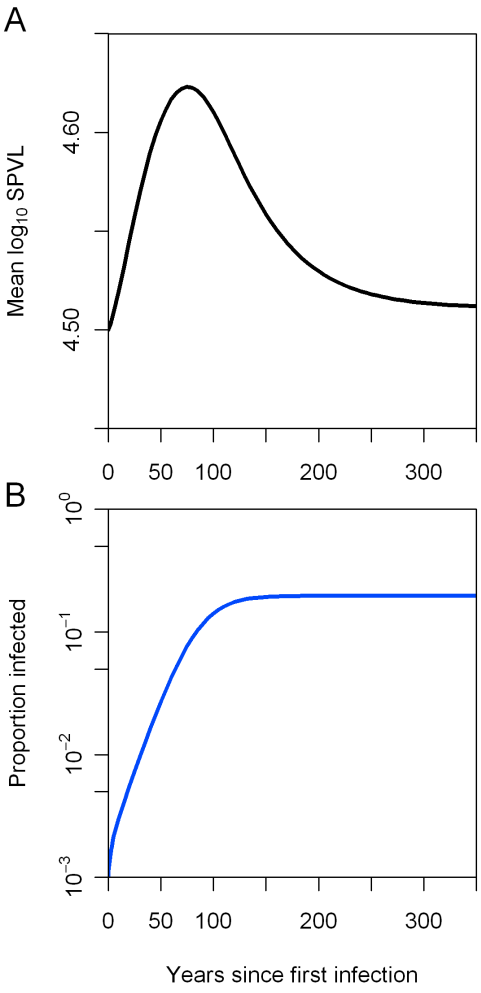


Figure 1.9: The evolution of the mean and the epidemic growth pattern. The founding genotype has μ=4.5, very close to the equilibrium mean, to illustrate changes in the mean in response to growth or shrinkage of the epidemic. a) The evolution of the mean log10 SPVL during b) epidemic growth.

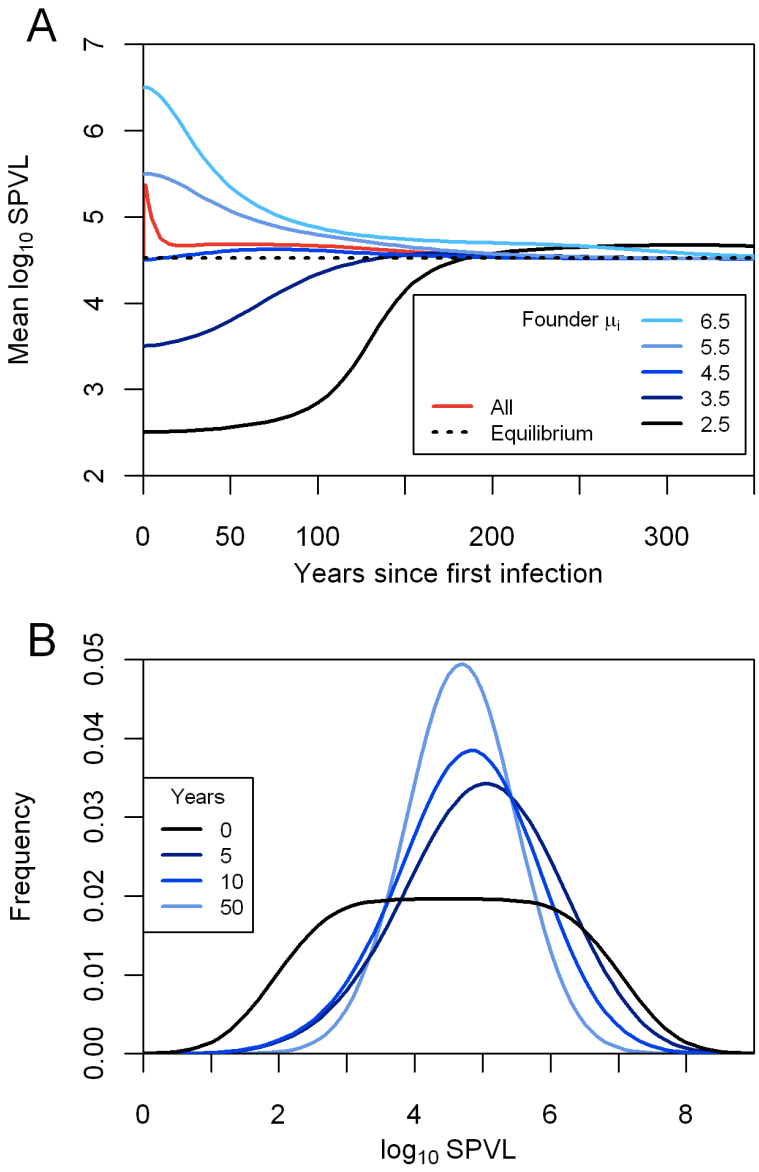


Figure 1.10: Evolution of SPVL from various scenarios of founding genotype. a) Mean log10 SPVL over time for different founding virulences. These range from $\mu = 2.5$ to 6.5 log10 SPVL, All (red) begins with all genotypes from $\mu_i = 2.0$ to 7.0 at equal prevalence, and Equilibrium (dashed black) is the SPVL value to which all scenarios in this figure are evolving. b) Evolution of SPVL distribution from high diversity scenario where all genotypes are equally represented at the start, corresponding to "All" in panel (a). The parameter values for both are maximum likelihood values, σ_M=0.12 and σ_E=0.66.

replacement of subtype D by subtype A, which has been observed in Uganda [42] and Greece [174]. Other noteworthy trends include the dominance of subtype C in southern Africa [218], which may be a result of an extended period of high viraemia in primary infection [164]. Taken together, these findings strongly suggest that HIV-1 virulence can change in ways not fully reflected by SPVL, and thus that more data are needed to identify other appropriate surrogate measures (or determinants) of virulence. More generally, the theoretical challenge is then to explain in terms of these other determinants of infectiousness and survival, how differences in virulence are maintained in different viral subtypes.

There are a number of other directions in which our model could be developed. In this study the mutational variance, the extent to which the viral genotype changed from one infection to the next, was considered independent of the age of infection (AOI). At first, this may seem a paradoxical choice, since mutation which occurs between hosts must be the result of mutations and selection occurring within the infected host. It would reasonable to suggest that the size of between-host mutation is positively related to the AOI, since nucleotide divergence from the founding strain has been shown to occur at a constant rate during infection [207]. If this were the case, the between-host mutation rate would be the same regardless of the generation time and consequently of the virulence of the virus. However, a study of within-host evolution over time found that the rate of divergence from the founding genotype was positively correlated with viral load [19], suggesting that higher virulence infections diverge more rapidly. A model with a mutational variance independent of the AOI allows for this, as a higher virulence virus will have faster transmissions and therefore more between-host mutation events per unit time.

An accurate functional representation of mutational variance as a function of AOI thus requires more detailed understanding than seems currently possible. To resolve this, and for the sake of parsimony, the two effects described above were assumed to cancel each other out, and thus the mutational variance is independent of AOI. To test the sensitivity to this assumption, the model was adapted to include AOI-dependent mutational variance (linearly increasing as a function of time), and the results were qualitatively and quantitatively similar (data not shown).

An additional problem with this model is that the data to which the model is fitted consists of transmission pairs, for most of whom the age of infection at which transmission occurs is unknown. Assuming an AOI-independent mutational variance considerably reduces the complexity of the analysis. There is however little doubt that extending the model to include a more detailed description of within-host processes and also resolving the effects of conflicting selection at the within and between host levels would be enlightening.

The pattern of mutation was modelled as a normal distribution. It may be reasonable to assume that the distribution is negatively skewed because deleterious mutations are much more frequent than beneficial ones, for example in the case of protease gene [176]. However, it is misleading to compare the between-host

mutation process to the mutation of individual viral genomes because deleterious mutations may be counterbalanced by within-host selection for viable viruses and there is no evidence for asymmetry in the net effect.

The host effect in this study was also modelled by a normal distribution which is justified if there are a large number of host effects and they are assumed to each have a multiplicative effect on SPVL. Host effects are known to account for a certain quantity of SPVL variation [196, 53, 111, 64, 79] and a very low estimate of the environmental variance would not be consistent with these studies. The maximum likelihood estimate of σ_E was encouragingly high (σ_E=0.66, Figure 1.2, contributing 71% of the total variance in SPVL. As more is understood about how the host contributes to variation in SPVL, this source of variance may be further decomposed [64].

While the normal distribution modeled for the host effect and mutation pattern may drive the final distribution of SPVL towards normality, this is not prescriptive since the distribution of genotypes can deviate from normal. The methods for model fitting also explored a very wide range of possible variance in the final distribution. The mean of the distribution was more rigidly imposed, but by bootstrapping the original data which was used to fit the dependencies of transmission rate and duration of infection on SPVL, we were able to show that the conclusions were similar even when those data were varied.

The epidemiological component of this model could be made more realistic. The model could for example be structured by age, sex, location, and HLA type, as well as sexual activity and include stochastic effects which are addressed in the next chapter. It is not clear what effect on virulence these heterogeneities would have, but they might help for example explain the persistence of diversity between subtypes and help provide reasonable initial conditions, since a stochastic model could elucidate which viruses are more likely to have started the epidemic. The analysis could be further developed by relaxing the assumption that the SPVL is at an evolutionary optimal equilibrium, though this assumption is consistent with data (Figure 1.4. The mean log10 SPVL and its heritability do not change substantially in the later stages of the epidemic (Figure 1.6a-c), and the mean log10 SPVL of the Ugandan data (4.51) is close to the predicted equilibrium value (4.52), suggesting that even if the observed data do not represent an equilibrium, they represent something close enough to render the maximum likelihood parameter estimations reasonable.

Despite being simple and parsimonious rather than detailed, this model provides a general framework that makes use of the most recent data on the heritability of SPVL, and that can be used to interpret past and predict future trends in virulence. One interesting trend is that the mean log10 SPVL can be observed to increase above the equilibrium value for a short while during the early stages of the epidemic. Epidemic growth is expected to favour a higher virulence than at equilibrium as a result of the cumulative advantage of rapid transmission when hosts are abundant [73, 25]. This is better demonstrated in Figure 1.9c, which shows the evolution of the mean log10 SPVL from a founding virulence

very close to the equilibrium mean. At this level of resolution the temporary spike in virulence can be seen, and this corresponds to the period of epidemic growth. As the number of susceptible individuals grows and the epidemic begins to slow, the virulence decreases in response towards equilibrium as longer-lived genotypes are favoured.

This suggests that if SPVL can evolve at the between-host level then a growing epidemic could select for higher virulence viruses. Bolker et al. [25] model this phenomenon and suggest that the peak of this transient virulence is likely to occur late within the first exponential growth phase of the epidemic, so if this were observable the virulence is likely still to be in this transient state above the equilibrium. Whether this phenomenon has contributed to the recent increase in virulence is difficult to ascertain [96]. A future slight decrease in virulence as an epidemic saturates would provide evidence for this hypothesis, if it could be identified [25]. The optimum virulence could also be shifted by a widespread intervention which affects the nature of transmission such as circumcision, vaccination, or antiretroviral therapy. This model may be used to predict such effects on virulence.

Recently published studies reporting the development of a reasonably effective vaccine [192] and many other potential preventive strategies (see Table 0.1) are promising in the fight against HIV transmission. Hypothetically, a vaccine may offer more protection against lower virulence genotypes and select for more virulent ones, or vice versa.

Antiretroviral therapy during asymptomatic infection reduces transmission rate [14, 40], presumably by reducing viral load [52, 194]. Antiretroviral therapy would therefore modify the relationship between SPVL, transmission and duration of asymptomatic infection, and it is possible to construct hypothetical scenarios that could select for either increased or decreased SPVL. In summary, this model could be used to predict (in general terms) the effects different interventions would have on virulence. These changes are expected to be relatively modest compared to gains obtained by curtailing transmission, but nonetheless some consideration should be given to the possibility of increased virulence and whether it could be mitigated.

1.6 Conclusion

These results support the hypothesis that the distribution of SPVL, and by implication of HIV-1 virulence, can plausibly be explained by selection for increased transmission in populations, though differences between viral subtypes needs to be elucidated in future work. This method disaggregates the effects of viral factors acting to determine SPVL, the effect of mutation (and thus indirectly within-host evolution), and other environmental and host factors. The best estimates indicate a relatively high proportion of SPVL explained by viral

factors (26%), as well as a modest rate of evolution of putative viral virulence factors. Reconciling these findings with data on within-host viral evolution may yet shed further light on the role of viral factors in HIV-1 pathogenesis.

Chapter 2

How virulent was the first strain of HIV?

2.1 Abstract

Background
There are many viral traits which affect the probability that a virus will emerge into an epidemic. Virulence is such a trait, and from its effect on the probability of emergence it may be possible to infer the likely virulence of the strain which made the transfer to humans, given that it has successfully emerged, and the timing of the transfer.
Methods
We explored the effect of unspecified viral genetic factors which influence virulence on the probability that a strain will emerge. An individual-based stochastic simulation with a high risk and a low risk group was used to model transmission within long-term partnerships and transmission by sex external to partnerships. The severity of infection was partly determined by the host and partly by the viral genotype which was heritable between infections and also subject to mutation of virulence factors.
Results
Viruses with intermediate virulence were much more likely to initiate epidemics. We also tested the effect of the various parameters in the model on the probability of emergence and persistence, and how these interacted with virulence. While several other parameters had a large effect, they did not modify the effect of virulence on the epidemic potential of viral strains. The viral virulence evolved towards intermediate values over time.
Conclusion
The preferential emergence of genotypes conveying intermediate virulence supports previous work that these genotypes will also dominate at evolutionary equilibrium, since they are optimised for transmission. This work suggests that, all else being equal, the epidemic is likely to have been founded by a strain of intermediate virulence.

2.2 Introduction

The virulence of a pathogen is important to its fitness. Fitness can be defined by its ability to spread within the host, or by its ability to transmit between hosts, and in this work the concern is with the latter definition. Qualities of a pathogen which determine its transmission fitness, as explored in the previous chapter, may be the same qualities which determine how likely it is to emerge in a new population, but there may also be differences. The term emergence refers to an organism becoming a successful pathogen of humans by infecting a substantial proportion of the population, and by persisting in that population over time. In this paper we explore how qualities of HIV, in relation to virulence, affect the probability of emergence following a single introduction into a human population.

There have been multiple introductions of HIV into the human population from different primate reservoirs. Most infections of HIV globally are HIV-1 group M, but HIV-1 has had two other introductions from SIVcpz in chimpanzees to form groups N and O. There have been eight introductions of HIV-2 from SIVsmm in sooty mangabey monkeys [97]. The seropositivity to another retrovirus, simian foamy virus, in individuals who are in contact with wild non-human primates suggests that zoonotic transmission of retroviruses is frequent enough to support multiple introductions of HIV, especially since the viruses involved had their origins in three different reservoir species [231]. Prevalence of SIVcpz [202] and SIVsmm [36] in wild chimps is variable across populations and ages but it is high enough to create opportunities for introduction.

Humans hunting primates and monkeys are exposed to many HIV-like viruses [231, 51], and it is likely that, as for many zoonoses [179], emergence as an epidemic is a rare event compared to the spillover of the virus. The majority of infections globally are HIV-1 group M, while HIV-1 groups N and O are rare, and similarly in HIV-2, only groups A and B have taken hold, while C-H are very rare in comparison [97]. Why have certain strains in particular been established in the human populations, while others have remained rare?

Many reasons external to the virus may contribute to epidemic success. Stochastic effects may play a large role which depends, for example, on the sexual connectedness of the first infected human. That individual may be part of a highly connected network, enabling the infection to spread from the point of origin, or may be comparatively isolated, so the infection dies out in spite of its epidemic potential. Other factors such as human population density [110], or high malaria prevalence [2] may also increase the probability of successful HIV emergence.

There are several viral characteristics which determine the epidemic potential. Many of these are qualitative and highly specific to the species of primate host. All HIV-1 viruses carry a specific Gag mutation which enables better replication in human CD4 cells [222]. Successful viruses also carry mechanisms to counteract the array of host anti-viral factors, such as APOBEC3G, TRIM5α and tetherin [209]. For example, HIV-1 group M uses the viral Vpu protein to degrade tetherin, which would otherwise prevent new virus from escaping the cell membrane. The Vpu protein in other HIV-1 groups show much less (group N) or no (groups O and P) tetherin degrading activity [203], which may explain their relatively low prevalence. This change in function of Vpu was required for a successful switch to humans: the viral Nef protein in most SIVs is able to degrade tetherins in the corresponding host species, while human tetherin has a deletion which prevents this [108]. HIV-2 group A achieved tetherin degradation through a different route involving gp41 [133].

In this chapter we investigated the effect of viral virulence on epidemic potential. The life history of HIV-1 infections is highly variable, as some progress much more rapidly to AIDS than others, and there is order-of-magnitude variation in viral load between individuals [152]. We propose that a key determinant of

the success of a new virus entering humans is the ability to generate substantial viral replication and be transmitted between hosts, and more directly that the SPVL generated by the first few infections is strongly related to the probability of epidemic emergence of a new virus.

In the previous chapter, a model was analysed which showed that natural selection for lifetime transmissibility can explain the distribution of SPVL we observe today [210]. The rate at which the evolution of the distribution of SPVL occurs is fast enough to have plausibly converged on current values starting from a wide range of virulence, within the available timescale of approximately 100 years since the proposed date of origin of HIV-1. HIV-1 group M, the viral lineage responsible for the current pandemic, is estimated from phylogenetic analysis to have emerged in humans in 1908 (95% confidence interval 1884-1924) [233].

However, the time taken for the distribution of SPVL to evolve to current levels is dependent on the within-host processes that govern the evolution of SPVL over time, and on the virulence of the founding strain, which initiated the whole epidemic. The parameter values that quantify the effect of within-host processes that change viral virulence factors from one infection to the next, and that quantify the influence of non-viral factors in generating variance in SPVL, were separately estimated in chapter 1. Here, we extend this to determine the probability of a virus causing an epidemic as a function of the virulence genotype, and hence to characterise the extent and nature of selection on the virulene characteristics of the founder virus of the HIV-1 epidemic.

The deterministic model in chapter one was extended to include stochastic effects, as these allowed examination of the circumstances in which an introduced viral strain will emerge and persist, or go extinct. Including stochasticity implies that birth, sexual behaviour, transmission and death are governed by random processes. We also aimed to determine whether findings from chapter one on the evolutionary dynamics of HIV virulence were robust to the consideration of stochastic effects, and to heterogeneity in host sexual mixing.

2.3 Methods

The model in the previous chapter was adapted to an individual-based framework, which implies that each individual host is modelled separately rather than considering populations of individuals subject to average effects. This allowed characteristics of individuals to be tracked more easily, in terms of their stage of infection, their sexual relationships and the strain they were infected with.

In the model, HIV-1 infection spreads within a dynamic contact network. Individuals are dichotomously classified into a core group of individuals with a large rate of casual sexual contact, and a non-core group of individuals with a

lower rate of casual contact and who also form long-lasting partnerships. Upon sexual maturation, the probability of entering into the core group is given by pc, and individuals remain in their group for life. The two sexes behave in exactly the same way and are not differentiated in the model. Transmission can occur within discordant long-lasting partnerships in the non-core group, and also due to casual contacts. Casual contacts occur at a higher per-capita rate in the core group, but also occur in the non-core group, and the latter also form long term partnerships.

The number of susceptible individuals, X, begins at X_0, and is made up of core group (X_c) and non-core group individuals (Xn), the latter of which can be either single (X_{ns}) or in a partnership (X_{nr}), although at the beginning of the simulation all non-core individuals are single. The numbers of infected individuals, Y, can be similarly divided with the same subscript notation (Y_c, $Y_n = Y_{ns} + Y_{nr}$).

Partnerships

The non-core group form partnerships at a fixed per-capita rate. The number of new partnerships in a given time step is picked from a binomial distribution 2.1 where D_s is the average time between partnerships, and Δt is the length of the time-step.

$$\text{Number of new partnerships} \sim Binomial\left(\frac{X_{ns}+Y_{ns}}{2}, \frac{1}{D_s}\Delta t\right) \tag{2.1}$$

When this number has been selected, the pairs to join the couple are picked randomly from the single population, $X_{ns} + Y_{ns}$. A breakup time 2.2 is then assigned to this partnership picked from an exponential distribution according to the average duration of partnership, D_r.

$$\text{Breakup time} \sim Exponential\,(D_r) \tag{2.2}$$

Partnerships will also break up if one of the individuals dies before the allotted breakup time. Partnerships can either be serodiscordant or seroconcordant, and in the latter case either both positive or negative. Individuals do not form multiple concurrent partnerships.

Virulence and set-point viral load

When any transmission occurs there is an infector and a recipient of the virus. Upon transmission, the recipient acquires a virus with virulence genotype, μ_x on which their SPVL is based. This is related to the virulence genotype of the infectors virus.

As in chapter one, it is assumed that the SPVL is influenced by the genotype of the virus that initially infects the host, but is not substantially affected by subsequent evolution of the viral quasispecies within the host. For example, humans carrying the HLA-B57 tend to have lower viral loads [7], but viral mutants which escape this response pay a fitness cost [156] and these mutations are not associated with increased viral load [57].

Thus, the log10 SPVL, V_x, is constant for the infected individuals lifespan, and is drawn from a normal distribution with mean μ_x, and variance ${\sigma_E}^2$ (distribution 2.3), which is a measure of the amount of variation in SPVL determined by factors external to the virus, principally host factors.

$$V_x \sim Normal\left(\mu_x, {\sigma_E}^2\right) \tag{2.3}$$

Evolution in the viral genotype can occur during the course of infection, eventually resulting in change in the VVG from one infection to the next. The new log10 viral genotype, μ_y, is picked from a normal distribution (2.4) centred on the infectors log10 virulence genotype, μ_x, with variance ${\sigma_M}^2$ which is measure of the per generation mutation rate of the virulence genotype.

$$\mu_y \sim Normal\left(\mu_x, {\sigma_E}^2\right) \tag{2.4}$$

This approach to modelling viral evolution matches that used in chapter one, which has the advantage that it allows evolution of SPVL in both directions, with a specified rate of mutation per generation. As it is normally distributed, the probability of an increase or decrease in virulence is equal. This seems like a reasonable if crude first approximation to the outcome of the complex interplay between selection for increased intrinsic replicative capacity and escape from a diverse immune response that shapes the viral quasispecies in the infected host.

Stages of infection

Once infected, the individual progresses through three stages of infection: primary (including acute and early infection), asymptomatic (chronic or latent infection), and disease (late-stage pre-AIDS/AIDS), before death occurs. Since the model is primarily intended to describe evolution in the pre-treatment era, the effects of diagnosis or treatment were ignored. The model for the duration of infection and rate of transmission and how these depend on SPVL was based on previous work, from which many of the parameter values were also taken [73], as in chapter one.

The rate of transmission for the primary and disease stages of infection are β_P and β_D, respectively, and are the same for all individuals. The mean duration of these stages are also the same for all individuals, and are $\bar{D}_P$ and $\bar{D}_D$ for primary and disease stage respectively, but for each individual the actual durations are picked from exponential distributions (2.5 and 2.6).

$$D_P(y) \sim Exponential\left(\bar{D}_P\right) \tag{2.5}$$

$$D_D(y) \sim Exponential\left(\bar{D}_D\right) \tag{2.6}$$

The rate of transmission during asymptomatic infection for an individual y is determined by the individuals SPVL value V_y according to a Hill function 2.7 with parameters β_{max}, β_{50}, and β_k that represent, respectively, the maximum rate of transmission during asymptomatic infection, the SPVL value at which infection rate is half the maximum, and a shape parameter. This gives a sigmoidal function of rate of transmission on log10 SPVL.

$$\beta_A(V_y) = \frac{\beta_{max}(10^{V_y})^{\beta_k}}{(10^{V_y})^{\beta_k} + \beta_{50}^{\beta_k}} \tag{2.7}$$

The expected duration of asymptomatic infection for the individual with SPVL V_y is also determined by a Hill function 2.8 with parameters D_{max}, D_{50} and D_k. This gives a reverse sigmoidal function of duration of asymptomatic infection on log10 SPVL.

$$\bar{D}_A(V_y) = \frac{D_{max}{D_{50}}^{D_k}}{(10^{V_y})^{D_k} + D_{50}^{D_k}} \tag{2.8}$$

For a given individual, the length of their asymptomatic infection is based on his/her expected duration, but is picked from a survival curve with that average duration. This is done by choosing the point on the survival curve (equation 2.9=) which corresponds to the random variable U which is picked from a uniform distribution between 0 and 1. The shape parameter for the survival function is given by ρ, and Γ is the standard gamma function.

$$D_A(y) = \bar{D}_A(V_y)\frac{(-\ln(U))^{\frac{1}{\rho}}}{\Gamma(1+\frac{1}{\rho})} \tag{2.9}$$

At the point of infection, the time at which an individual will progress through each of the stages is determined at the start according to $D_P(y)$, $D_A(y)$ and $D_D(y)$ and these progression events occur in continuous time.

Transmission within partnerships

Heterogeneity in sexual behavioural with a highly active core group and a less active group, and a degree of assortativity between them, is a model structure that has a long tradition [80]. Modelling work has emphasised the additional importance of multiple concurrent partnerships on the spread of HIV [157],

particularly when enhanced infectivity during primary infection is taken into account [59]. However, a casual transmission process occurs simulateneously with the partnership transmissions (such as in [184]. A wide range of parameter values which determine the partnership dynamics, the contact rates for casual transmission, as well as the assortativity and size of the core group, are considered. This is judged to create a sufficiently flexible range of sexual mixing scenarios that the inclusion of explicit concurrent partnerships is not required to address the questions in this analysis.

When a partnership forms, it is concordant if both individuals are uninfected, or both infected. It is discordant if it consists of an infected and an uninfected partner. A partnership may also become discordant if both partners were initially uninfected and one is subsequently infected by casual transmission from an infector outside the partnership. Upon initiation of a discordant partnership, or the progression to the next disease stage of the infected individual in a discordant partnership, a time for future transmission is scheduled. The time to transmission from infected partner y to susceptible partner x, t_{xy}, is picked from an exponential distribution (2.10) according to the rate of transmission, β_y, of partner y, at that time.

$$t_{xy} \sim t + Exponential\left(\frac{1}{\beta_y}\right) \tag{2.10}$$

This transmission will occur only if it occurs before the next disease progression, breakup or death of the infected partner, or the uninfected partner acquires infection casually.

Casual transmission

In order to incorporate the effect of high risk individuals and the possibility of infection occurring by casual contacts with individuals also involved in long-term partnerships (e.g. extra-marital contacts), casual transmission is incorporated which can occur between any pair of individuals in the population. The core group have a contact rate for casual transmission, λ, while the lower contact rate for the non-core group is given by $\lambda\kappa$, where $0 < \kappa < 1$. An assortativity factor, ϵ, determines the extent to which individuals preferentially contact members of their own sexual behaviour group rather than at random between the two groups.

The probabilities, p_{fg}, of an individual in the group f being infected by a member of group g, is determined by equation 2.11.

$$\begin{aligned} p_{fg} = \Delta t \cdot\ & \text{contact rate}_f \cdot\ \text{contact rate}_g \\ & \cdot\ \text{total transmission rate}_g \cdot\ p(\text{possible contacts}_f^{-1} | g) \end{aligned} \tag{2.11}$$

The contact rates are fixed (equation 2.12). The total transmission rate is the sum of all the infectivity values of individuals in group g (equation 2.13).

$$\begin{pmatrix} \text{contact rate}_n \\ \text{contact rate}_c \end{pmatrix} = \begin{pmatrix} \lambda\kappa \\ \lambda \end{pmatrix} \tag{2.12}$$

$$\text{total transmission rate}_g = \sum_{y_g} \beta_{y_g} \tag{2.13}$$

The reciprocal of the number of possible contacts for group f is the reciprocal of the number of all contacts made in the population, multiplied by the probability of random mixing $(1-\epsilon)$ (equation 2.14), and if $f = g$ this is added to the reciprocal of the number of possible contacts in group f, multiplied by the probability of assortative mixing, ϵ (equation 2.15).

$$p(\text{possible contacts}_f^{-1} | f \neq g) = \frac{1-\epsilon}{\lambda\kappa(Y_n + X_n) + \lambda(Y_c + X_c)} \tag{2.14}$$

$$p(\text{possible contacts}_f^{-1} | f = g) = \frac{1-\epsilon}{\lambda\kappa(Y_n + X_n) + \lambda(Y_c + X_c)} + \frac{\epsilon}{\text{contact rate}_f(Y_f + X_f)} \tag{2.15}$$

For each susceptible individual in group f there are then three possibilities: he/she is infected by the non-core group (with probability p_{fn}), by the core group (with probability p_{fc}), or he/she is uninfected (with probability $1 - p_{fn} - p_{fn}$).

When the infection status of a given individual has been selected as well as the which group he/she has been infected from, the individual is then assigned an infector from the appropriate group, with a probability proportional to the rate of transmission of all individuals in the transmitting group.

Deaths and births

New individuals are recruited into the sexually active population at a constant rate, α in each time step Δt according to a Poisson distribution (2.16).

$$\text{New individuals} \sim Poisson\,(\alpha\Delta t) \tag{2.16}$$

Individuals leave the sexually active population either at the end of the disease stage of infection, or because of background mortality which occurs in each time step at per capita rate ω. We refer to this as death, but this encompasses other

reasons for leaving the sexually active group. This background death is also a Poisson distribution (2.17).

$$\text{Background deaths} \sim Poisson\left((Y_c + X_c + Y_n + X_n)\omega\Delta t\right) \tag{2.17}$$

In order to maintain the initially susceptible population at size X_0, the value of α is calculated according to equation 2.18.

$$\alpha = \frac{X_0}{\alpha} \tag{2.18}$$

Initial infection

The epidemic is initiated after a burn-in period during which time the non-core individuals form partnerships. This was set at ten years after it was established that this time period is sufficient for the numbers of partnerships to stabilise. At the maximum duration between partnerships, 2 years, more than 99% of individuals, on average, would have formed at least their first partnership within this time. Following this burn-in, a number of individuals, Y_0, are infected at random. This value is normally set at one as all epidemics from a single introduction will begin with a single infection. The founding individual(s) is/are infected with a strain of virulence genotype μ_0, which is specified and is the principal variable under study. The founders belong either entirely to the core group or entirely to the non-core group, and the founding group is chosen by probability P_c.

Reporting

An epidemic is considered to have emerged successfully if more than 2% of the population are infected at any point in the 100 years since the infection was introduced, and we refer to this as a threshold epidemic. The value was chosen as it was the minimum prevalence which was not already reached by the first infection in the smallest modelled population of 100 individuals. It is considered to have persisted if there are still more than one infected individual at the end of the period of 100 years, and we refer to this as a survival epidemic. The 100 year period was chosen as the approximate time elapsed since the origin of HIV-1 group M [233]. Each set of variables is run for an appropriate number of repeats, R, which is normally 10,000. For a given variable set the probability of starting an epidemic is calculated by how many were successful, R_E, divided by the number of trials R (equation 2.19). The confidence intervals were calculated according to the Clopper-Pearson method [39] to allow for more accurate limits at low or high probabilities of success. This is shown in equation 2.20, where $Beta^{-}1$ refers to the inverse of the beta cumulative density distribution, with

the probability and two shape parameters given subsequently in the parentheses.

$$\text{Proportion successful} = \frac{R_E}{R} \tag{2.19}$$

$$\begin{aligned} \text{lower bound} &= \begin{cases} 0 & \text{if} R_E = 0 \\ 1 - Beta^{-1}(0.975, R - R_E + 1, R_E) & \text{if} R_E > 0 \end{cases} \\ \text{upper bound} &= \begin{cases} 1 - Beta^{-1}(0.025, R - R_E, R_E + 1) & \text{if} R_E < R \\ 1 & \text{if} R_E = R \end{cases} \end{aligned} \tag{2.20}$$

Parameter Sampling

Parameter values that determine the duration and rate of transmission, and the evolution of SPVL, were taken directly from the literature (Table 2.1), as this study should be viewed in the context of those data and parametric inferences. The other parameter values that determine sexual mixing behaviour, demographics and the founding virulence were under investigation in this study (Table 2.2).

To find appropriate input parameters, a systematic search in parameter space was performed. With the exception of Y_0, which was assumed to be 1, values for each input parameter were drawn using Latin Hypercube Sampling using the *tgp* package in R [87, 88, 217] according to their plausible ranges as shown in Table 2.2.

A single simulation was performed for each of 100,000 combinations of parameter values. Combinations which yielded prevalence values between 15 and 30% were selected, since this was considered to be congruent with a generalised HIV epidemic, and we examined the distribution of parameter values within this successful subset.

A univariate analysis was then performed to test the effect of each parameter on the probability of emergence. Each parameter was tested for a range of values at regular intervals across its range. For each value 10,000 simulations were performed, and the other parameter values in each simulation were sampled randomly from the sets of values which were successful in the parameter estimation step. This enabled the effect of each parameter to be examined against a varied background of suitable parameters.

We investigated how the effect of virulence on probability of emergence is modified by the other parameters. For each of the parameters excluding μ_0, three values were selected which spanned their range, and for each value we tested the effect of the range of μ_0 values (1.5, 2.0, 2.57.5). For each value of the parameter under study and each value of μ_0, 1000 simulations were performed, with the

Table 2.1: Parameter values taken from the literature.

Symbol	Description	Value	Reference
β_P	Unadjusted rate of transmission during primary infection.	2.76 per year	[103]
β_D	Unadjusted rate of transmission during high transmission disease stage.	0.76 per year	[103]
$\bar{D}_P$	Average duration of primary infection	0.25 years	[103]
$\bar{D}_D$	Average duration of high transmission disease stage	0.75 years	[103]
β_{max}	Average rate of transmission during asymptomatic stage for highest SPVL	0.317 per year	[73]
β_{50}	SPVL at which infectiousness is half maximum	13,938 copies per ml	[73]
β_k	Hill coefficient: steepness of increase in infectiousness as a function of SPVL	1.02	[73]
D_{max}	Average duration of asymptomatic stage for lowest SPVL	25.4 years	[73]
D_{50}	SPVL at which duration of asymptomatic infection is half maximum	3,058 copies per ml	[73]
D_k	Hill coefficient: steepness of decrease in duration as a function of SPVL	0.41	[73]
ρ	Shape coefficient for the Weibull survival distribution	3.46	[73]

other parameters taken at random from sets of values which were successful in the parameter estimation step.

We attempted to reproduce the results from chapter one and previous work [210] in which we investigated the evolution of SPVL over time under different mutation rates. In this chapter we simulated epidemics over 500 years with the mutational and environmental standard deviation (σ_M and σ_E) set at 0.12 and 0.66 respectively, in accordance with their maximum likelihood estimates in chapter one 1.2. We also tested an alternative scenario representing the highest mutation rate, σ_M=0.39, which still fell within the 95% confidence boundary as estimated in the previous chapter, with σ_E=0.55 as the corresponding environmental standard deviation.

Table 2.2: Parameters under investigation.

Symbol	Description	Range
Y_0	Number of original infections.	1-10 people
X_0	Initial population size.	$10^2 - 10^5$ people
ω	Background rate of leaving sexually active population	0.001-0.1 person^{-1} year^{-1}
$\bar{D}_r$	Average duration of partnership	0-5 years
$\bar{D}_s$	Average time between partnerships	0-2 years
P_c	Probability of entering core group upon sexual debut.	0-1
λ	Core group contact rate.	0-1
κ	Ratio in contact rate between non-core and core group.	0-1
ϵ	Assortativity factor: probability that an individual prefers contacts within their own group, rather than at random.	0-1
σ_E	Environmental standard deviation of log10 SPVL.	0-1 log10 SPVL
σ_M	Mutational standard deviation of log10 SPVL.	0-1 log10 SPVL
μ_0	Mean log10 SPVL of founding genotype.	1.5-7.5 log10 SPVL

2.3.1 Basic reproduction number, R_0

The most common and useful measure of epidemic potential is the basic reproduction number R_0, defined as the average number of individuals infected by a typical index case during the exponentially growing phase of an epidemic and when the population is fully susceptible, measured over the whole course of their infectiousness [50]. It is a crucial measure in determining the epidemic potential of a pathogen because if $R_0 > 1$ the incidence is expected to increase, or decrease if $R_0 < 1$.

Due to the complexity which accompanies the inclusion of both long term partnerships and casual mixing in our model, deriving R_0 analytically for a given strain in this model is, to the best of our ability, an intractable problem. We have therefore taken a different approach and attempted to approximate it numerically. This can be done relatively intuitively in an individual-based model by measuring the number of secondary infections which a typical infected individual produces when the population is effectively susceptible, and averaging over multiple simulations.

To avoid any saturation of the susceptible population with infected individuals, the ideal individual in whom to measure R_0 would be one who was infected as early as possible, especially since the population of interest is small. One

approach would be to measure the number infected directly by the first case in the population.

However, this individual deviates from being a typical representative of infected individuals in the outbreak for two reasons. The first reason is that a typical infected individual in the non-core group may have been infected by his/her partner, and therefore necessarily spends some time between transmission and break-up in a relationship with an infected person, to whom he/she cannot transmit. The first case is not able to be in a relationship with an infected person, so estimating R_0 from that individual will be an overestimate.

The second reason is that a typical infected individual should be in the core group according to the probability that a core group individual becomes infected, whereas the first individual is in the core group according to their proportion in the population. Since the core group have a higher contact rate, a typical infected individual is more likely to be in the core group than their proportion in the population suggests. The lack of population structure in this model may exacerbate this problem.

We address this problem by using the second case, or the first infectee, as a typical case, and measuring the number of his/her transmissions which we refer to as R_1. This avoids the first problem, since he/she is an infectee. It also mediates the second problem, because by virtue of having become infected, he/she is selected on the basis of the likelihood of becoming a case.

It is likely that individuals infected later in the epidemic, for example in the third or fourth generation, are more likely to be in the core group and therefore more typical, so R_1 may be an underestimate. However, this would depend on the assortativity, the relative contact rates and the partnership dynamics, so it would be difficult to say in advance which generation would be best in which to measure the reproductive number. Furthermore, measuring in later generations is conditional on the epidemic reaching that many generations, so a large proportion of important simulations would have to be ignored. For these reasons, as well as for parsimony and computational ease, I use the R_1 measure as a proxy for R_0, but with these caveats as discussed.

When mixing between the core and non-core groups is not too assortative, it is approximately, but not exactly equal to R_0; in all cases it is a similar cumulative measure of transmissibility, and a monotonic function of R_0. It is expected to be an underestimate of R_0, since the first infectee is less likely to be in the core-group than a typical infectee, but the extent of the discrepancy is hard to predict.

2.4 Results

The principal aim was to investigate how the viral genetic determinants of virulence affect the probability of emergence in a finite population, and how the other characteristics of the population interact with this relationship. We did not consider specific virulence determinants in the viral genotype, but rather the viral contribution to virulence as a numerical quantity. The secondary aim was to show that the evolution of SPVL towards intermediate values as shown in chapter one was robust to consideration of stochastic effects and heterogeneity in sexual mixing patterns.

We found that the probability that an HIV-1 virus, entering the human population, can emerge and launch an epidemic, is strongly dependent on the virulence of the founder virus, with intermediate virulence genotypes being much more likely to produce an epidemic. Combinations of parameters were selected if they ended in an epidemic of between 15 and 30% prevalence after 100 years after infection was introduced, and 2.8% of the sets of parameter values fulfilled these criteria. The successful combinations were plotted in a bivariate fashion to examine whether or not there were correlations between these values (Figure 2.1).

The initial virulence, μ_0, had a large impact on the success of a parameter combination, as the points in this figure are not evenly distributed. This was also true of the core group contact rate, λ, and the ratio between the non- core and core group contact rate, κ. The distributions of these three parameters are shown in 2.2. The middle row represents the approximate posterior distribution of the parameter in question, as produced by the ABC method. The two other rows are included in order to differentiate between values which are excluded because they produce epidemics which are too small (top row) or too large (bottom row).

A univariate analysis was performed to show the effect of individual parameters on the probability of an epidemic (Figure 2.3). The three parameters shown in Figure 2.2 had a marked effect, but several other parameters also affected the probability of emergence, while some had an almost negligible effect (Figure 2.3), i, j and k). The effect of each parameter in most cases was very similar whether the epidemic was measured by survival of the infection at 100 years (Figure 2.3) or by a prevalence of >2% occurring at any time within 100 years (Figure 2.4). We hypothesized that the increase in probability of an epidemic as Y_0 increased could be explained by a simple probability model of the probability of an whereby the probability of emergence when $Y_0 > 1$ could be explained by a function of the probability at $Y_0 = 1$ (Figure 2.5)

We also examined in detail the relationship between the initial virulence, μ_0 and the probability of emergence, in the context of each of the other parameters. For each parameter three values were tested and the effect of virulence on the probability of the epidemic surviving at 100 years was plotted for each value

(Figure 2.6). The same is also shown but for the probability of the epidemic crossing the 2% prevalence threshold (Figure 2.7). For many of the plots the coloured lines are superimposed suggesting that this parameter has very little effect on the probability of emergence. Some parameters affected all values of μ_0 in a uniformly multiplicative (e.g. Figure 2.6), g) or additive (e.g. Figure 2.6, h) fashion, but interestingly some affected different values of μ_0 differently (e.g. Figure 2.6, d).

Since the probability of emergence is associated with the basic reproductive number, we examined the relationship between probability of emergence and the reproduction number of the first infectee, R_1. From the initial combinations selected by Latin Hypercube Sampling, 1000 combinations of parameters were selected, and for each 100 simulations were performed to give, for each combination, an average R_1 and a probability of emergence, in order to examine the relationship between the two. They exhibit a strong correlation, but not an apparent threshold at R_1=1 (Figure 2.8, a and c), as a result of its being an approximation of the true basic reproductive number. We investigated the effect of different mutation rates on the relationship between R_1 and the probability of emergence, in order to compare this work with that of Antia et al. [10], who modeled the emergence of mutating pathogens under stochastic effects in the context of R_0, and we found that under our model the mutation rate had little effect (Figure 2.8, b and d).

Finally, Figure 2.9 is a reproduction of the result from the deterministic model in the previous chapter (Figure 1.10a). This was done to show that the virulence will evolve over time towards intermediate values even when stochastic effects and complex sexual mixing patterns are included. Figure 2.9 shows the evolution of the virulence over time, averaged over 10,000 simulations, and for different values of σ_M and σ_E. The figures show a clear trend of evolution towards intermediate SPVL regardless of the initial value of μ_0, and this occurs faster at higher values of σ_M.

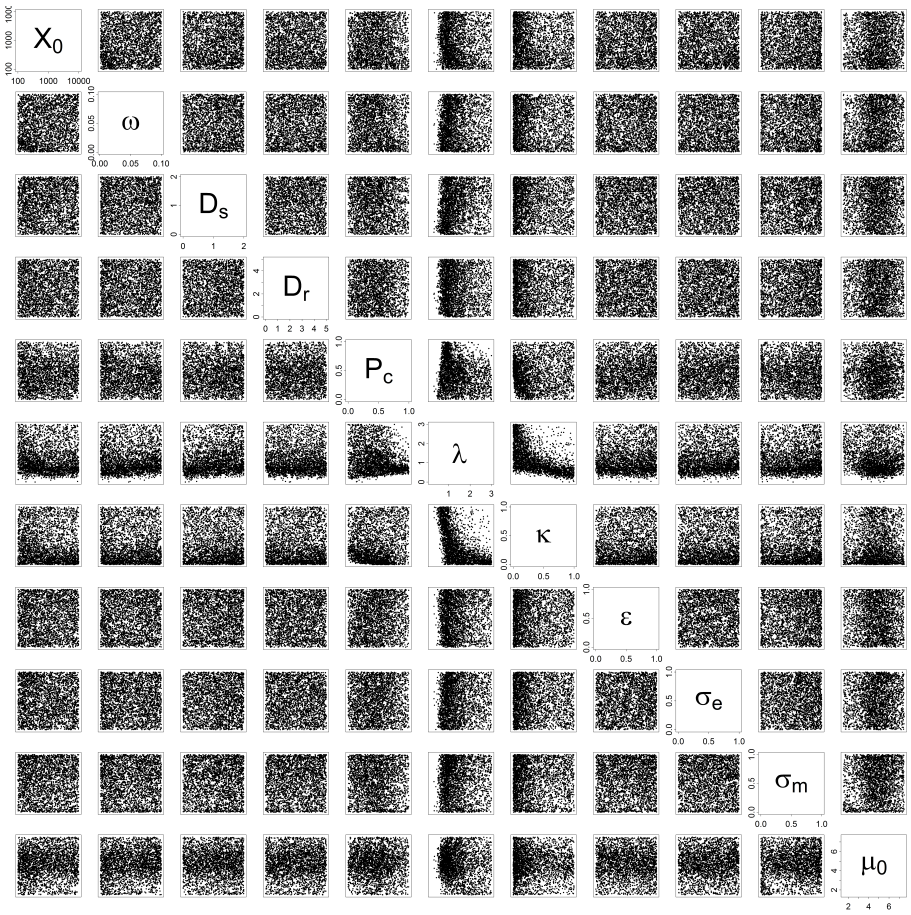


Figure 2.1: Parameter estimation. The 100,000 initial sets of values were selected by Latin Hypercube sampling from the ranges in Table 2.2), excluding Y_0 which was set at one. Each combination was simulated, and 2789 of these produced an epidemic with a final prevalence between 15 and 30%. These points are plotted in this figure, with each panel showing the same set of simulations. In a particular row, the same parameter is on the y-axis, while in each column the same parameter is on the x-axis. The parameters signify: X_0, initial population size; ω, the death rate; D_s, the average time between infections; D_r, the average duration of relationships; P_c, the proportion entering the core group; λ, the core group contact rate; κ, the ratio between non-core and core group contact rates; ϵ, the assortativity; σ_E, the environmental standard deviation; σ_M, the mutational standard deviation; μ_0, the virulence of the founding strain.

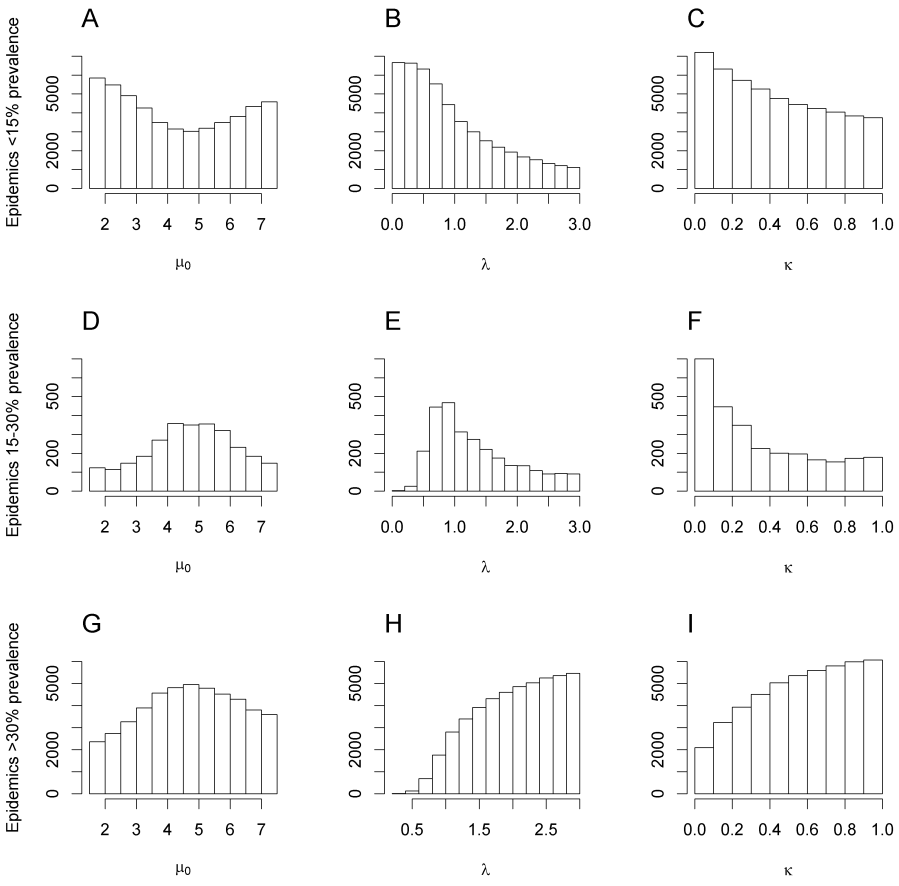


Figure 2.2: Distribution of key parameters. The distribution of the parameters within simulations with a final prevalence of <15% (a, b and c), 15-30% (d, e and f) and >30% (g, h and i). The parameters are the initial virulence, μ_0 (a, d and g), the core group contact rate, λ (b, e and h), and the ratio of non-core to core group contact rate, κ (c, f and i).

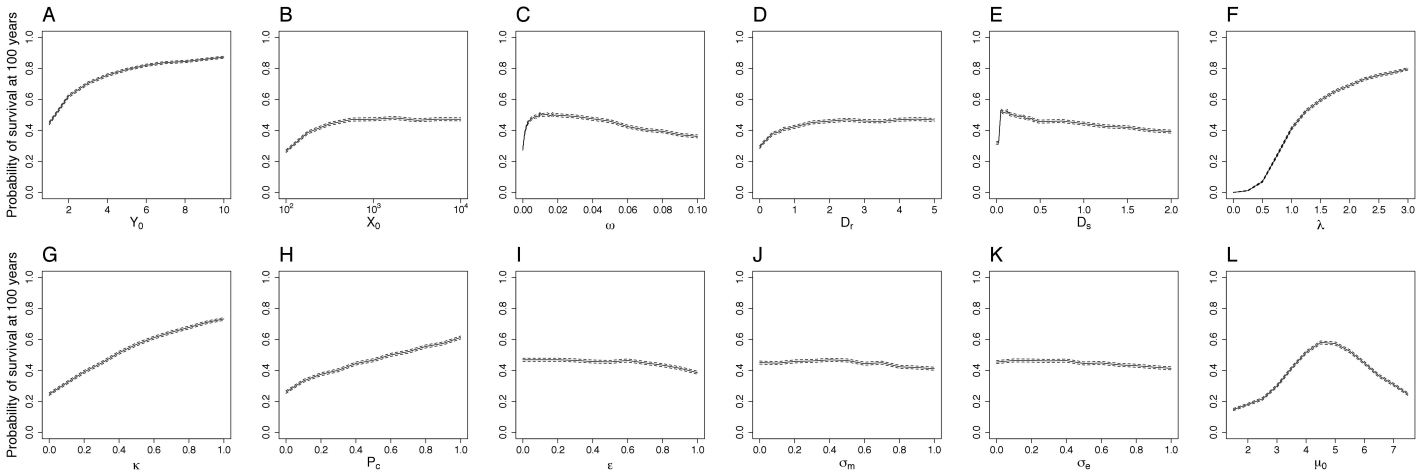


Figure 2.3: Univariate analysis of the effect of each parameter on probability of emergence, measured as the proportion of simulations in which there are at least two infected individuals surviving at the point 100 years after the initial infection. For each value of each parameter tested, 10,000 simulations were performed. The dotted lines represent 95% confidence intervals. The parameters under investigation in each figure are are a) number of founders, b) population size, c) death rate, d) duration of relationships, e) time between relationships, f) core group contact rate, g) non-core to core group contact rate ratio, h) proportion entering core group, i) assortativity, j) mutational standard deviation, k) environmental standard deviation, l) initial virulence. The range of X_0 values (b) was capped at 10^4 because an increase after this point did not affect the probability of emergence (data not shown) but became very computationally intensive. The death rate, average duration of partnership and average time between partnership (c, d and e) were more densely sampled at lower values to examine the outcome as they tended to zero.

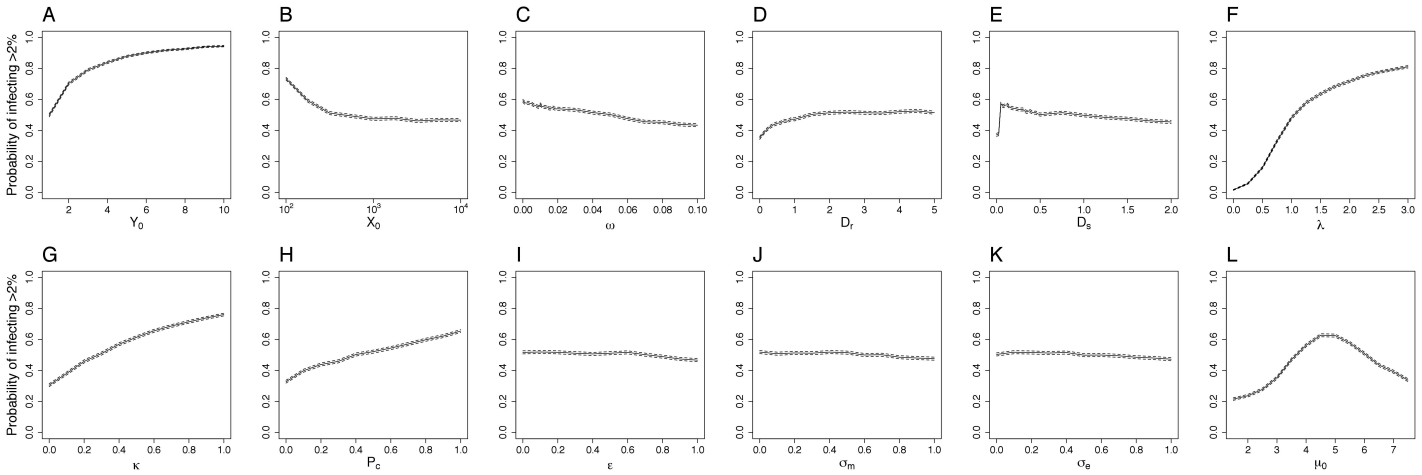


Figure 2.4: Univariate analysis of the effect of each parameter on probability of emergence, measured as the proportion of simulations in which the infection reaches at least 2% prevalence within 100 years after the initial infection. For each value of each parameter tested, 10,000 simulations were performed. The dotted lines represent 95% confidence intervals. The key parameter in each plot corresponds to that in Figure 2.3.

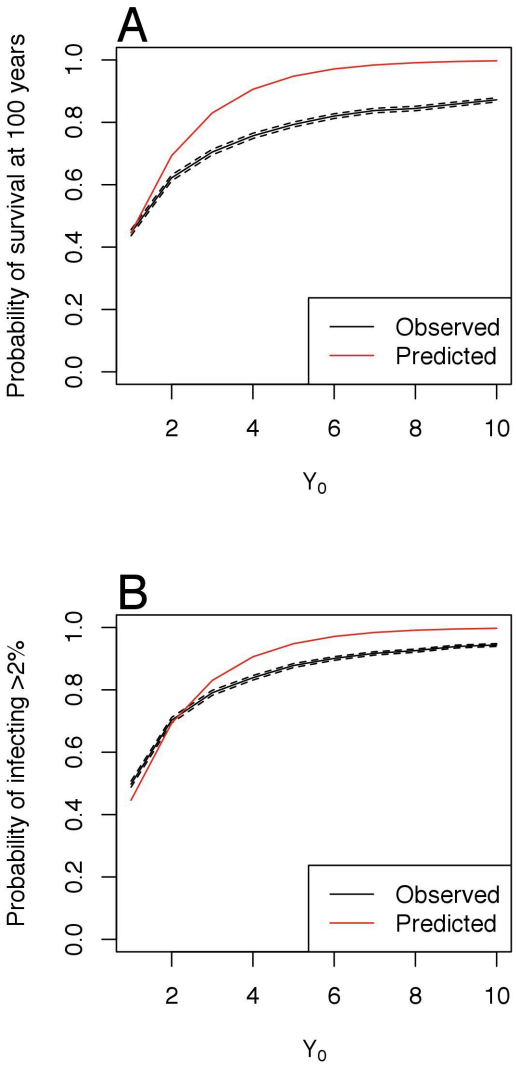


Figure 2.5: The increase in the probability of an epidemic by a) survival at 100 years or b) crossing the 2% threshold, as the number of founders increases. The probability observed in the model (black lines), is compared with the increase in epidemic probability based on Y_0 independent trials. The probability of emergence with trials. The probability of emergence with Y_0 founders, p_{Y_0}, is given by $1-(1-p_1)^{Y_0}$ where p_1 is the probability of an epidemic when $Y_0 = 1$.

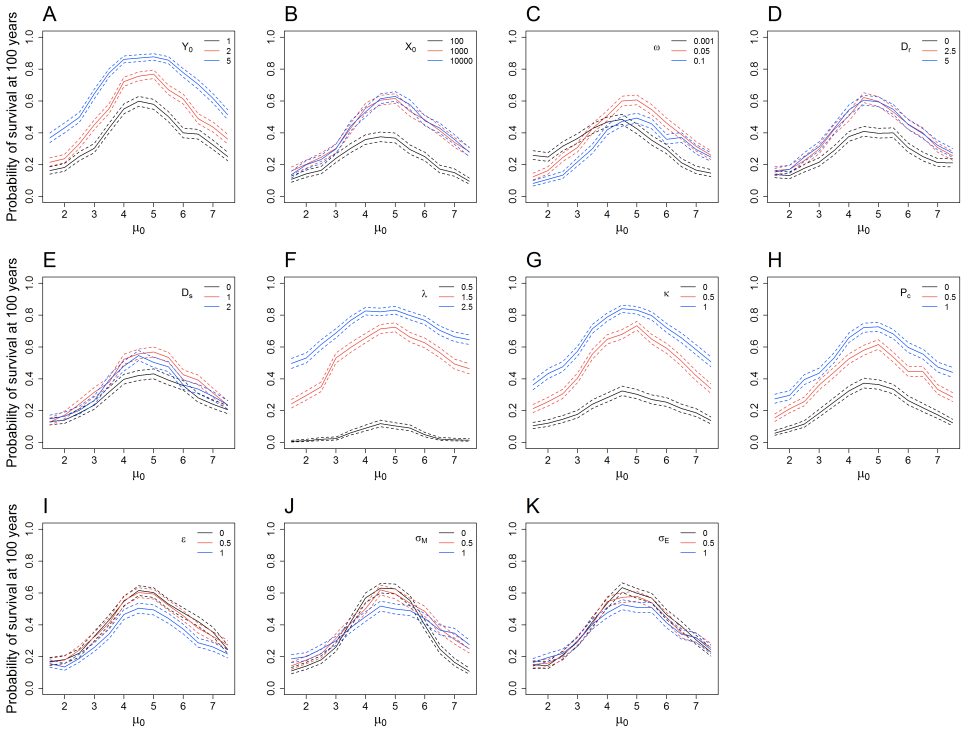


Figure 2.6: The interaction of each parameter with the effect of virulence on emergence. Emergence is measured by survival of the infection after 100 years. For each value of the parameter under scrutiny, indicated by the legend, a range of values of μ_0 were tested, and for each combination 1000 simulations were performed. The parameters under investigation in each figure are a) number of founders, b) population size, c) death rate, d) duration of relationships, e) time between relationships, f) core group contact rate, g) non-core to core group contact rate ratio, h) proportion entering core group, i) assortativity, j) mutational standard deviation, k) environmental standard deviation.

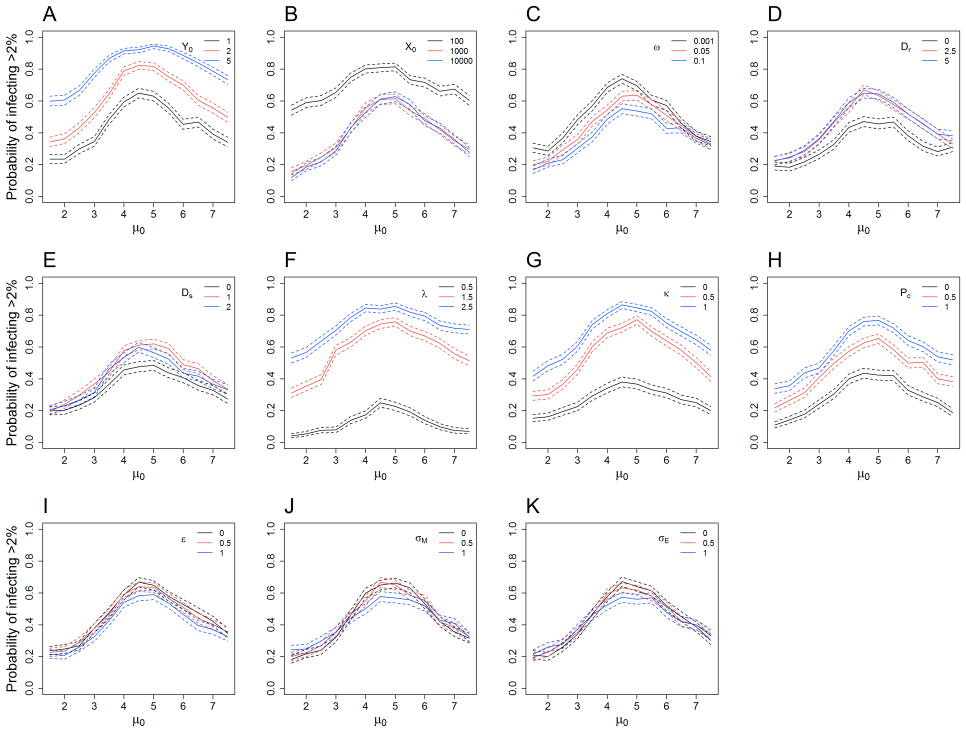


Figure 2.7: The interaction of each parameter with the effect of virulence on emergence. Emergence is measured by the infection reaching 2% prevalence at any point within 100 years. For each value of the parameter under scrutiny, indicated by the legend, a range of values of μ_0 were tested, and for each combination 1000 simulations were performed. The parameters under investigation in each figure correspond to those in Figure 2.6.

2.5 Discussion

In this work we explored the effect of founder selection on HIV-1 viral virulence. Specifically, we have studied the effect of viral virulence on the per-introduction probability of emergence of HIV infection, and how other parameters governing population and sexual behaviour interact with that relationship. We also studied the effect of between-host selection on the evolution of virulence over time under stochastic fluctuations.

The principal result is that a virus with intermediate virulence (μ_0) is more likely to found an epidemic than one with a high or low virulence genotype (Figure 2.3 and 2.4, l), and this trend occurs regardless of large changes in the other parameters (Figure 2.6 and 2.7). This confirms previous work which suggested that these intermediate viruses are fitter and will therefore be naturally selected in the population due to their higher basic reproductive number (R_0) [73].

The apparent strength of founder selection at the origin of HIV-1 also applies to any further epidemiological bottlenecks that HIV underwent as it spread globally. The practical implication is that, subject to the functional dependencies on viral load used to quantify this model, there should not be order of magnitude differences in virulence between different HIV strains and subtypes, nor between recent or ancient human isolates. Conversely, where major differences in virulence do exist, such as between HIV-1 group M and HIV-2, it is likely that the functional relationships linking SPVL, infectiousness and survival are different. We have also explored the extent to which host mixing characteristics affect founder selection.

In this work two different definitions of an epidemic are used, defined as survival of at least two infected individuals after 100 years, and a prevalence of greater than 2% at any point within 100 years. Both were explored with a view that both outcomes were important. An epidemic which lasted 100 years with prevalence <2% would be important, as would one which explosively infected many and then died out. In most cases reported here, the outcomes are identical or very highly related. Very low population sizes are an exception: a lower population size increases the probability that the infection will cross the 2% threshold (Figure 2.4, b), because a single founding infection represents a greater percentage of the population. However, a low population is also subject to more stochastic effects and this increases the probability that the infection will die out before 100 years (Figure 2.3, b).

The relatively small effect of the parameters governing length of (D_r) and time between (D_s) relationships suggests that this part of the sexual mixing behaviour is of less importance for the questions asked in this work (Figure 2.3 and 2.4, d and e). However, increasing the length of time involved in the partnership transmission also increases the impact of virulence on the probability of emergence (Figure 2.6 and 2.7, d and e).

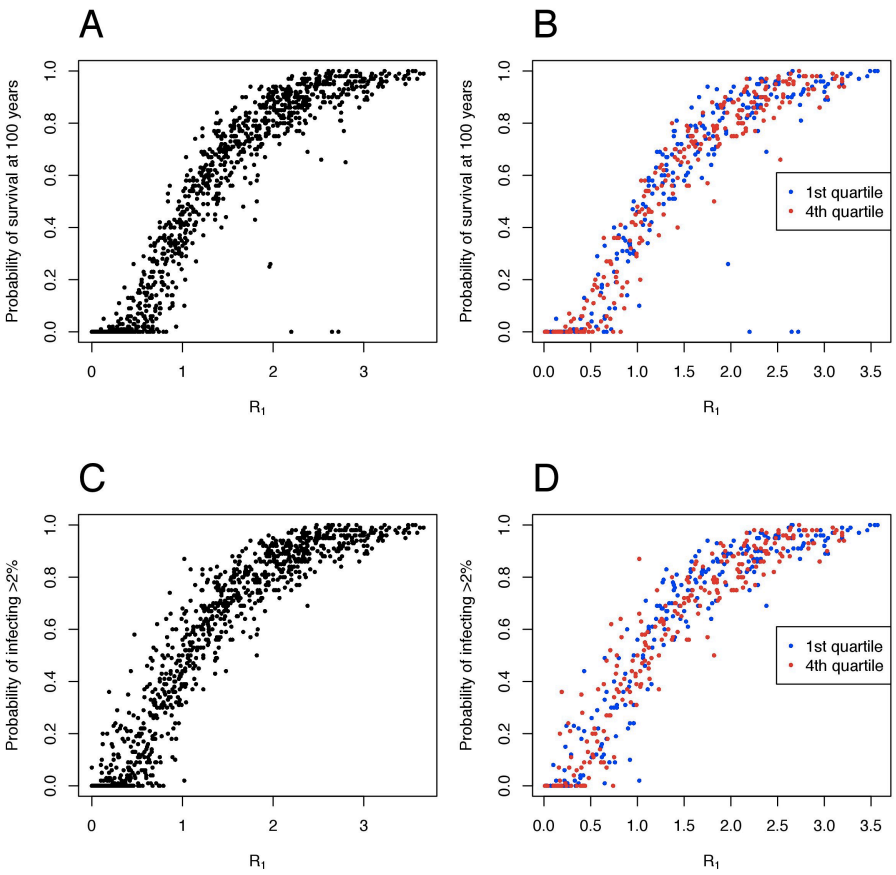


Figure 2.8: The relationship between the number of transmissions arising from the first infectee, R_1, and the probability of emergence. For each of 1000 sets of parameter values selected in the initial parameter selection step, 100 repeats were performed in order to measure the average R_1 and the probability of emergence as measured by survival after 100 years (a and b) and by infection of greater than 2% (c and d). Black points represent all 1000 simulations (a and c), while the blue and red points represent simulations with the lowest and highest quartiles of the mutational standard deviation σ_M (b and d).

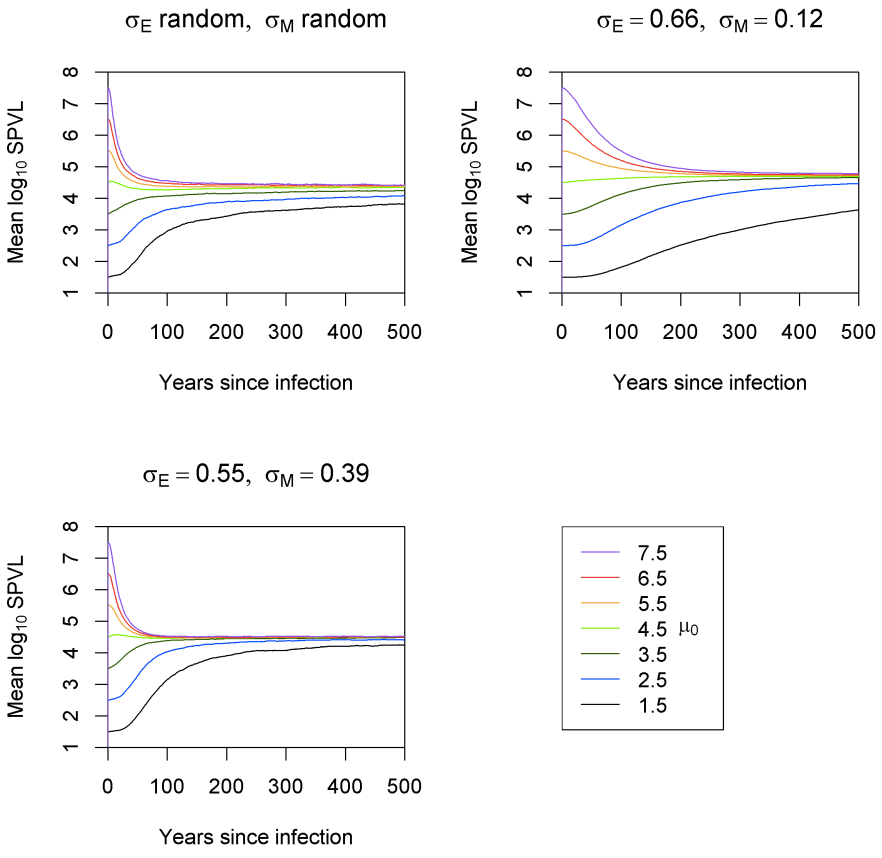


Figure 2.9: The evolution of SPVL over time from a range of different values of μ_0, averaged over 10,000 simulations. Parameter combinations were chosen from the initial parameter selection for all three panels, but in panel b the values of σ_E and σ_M were replaced with 0.66 and 0.12 respectively, and in panel c the values were 0.55 and 0.39 respectively. These values were chosen to be in accordance with values examined in chapter one.

Higher values of the core group contact rate, λ, are much more likely to produce an epidemic, and the same is also true of the contact rate ratio in the non-core group, κ, which is also positively related to the probability of emergence (Figure 2.3 and 2.4, f and g). This suggests that transmission dynamics are dominated by infection through casual rather than long-term partnerships. This is also supported by the greater probability of emergence when the core group makes up a greater proportion of the population (Figure 2.3 and 2.4, h). Interestingly the assortativity, ϵ, has little effect on emergence, except for a slight decrease in emergence probability when assortativity is extremely high (Figure 2.3 and 2.4, i), which may indicate that until the population is almost entirely assortative there is sufficient mixing to spread the infection from one group to the next. The parameters governing casual transmission do not appear to interact greatly with the effect of virulence, μ_0, on emergence, although high values of λ and κ may flatten the relationship between μ_0 and probability of emergence (Figure 2.3 and 2.4, f and g).

The relationship between R_1 and the probability of an epidemic shows a strong link (Figure 2.8, a and c). In spite of this, there does not appear to be a threshold of the reproductive number at $R_1 = 1$, suggesting that R_1 underestimates true R_0 by a non-trivial amount. This is most likely to be the result of the inadequacy of the value R_1 as a proxy for the true basic reproductive number R_0. The first infectee is more likely to be in the core group than an individual picked at random, but is less likely to be than a typical infectee, for whom initial conditions are no object. The value of R_1 is therefore expected to be an underestimate of R_0, and the results shown here support that, since there is a threshold effect, but not at $R_1 = 1$.

We also attempted to reproduce a result from work by Antia et al. (Antia, Regoes et al. 2003), who explored the relationship between mutation, R_0 and emergence in a stochastic model. They modelled a mutating pathogen population with a fitness close to the epidemic threshold, $R_0 = 1$, but in which the R_0 is slightly less than one. In a deterministic model this infection would die out, but under stochastic effects it is able to spread sufficiently that a fitter mutant is produced before extinction that then results in an epidemic. Figure 2.8 (b and d) shows that the relationship between R_1 and virulence is not modified by differences in mutation rate. The relationship does not appear to be different between the lowest (blue points) or highest quartile (red points) of mutation rates. This suggests that mutation rate plays less of a role in the epidemic potential of a strain than in the work by Antia et al. [10].

The model described in this paper supports the result from chapter one (Figure 1.10a) in which it was shown under a deterministic model that the mean log10 SPVL evolves towards intermediate values over time Figure 2.9.

A higher number of founders increases the probability of an epidemic by either measure (Figures 2.3 and 2.3), which is expected, as is the saturating relationship between number of founders and epidemic probability since there will be limiting returns as additional founders are added. A very simple probabilistic

model assuming each founder represents an individual trial exhibits the same qualitative pattern (Figure 2.5) although the fit is much better when an epidemic is defined by threshold prevalence (Figure 2.5b). While the epidemic of each major strain (HIV-1M, N & P, and HIV-2) was likely started by a single founder, considering a higher number of founders is particularly relevant when examining founder selection in epidemic bottlenecks, for example associated with the spread of subtype B virus into the USA [83]. Our results show if there are multiple founders, this relaxes selection on the virulence genotype to produce an epidemic (Figures 2.6 and 2.7, a), particularly when defining an epidemic as crossing the 2% prevalence boundary at any time. These results suggest that multiple introductions may result in a greater diversity of HIV virulence in the new population, but understanding how contemporary these introductions must be for such a phenomenon to occur, and how long this higher diversity would last before balancing selection removed these extreme genotypes, would require further work.

A major question of interest with regard to the origin of HIV is why it has emerged so recently in human history. Changes in hunting practices, such as the introduction of guns, deforestation, and urbanisation of the tropics may have contributed to increased contact with non-human primate bodily fluids, and increased the opportunities for SIV to cross the species barrier. A higher number of founders has been shown in this study to increase the probability of emergence (Figures 2.3 and Figure 2.4, a), and even if extra founders increase the probability of emergence in a limiting fashion, the increase might be even further enhanced if the two individuals are infected with different candidate strains.

2.6 Conclusion

Founder selection has likely have played an important role in shaping the virulence on HIV-1 infection right from the start of the epidemic, and selects for very similar viruses as selection for optimal spread later on in the epidemic. Our findings reiterate, in the specific context of HIV, the general prediction that major zoonotic epidemics are most likely to be caused by pathogens that are already quite transmissible [10], or which are introduced into a risk population where contact rates are such that transmission is highly amplified.

Chapter 3

How effectively can phylogenies be used to measure heritability?

3.1 Abstract

Introduction
For any infectious disease, the heritability (or similarity) of prognostic markers, measured from one infection to the next, is a measure of how much control the pathogen genotype has over disease severity. The severity of HIV-1 infection, as measured for example by SPVL, is highly variable between infected individuals. While many studies have addressed the human covariates of severity, comparatively few have examined its viral covariates. Heritability in viral load can be used to quantify the net impact of as yet unknown viral factors on HIV-1 viral load. To date, heritability estimates have been found to vary widely between studies, and within subpopulations within studies.
Methods
In this study we set out to systematically compare methods for measuring heritability in HIV-1 SPVL, with particular emphasis on methods that use viral phylogenies to estimate heritability. We present mathematical and computational results that allow different methods to produce directly comparable estimates. We also present two new phylogenetic comparative methods, which either cluster individuals into pairs ("Phylogenetic Pairs") or into clusters of various sizes ("Hierarchical Clustering").
Results
Using a wide range of existing and new methods, we re-analysed patient data from two studies, the Rakai community cohort from Uganda, and the Swiss

HIV Seroconverters Cohort. The methods differed in their ability to detect heritability, and this is dependent on which data are included in the analysis. We also simulated evolution on the phylogeny for a range of heritability values to examine how well a given method detects and quantifies true heritability. In this study we used the correlation coefficient between the SPVL of an index and a secondary case as the measure of heritability, labelled ρ. This is a lower bound on the proportion of SPVL caused by viral factors. We also estimated confidence intervals of heritability using a method that is robust but computationally intensive. Depending on the phylogeny, several of the methods could consistently detect ρ above 0.4, and none below. Of the methods we tested, we found that Pagel's λ and Phylogenetic Pairs were the most robust and sensitive. These methods produced variable estimates of heritability, which were generally consistent with results from previous studies.

Conclusion

Our results suggest that in some populations, more than half of variation in SPVL is controlled by viral factors, but that in others, this proportion may be considerably lower. We also show that existing phylogenetic methods are ill-suited for detecting levels of heritability below 0.4, which may nonetheless be biologically important.

3.2 Introduction

HIV has a high mutation rate [107, 5] and daily turnover [100, 227] and therefore adapts rapidly under selective pressure. Adaptation to the host environment has been extensively studied in the context of immune escape [232, 27], pathogenesis [165] and drug resistance [197, 205]. Increasingly there is interest in viral adaptation to the population, for example through transmission of escape [136] or drug resistance mutations [236]. The impact of HIV on mortality and morbidity and its heterogeneity in severity of infection make it an important topic for study. Understanding the effect of public health interventions which affect transmission on virulence evolution is important when evaluating such programmes [77].

The heritability of a trait determines the rate at which it will evolve under directional selection. The phenotype of any organism is controlled partly by its genome, and partly by its environment. The heritability, h^2, is defined in equation 3.1 as the proportion of total phenotypic variance (${\sigma_P}^2$) which is ascribed to genetic variance (${\sigma_G}^2$) [70].

$$h^2 = \frac{\sigma_G^2}{\sigma_P^2} \tag{3.1}$$

SPVL is partly determined by host factors, both genetic such as HLA type (reviewed in [31]), sex [53], and non-genetic factors such as age [196], and co-infections [2, 111]. Specific human genetic factors have been identified to which 13% of SPVL variation can be attributed, with a further 9% explained by age, sex and population structure [64], but the total environmental component of variance (${\sigma_E}^2$, which in this context includes all host effects) may still be much greater than this. The remaining variation in SPVL is the result of variation in the VVG. Subtype-specific differences in virulence (reviewed in [81]) provided the first evidence for its control by the viral genotype. Subsequently a number of studies have measured heritability directly by comparing the variance in SPVL between individuals with closely related virus, and with the variance between those with distantly related virus. The wide range of heritability estimates (Table 3.1, adapted from [158]) may result from the diversity in the datasets, or from chance, but may result from differences in the methods. Wide confidence intervals also suggest large uncertainty in the estimates.

Table 3.1: Results from previous studies in heritability of set-point viral load.

Study	**Method**	**Data source**	**Population**[a]	n	**Heritability**
Tang et al. 2004 [215]	Pearson correlation coefficient[b]- r	Zambia couples	HET	115×2	20% (1%,37%)
Hecht et al. 2010 [94]	Pearson correlation coefficient[b]- r	USA couples	MSM	24×2	55% (19%,78%)
van der Kuyl et al. 2010 [130]	Pearson correlation coefficient[b]- r	Netherlands couples	MSM/HET	56×2	25% (-1%,48%)
Hollingsworth et al. 2010 [104]	ANOVA coefficient of determination[c]- R^2	Uganda couples	HET Strong	29×2	37%[d]
			HET Weak/Strong	97×2	23%[d]
Alizon et al. 2010 [4]	Independent contrasts - K	Switzerland general population	MSM Strict	134	K=0.59 (0.45,0.73)[e]
			All Strict	230	Not significant
			MSM Liberal	404	K=0.09 (0.02,0.16)[e]
			All Liberal	661	Not significant

[a] HET: heterosexuals; MSM: men who have sex with men.
[b] Correlation coefficient is equivalent in these scenarios to the percentage of variation explained by the partnership [71].
[c] Coefficient of determination is equivalent to the correlation coefficient if the viral loads of the source and the recipient are normally distributed with identical variance [158], but in the ANOVA method it is easier to adjust for cofactors [71].
[d] Confidence intervals not available.
[e] Confidence intervals based on twice the standard deviation (0.07).

Most of the studies examined SPVL correlation between pairs of individuals where one had infected the other (transmission pairs). These studies required the identification and study of couples where transmission was epidemiologically suspected and confirmed by phylogenetic analysis, and so the sample sizes are as a result relatively small. The phylogenetic comparative study by Alizon et al. [4] had the advantage that any sample of well characterised patients could be analysed, as the study used only the phylogenetic relationship between HIV sequences to infer genetic similarity, allowing a larger sample size, and potentially the application of the method to a much wider range of data sources. However, detection and measurement of heritability in the phylogenetic study was found to be strongly dependent on the inclusion criteria used. Contrary to expectations, the most exclusive criteria and therefore the smallest sample size exhibited the greatest level of inferred heritability. The authors did not take into account the effect of cofactors such as age and co-infections which may influence the SPVL and which may also cluster together on the phylogenetic tree. It is also uncertain exactly how the quantity measured by the two methods used (Pagels λ and Blombergs K) should be interpreted in the context of heritability. We herein propose a new approach which, as in [4], uses only the phylogenetic distances to determine genetic proximity of viruses, but allows inclusion of cofactors and also links the results to true heritability.

The aim of this study was to test this new approach, and compare it with other available methods for measuring trait heritability on a phylogenetic tree.

3.3 Methods

We tested the effectiveness of methods which measure heritability by using the phylogeny to reconstruct the transmission network. Each method produces a single statistic which is a measure of phylogenetic signal. We examined how well each statistic could be used to identify and quantify heritability.

We developed two methods, the Phylogenetic Pairs (PP) method and the Hierarchical Clustering (HC) method. They use the phylogeny to identify transmission clusters on the tree and then test for similarity of SPVL within the clusters using analysis of variance (ANOVA). We also describe and test several other statistics developed elsewhere to measure phylogenetic signal (the Mantel test [147]; Blomberg's independent contrasts [24]; Pagel's λ transformation [75]; and the Abouheif-Moran tests [178]).

Phylogenetic Pairs (PP)

The PP method analyses putative transmission pairs on the phylogenetic tree, without requiring behavioural data. These pairs are identified as two tips on the phylogenetic tree that are each others closest neighbour. They are located by testing each tip i in turn. For each i its nearest neighbour j is located. If i is also the nearest neighbour of j, then they are defined as a pair. There

Table 3.2: Characteristics of the individuals from the Rakai cohort which are included in this study.

Factor	**Category**	**Number**
Sex	Male	164 (44.4%)
	Female	205 (55.6%)
Age	15-24	53 (14.4%)
	25-29	65 (17.6%)
	30-39	93 (25.2%)
	40+	39 (10.6%)
	Missing data	119 (32.2%)
Genito-urinary discharge (GUD)	Yes	64 (17.3%)
	No	40 (10.8%)
	Missing data	265 (71.8%)
Subtype	A	55 (15.9%)
	D	240 (65.0%)
	Other/missing	74 (20.1%)

is no maximum distance between them, nor must they form a monophyletic group.

The ANOVA is then performed as in equations 3.2 and 3.3 where Pair indicates the pair number and in each case the adjusted R^2 is used as a measure of the variance explained. The additional variance explained by the model which takes into account the pairs becomes the test statistic PP (equation 3.4). Individuals not included in a pair are excluded. The cofactors shown here are those available in the Rakai dataset (Table 3.2) where GUD is Genito-Urinary Discharge. In the Swiss data sex, age and risk group are used (Table 3.3).

$$\text{SPVL: Sex+Age+GUD+Pair } \rightarrow \text{adj}{R_c}^2 \tag{3.2}$$

$$\text{SPVL: Sex+Age+GUD } \rightarrow \text{adj}R^2 \tag{3.3}$$

$$PP = \text{adj}{R_c}^2 - \text{adj}R^2 \tag{3.4}$$

Cofactors are only included for the randomisation test, and not in the simulations, because although they affect SPVL, it is not clear how to incorporate these effects in simulations. One possible method would be to calculate

Table 3.3: Characteristics of the individuals from the Swiss cohort which are included in this study.

Factor	**Category**	**Number**
Sex	Male	554 (83.8%)
	Female	107 (16.2)
Age	15-24	53 (8.0%)
	25-29	110 (16.6%)
	30-39	280 (42.4%)
	40-49	151 (22.8%)
	50+	49 (7.4%)
	Missing data	18 (2.7%)
Genito-urinary discharge (GUD)	Yes	64 (17.3%)
	No	40 (10.8%)
	Missing data	265 (71.8%)
Risk group	Blood	2 (0.3%)
	Heterosexual	142 (21.4%)
	Injecting drug use	97 (14.7%)
	MSM	404 (61.1%)
	Other	4 (0.6%)
	Missing	12 (1.8%)
Viral load variability	Liberal only	431 (65.2%)
	Strict	230 (34.8%)

the fixed effects of the cofactors using a linear regression, simulate the residuals, and subsequently add the fixed effects, but this requires using residuals as data, which is problematic [74]. We therefore ignore the cofactors for the purposes of simulation, and they are not included in analyses which used simulated data. In this case the ANOVA is performed by regression of SPVL on Pair alone (equation 3.5), and the test statistic PP_s was the resulting adj${R_s}^2$ (equation 3.6).

$$\text{SPVL: Pair} \rightarrow \text{adj}{R_s}^2 \tag{3.5}$$

$$PP_s = \text{adj}{R_s}^2 \tag{3.6}$$

Hierarchical Clustering (HC)

The HC method is similar to the PP method but it combines individuals into clusters of two or more according to a threshold distance. Short distances give small clusters and many individuals are not included in a cluster, while larger distances give larger clusters and more inclusion. There is no obvious cluster size, proportion included or number of clusters to use, so the method integrates over the range of clustering distances. There are also 4 possible clustering algorithms, implemented using the hclust method in R [217], although the 'ward' method has been used as the default for this analysis.

1. The complete method joins an individual to a cluster if it is linked to (within the threshold distance of) every other member of the cluster.
2. The single method joins an individual to a cluster if it is linked to any other member of the cluster.
3. The average method joins an individual to a cluster if the average distance to every other member of the cluster is lower than the threshold.
4. The ward method calculates the least squares cost of each cluster, based on the distances between its members. If the cost of joining two clusters is less than the threshold then they will be combined.

Once individuals are assigned clusters at a given distance d, a pair of ANOVAs is performed, as in equations 3.7 and 3.8. The value of adjR^2 is also dependent on d since only individuals who are in clusters at that distance are included. The value of adj${R_c}^2(d)$ adj$R^2(d)$ is calculated for the range of distance, and the sum over all d, multiplied by its increment, Δd=0.01, is the test statistic, HC 3.9. This method therefore measures the degree of clustering at all depths of the tree. While a nested ANOVA would have been a desirable way to approach this problem, there was insufficient data and too many parameters to estimate for such a complex nesting model.

$$\text{SPVL: Sex+Age+GUD+Cluster}(d) \rightarrow \text{adj}{R_c}^2(d) \tag{3.7}$$

$$\text{SPVL: Sex+Age+GUD} \rightarrow \text{adj}R^2(d) \tag{3.8}$$

$$HC = \Delta d \sum_d \left(\text{adj}{R_c}^2(d) - \text{adj}R^2(d)\right) \tag{3.9}$$

As with the PP method, cofactors are only included in the randomisation test, so when simulations are involved a single ANOVA is done for each clustering

distance 3.10 and the test statistic, HC, is the sum over each value of adj${R_s}^2(d)$ 3.11.

$$\text{SPVL: Cluster}(d) \rightarrow \text{adj}R^2(d) \tag{3.10}$$

$$HC_s = \Delta d \sum_d \text{adj}{R_s}^2(d) \tag{3.11}$$

Mantel test

The Mantel test measures correlation between two distance matrices [147], in this case the phylogenetic distance matrix, and the SPVL trait difference matrix. The statistic is the sum of the product of each cell and the corresponding cell in the other matrix (excluding the diagonal). The method was implemented in R [217] using the *ade4* package [56].

Blombergs' K

In this method pairs of sister tips are contrasted independently by dividing their difference in trait value by the square root of the branch length separating them [66, 4]. These tips are then removed, and their parent node ascribed a new trait value from an average of the value at each daughter tip, weighted by the reciprocal of the branch length from the parent node to each tip. This new tip can then be compared with its sister tip, and this continues until every pair has been contrasted. The variance of these values, the Phylogenetic Independent Contrasts (PICv) is high (low) when phylogenetic signal is low (high).

The PICv can be converted to the K statistic by dividing it by its expected value under a Brownian motion model of evolution, which depends on the shape and size of the phylogeny. If K=0, this indicates no phylogenetic signal, while K=1 implies evolution according to Brownian motion. K can also exceed one, for example if the trait is selected in different directions on different branches of the tree. This statistic was measured in R [217] using the phylosignal function from the package *picante* [118].

Pagel's λ

The shared evolutionary history of the tips on a phylogeny can be expressed as a variance-covariance matrix of size nxn where n is the number of tips. The values along the diagonal represent the distance between each tip and the root, and the off-diagonal values give the distance between the root and the most recent common ancestor of the two tips. This value therefore captures the

shared evolutionary history between these two tips, with higher values representing a greater amount of shared history, and a consequent similarity in trait values.

These covariances can be used to simulate these trait values, which can be fit to the true data. The λ statistic is a multiplier which transforms the covariances to maximise the fit between the simulated and the real tip data. If λ=0, the phylogeny has no effect . If λ=1, trait evolution can be best described by a Brownian motion model, and $\lambda > 1$ is also possible [75]. It was implemented using the phylosig function [169, 193].

Abouheif Moran test

The Abouheif Moran statistic is the sum of the differences between the trait values for each pair of tips, weighted by the reciprocal of the distance between them. There are five different methods for measuring the distance between tips, although number two, 'Abouheif', has been used as the default for this analysis. These tests were conducted in the *adephylo* package [109] in R [217].

1. 'oriAbouheif': derived from the test for serial independence [1], this measure is the inverse of the product of the number of branches descending from each node in the path connecting the two tips [178].
2. 'Abouheif': as above but the diagonals in the proximity matrix are set to zero [178].
3. 'sumDD': as in the "Abouheif" method, but it measures the inverse of the sum of the number of descending branches, rather than the inverse of the product.
4. 'nNodes': The inverse of the number of nodes on the path connecting the two tips.
5. 'patristic': The inverse of the branch length on the path separating two tips.

3.3.1 Significance tests

For each statistic a randomisation test is performed in which the patient data associated with the tips are randomly permuted on the tree 1000 times, and the statistic is recalculated each time. Randomisation scrambles phylogenetic signal, giving an expected distribution of the statistic under no heritability, which is the null hypothesis. The proportion of randomisations which give a statistic higher than the true value is the one-tailed p-value. When the randomisation test is performed for the PP and HC statistics the cofactors remain with their corresponding tips.

3.3.2 Quantifying heritability

In order to examine the relationship between true heritability and each of the available statistics, we simulated evolution of SPVL on phylogenetic trees with a specific input value ρ as a measure of heritability, in a manner derived from previous work [4]. This was done to test the sensitivity of each statistic to different levels of heritability, and also to estimate the true heritability.

The simulation begins at the root node, and each daughter node is given a SPVL value based on that of its parent node, and on the SPVL distribution in the whole population. The higher the heritability, the more the value depends on the parent. The log10 SPVL at the root node, μ_R, is given by the mean of the true SPVL data (equation 3.12.

$$\mu_R = \bar{\mu} \tag{3.12}$$

Let M be a random variable normally distributed with the mean and variance given by that of the true log10 SPVL data (equation 3.13.The SPVL at each subsequent daughter node, V_D, is derived from the parent node, V_P, the random variable M, and the value ρ (equation 3.14). This method is derived from previous work [4] and is a variation on the Ornstein-Uhlenbeck process [220, 65] which allows Brownian motion to occur while constraining the distribution of the population.

$$M \sim N(\bar{\mu}, {\sigma_P}^2) \tag{3.13}$$

$$V_D = \rho V_P + M\sqrt{1-\rho^2} \tag{3.14}$$

The result is a set of simulated data at the tips of the tree which is used for analysis. The value ρ is the correlation coefficient between an index (P) and a secondary case (D), since equation 3.14 can also be described as in equation 3.15.

$$V_D|V_P \sim N\left(V_P, (1-\rho^2){\sigma_P}^2\right) \tag{3.15}$$

3.3.3 Alternative method of simulation

The simulation method described above assumes that a branch length between two nodes on the phylogeny represents a single transmission, and the expected evolution of virulence is the same along any branch, regardless of its length. In an alternative method, a branch between two nodes could represent multiple generations, with longer branches allowing more transmissions.

The number of generations between each parent and daughter node (P and D respectively) is represented by g_{PD} (equation 3.16). At least one generation is assumed, but the number in excess of this is taken from a Poisson distribution where l_{PD} is the distance between the two nodes in mean substitutions per site, r is the rate of evolution on the phylogeny in substitutions per site per year, and T_g is the mean generation time of HIV-1.

$$g_{PD} = 1 + Poisson\left(\frac{l_{PD}}{rT_g}\right) \tag{3.16}$$

The rate of evolution in the Rakai dataset was measured at 6.8×10^{-3} per site per year using Path-O-Gen [189]. The mean generation time in this dataset was measured as 4.3 years [103].The SPVL at each daughter node is then given as in equation 3.14 but allowing for more than one generation. The random variable, M_b, is picked from a normal distribution, but represents the sum of g_PD normal distributions, and is therefore distributed with lower variance (equation 3.18.

$$V_D = \rho^{g_P D} V_P + M_b \left(\sum_{i=0}^{g_{PD}-1}\right) \sqrt{1-\rho^2} \tag{3.17}$$

$$M_b \sim N(\bar{\mu}, \frac{{\sigma_P}^2}{g_{PD}}) \tag{3.18}$$

In this study we will describe heritability in terms of the correlation coefficient, ρ, and discuss how this term can be interpreted in terms of traditional heritability (h^2, equation 3.1) to which it is strongly related.

Although in theory the range of a correlation coefficient is from -1 to +1, in reality it is problematic to talk of negative correlation between a series of individuals, since negative correlation with ones nearest neighbour may entail positive correlation with their next neighbour. Thus when we use ρ as a summary measure of an entire population, we consider only the range from 0 to +1.

3.3.4 Estimating true heritability

In order to estimate the value of true heritability in the data, and the confidence intervals surrounding it, we used data simulated with given heritability to identify the relationship between the input values of ρ and the output values of a given statistic. The true value of the statistic could then be examined in light of this relationship to establish true heritability.

We used an Approximate Bayesian Computation approach [45] to identify the true heritability distribution. Values of ρ were drawn from a prior distribution,

the data were simulated and a test statistic Z was calculated and compared with the test statistic from the true data $\hat{Z}$ with a tolerance parameter ϵ If the inequality 3.19 is satisfied, that value of ρ is accepted and forms part of the approximate posterior distribution of true ρ. In this study ϵ for a given statistic was set at 10% of the standard deviation of values for that statistic resulting from a set of preliminary simulations with ρ drawn from a uniform between 0 and 1 (data not shown).

$$\hat{Z} - \epsilon < Z|\rho < \hat{Z} + \epsilon \tag{3.19}$$

The prior distribution of ρ had a large effect on the distribution of the posterior value. It is important to identify whether or not heritability was significantly different from 0, but if all values of ρ were drawn from a uniform distribution (3.20) then the lower boundary of the confidence interval would always exceed 0.

$$\rho \sim U(0,1) \tag{3.20}$$

In order to allow the boundary of the confidence interval to include 0, we included values of 0 in the prior of ρ. We defined a further parameter, π, which represented the probability that a simulation had ρ=0 (3.21). Rather than drawing ρ from a uniform prior, instead π was drawn from U(0,1), which implied that 50% of values of ρ in the prior were drawn from a uniform distribution and 50% were set at ρ=0.

$$\rho \sim \begin{cases} 0 & \text{with probability } \pi \\ U(0,1) & \text{with probability } 1-\pi \end{cases} \tag{3.21}$$

Values of ρ were drawn from this prior until 1000 were successful, and these values could then be plotted as representing the distribution of heritability values, and their confidence intervals could be calculated by looking for the 2.5 and 97.5 percentiles.

3.3.5 Sensitivity testing

To test the sensitivity of each statistic we simulated 100 datasets with a range (0-1) of ρ values. For each dataset we performed a test with 100 randomisations, to measure how well each statistic detected significant heritability at different values of ρ. The method was computationally intensive, so 100 simulations and 100 randomisations for each were considered to be the minimum number adequate.

3.3.6 Data

Rakai

The study population was enrolled in the Rakai Community Cohort Study in the rural Rakai District of south-western Uganda. The study methods for this cohort have been outlined elsewhere [224, 225], but the specific selection of the cohort under study in this work has been described in further detail in more recent work [104], which also provided the phylogeny constructed from all sequences 3.1, inferred using *env* and *gag* sequences. Characteristics of the cohort are shown in 3.2. The phylogenetic inference was performed using the concatenated *env* and *gag* genes. The analysis was conducted using the GTR+Γ+I method in RAxML [213]. The outgroup was from an SIV sequence (Accession number: AB177846.1).

We also separated the subtype A and subtype D sequences, performed individual phylogenetic inference on each using the method specified for the entire dataset, and then used the subtype A and subtype D datasets as alternative subdivisions of the data (n=55 and n=240, respectively).

We also analysed subdivisions used in previous work [104]. The dataset contained several pairs who were identified as cohabiting couples, initially with one infected and one uninfected, in which the uninfected partner subsequently seroconverted. There were 16 couples for whom both gag and env were available, whose HIV sequences formed a monophyletic group on the phylogeny with >80% bootstrap support, and these individuals are referred to as the strongly linked couples (n=32). A third selection contained the strongly linked individuals and an additional 15 couples who were weakly linked, since they were monophyletic but with <80% bootstrap support (n=62).

Swiss

The Swiss data was taken from the Swiss HIV Cohort Study which has been described elsewhere [134, 204] and the specific selection has been analysed in previous work [4], from which we also took the phylogeny based on the *pol* gene 3.3. This was inferred using PhyML [92] from subtype B infected individuals for whom at least three viral load measurements were available after primary infection and before anti-retroviral therapy. Characteristics of the cohort are shown in Table 3.3.

The entire dataset (All Liberal, n=661) fit the liberal definition for viral load variability, meaning that three consecutive viral load measurements were available which remained within a one-log band during the period of asymptomatic infection. We also analysed three smaller more subdivisions of the data as in previous work [4]. The MSM Liberal subdivision (n=404) includes only individuals who are men who have sex with men. The All Strict subdivision (n=230) includes only individuals who fit the strict definition of viral load variability,

meaning that in addition to having three consecutive viral load measurements within a one-log band, they do not have any measurements within the time period (180 days after infection to 3 years after the first positive viral RNA) which are outside this band. The MSM Strict subdivision includes individuals who fit both criteria (n=134).

Trees for simulations were read and manipulated using the *ape* package [173] in R [217], which was also used to plot the trees.

3.4 Results

3.4.1 Testing significance in the true data.

The significance of each of the statistics when applied to each of the three subdivisions in the Rakai data was tested by randomisation (Table 3.4). There was more evident heritability in the Swiss data (Table 3.5). In the Swiss data the smaller subdivisions exhibited more significant heritability than larger ones. While this was assumed to be a result of the shared characteristics of the smaller subdivisions, but it is possible that this is the effect of sample size itself. We artificially created a smaller dataset which did not share characteristics by randomly sampling without replacement from the larger dataset. We thus produced a smaller dataset which was the same size as the MSM Strict dataset. Significance tests were then performed on these resampled datasets and the results compared with those from the targeted subdivisions which do share the characteristics 3.6. In almost every case the level of significance detected was much higher in the targeted than the random subdivision.

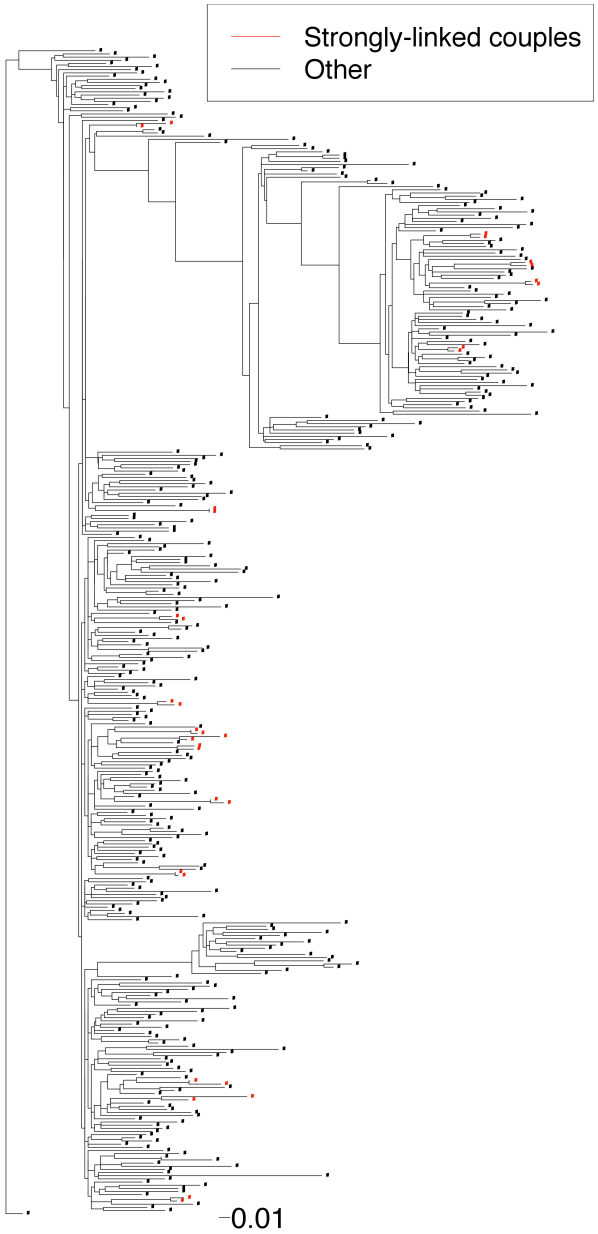


Figure 3.1: Maximum likelihood phylogenetic tree for the entire Rakai cohort used in this work, showing branch lengths and the strongly-linked couples are in red.

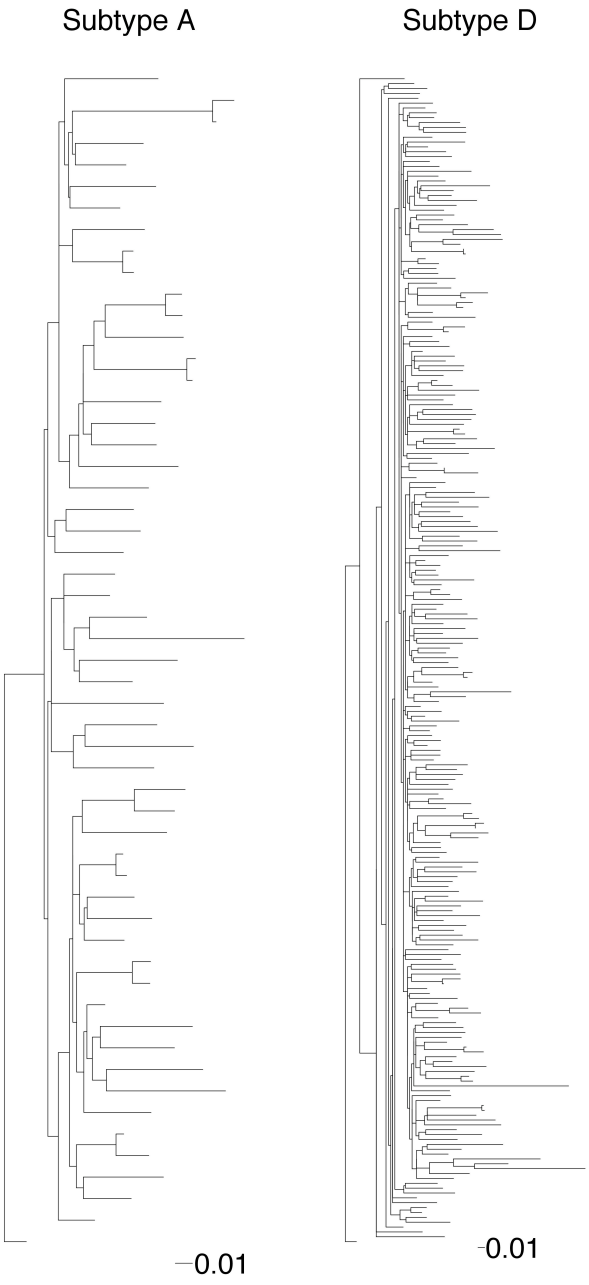


Figure 3.2: Phylogenetic tree for the two dominant subtypes, A and D, in the Rakai cohort, showing branch lengths.

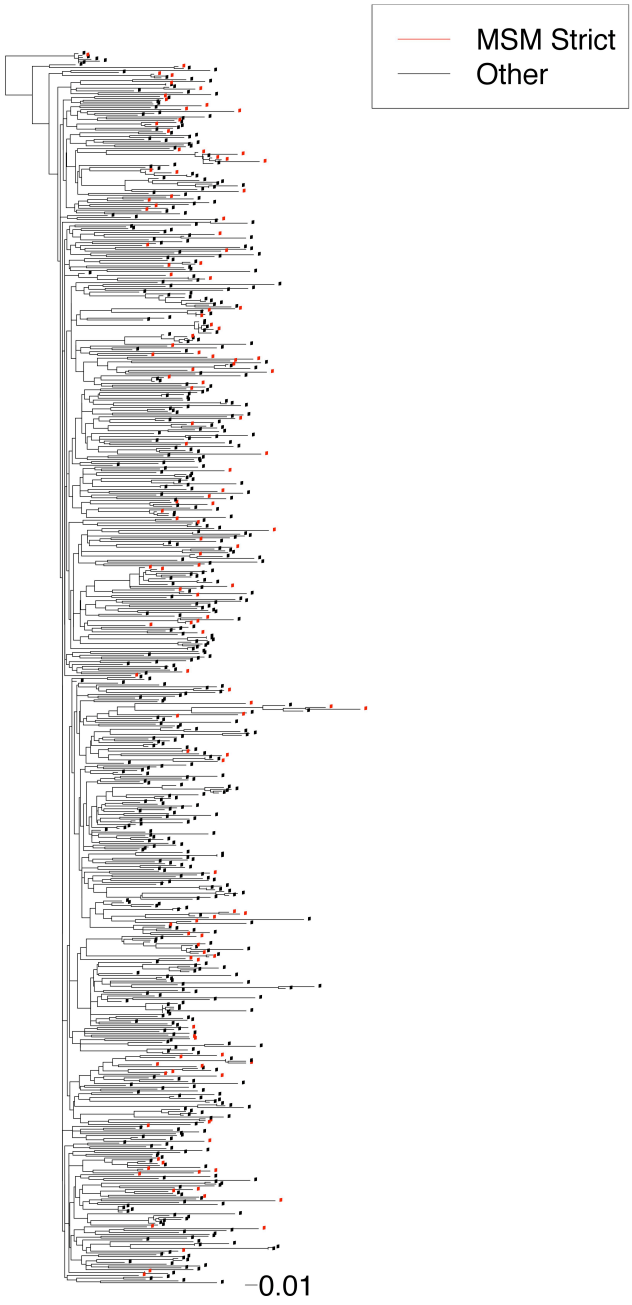


Figure 3.3: Phylogenetic tree for the Swiss cohort, showing branch lengths and the MSM Strict group are in red.

Table 3.4: The statistics (Z) and p-values (one-tailed) from analysis of the Rakai data including strongly linked couples only, strong and weakly linked couples, subtypes A and D or the entire dataset. P-values showing borderline significance (p<0.1) are in blue, and formal significance (p<0.05) is in red.

	Couples (n=32)		**Strong/Weak (n=62)**		**Subtype A (n=55)**		**Subtype D (n=240)**		**All (n=369)**	
Method	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**
PP	0.3774	0.083	0.3597	0.035	0.6213	0.136	0.4979	0.014	0.2817	0.024
HC complete	0.0712	0.088	0.0800	0.039	-0.004	0.616	0.0199	0.196	0.0208	0.146
HC single	0.0571	0.104	0.0317	0.158	0.0087	0.414	0.0127	0.250	0.0097	0.294
HC average	0.0644	0.096	0.0680	0.043	0.0089	0.464	0.0215	0.154	0.0143	0.222
HC ward	0.1137	0.113	0.1118	0.064	-0.0096	0.650	0.0247	0.168	0.0146	0.267
Mantel	-0.0539	0.680	-0.0197	0.613	0.1067	0.099	0.0223	0.288	-0.0216	0.795
Blomberg's K	0.0016	0.829	0.0003	0.819	0.4938	0.380	0.7202	0.094	0.0001	0.836
PICv	2241	0.828	13650	0.826	20.33	0.568	15.19	0.406	27290	0.832
Pagel's λ	0.3442	0.125	0.4481	0.074	0.0001	0.412	0.4395	0.321	0.7549	0.104
AM oriAbouheif	0.1635	0.091	0.1398	0.058	-0.1341	0.935	0.0043	0.420	0.0335	0.182
AM Abouheif	0.1266	0.106	0.1327	0.046	-0.1477	0.928	-0.0073	0.492	0.0308	0.182
AM sumDD	0.0191	0.105	0.0305	0.023	-0.0589	0.965	-0.0039	0.433	-0.0016	0.349
AM nNodes	0.0191	0.105	0.0305	0.023	-0.0589	0.965	-0.0039	0.433	-0.0016	0.349
AM patristic	0.1204	0.060	0.0659	0.036	-0.0267	0.719	-0.0017	0.084	0.0048	0.094

PP – Phylogenetic Pairs
HC – Hierarchical Clustering
PICv – Variance of Phylogenetic Independent Contrasts
AM – Abouheif Moran tests

Table 3.5: The statistics (Z) and p-values (one-tailed) from analysis of the Swiss data including subdivisions according to risk category (MSM or All) and variability of viral load (Strict or Liberal). Borderline significance ($p<0.1$) is in blue, and true significance ($p<0.05$) is in red.

	MSM Strict (n=134)		**All Strict (n=230)**		**MSM Liberal (n=404)**		**All Liberal (n=661)**	
Method	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**
PP	0.5948	0.005	0.3757	0.020	0.0094	0.492	-0.0892	0.817
HC complete	0.0494	0.032	0.0445	0.038	0.0195	0.096	0.0082	0.214
HC single	0.0343	0.040	0.0298	0.038	0.0082	0.175	0.0056	0.159
HC average	0.0462	0.024	0.0387	0.043	0.0158	0.113	0.0041	0.315
HC ward	0.0634	0.023	0.0603	0.020	0.0356	0.036	0.0077	0.315
Mantel	0.0499	0.207	0.0400	0.201	-0.0049	0.516	-0.014	0.712
Blomberg's K	0.5933	0.080	0.0251	0.366	0.0914	0.038	0.0023	0.305
PICv	5.579	0.002	127.2	0.345	40.22	0.052	1594	0.307
Pagel's λ	0.6459	0.016	0.2155	0.030	0.1534	0.025	0.1048	0.016
AM oriAbouheif	0.1273	0.023	0.0161	0.361	0.0383	0.141	-0.0126	0.680
AM Abouheif	0.1256	0.019	0.0154	0.338	0.0365	0.138	-0.0138	0.671
AM sumDD	0.0133	0.016	0.0006	0.168	0.0007	0.136	0.0000	0.166
AM nNodes	0.0133	0.016	0.0006	0.168	0.0007	0.136	0.0000	0.166
AM patristic	0.0002	0.054	0.0017	0.155	0.0046	0.119	0.0018	0.320

PP – Phylogenetic Pairs
HC – Hierarchical Clustering
PICv – Variance of Phylogenetic Independent Contrasts
AM – Abouheif Moran tests

Table 3.6: The effect of sample size reduction on significance. For each of the three smaller subdivisions, the impact of their lower sample size on their significance was assessed by creating three analogous samples of the same size as these, but which were drawn at random from the entire dataset rather than being selected by their shared characteristics. The targeted subdivisions which have been previously analysed (MSM Strict, All Strict, MSM Liberal and All Liberal) have been reproduced from Table 3.5. Each of the smaller subdivisions has been matched by a randomly resampled dataset of the same size. Borderline significance ($p<0.1$) is in blue, and true significance ($p<0.05$) is in red.

	n=134				n=230				n=404			
	MSM Strict		**Random**		**All Strict**		**Random**		**MSM Liberal**		**Random**	
Method	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**	**Z**	**p**
PP	0.5948	0.005	0.6686	0.200	0.3757	0.020	-0.1651	0.781	0.0094	0.492	-0.1631	0.905
HC complete	0.0494	0.032	0.0566	0.102	0.0445	0.038	-0.0314	0.890	0.0195	0.096	-0.0151	0.844
HC single	0.0343	0.040	0.0380	0.086	0.0298	0.038	-0.0216	0.888	0.0082	0.175	-0.0062	0.745
HC average	0.0462	0.024	0.0483	0.101	0.0387	0.043	-0.0287	0.888	0.0158	0.113	-0.0098	0.775
HC ward	0.0634	0.023	0.0635	0.101	0.0603	0.020	-0.0375	0.902	0.0356	0.036	-0.0154	0.791
Mantel	0.0499	0.207	0.0275	0.299	0.0400	0.201	-0.0543	0.879	-0.0049	0.516	-0.0398	0.884
Blomberg's K	0.5933	0.069	0.4678	0.389	0.0252	0.375	0.0014	0.807	0.0914	0.031	0.0021	0.594
PICv	5.579	0.002	7.167	0.286	127.2	0.345	2397	0.764	40.22	0.052	1500	0.552
Pagel's λ	0.6459	0.016	0.0001	0.241	0.2155	0.030	1.456	0.000	0.1534	0.025	0.0001	0.255
AM oriAbouheif	0.1273	0.023	-0.0397	0.732	0.0161	0.361	-0.0351	0.780	0.0383	0.141	-0.0382	0.871
AM Abouheif	0.1256	0.019	-0.0441	0.723	0.0154	0.338	-0.0402	0.784	0.0365	0.138	-0.0414	0.878
AM sumDD	0.0133	0.016	-0.0126	0.680	0.0006	0.168	-0.0047	0.522	0.0007	0.136	-0.0032	0.612
AM nNodes	0.0133	0.016	-0.0126	0.680	0.0006	0.168	-0.0047	0.522	0.0007	0.136	-0.0032	0.612
AM patristic	0.0002	0.054	0.0004	0.037	0.0017	0.155	-0.0048	0.583	0.0046	0.119	0.0003	0.277

PP – Phylogenetic Pairs
HC – Hierarchical Clustering
PICv – Variance of Phylogenetic Independent Contrasts
AM – Abouheif Moran tests

3.4.2 Estimates of heritability

True heritability was estimated by using the simulated heritability results to interpret the value of the statistic from the true data. Simulation of evolution on the tree under study was combined with an ABC method to measure the approximate posterior distribution of the true value of ρ in the data. The distributions of heritability values are plotted, for the entire Rakai dataset (Figure 3.4) as well as the analysis on the individual subtypes A and D (Figures 3.5 and 3.6), and on the entire Swiss data (Figure 3.7). The result from analysis of the MSM Strict subdivision was also plotted (Figure 3.8), as the subdivision with the most unambiguous heritability as demonstrated by the randomisation test (Table 3.5).

Tables 3.7 and 3.8 show the estimates of ρ. The distributions of heritability values were used to estimate the best estimate of ρ, which was taken to be the median because the distributions were asymmetric. The distribution was also used to calculate two-tailed 95% confidence intervals for the value of ρ. In none of the combinations of test statistics and data subdivision were any of the confidence bounds above zero. The tables also show the proportion of the posterior distribution of ρ which is zero. As expected, the data subdivision exhibiting the lowest proportion of zeros is the MSM Strict group from the Swiss data.

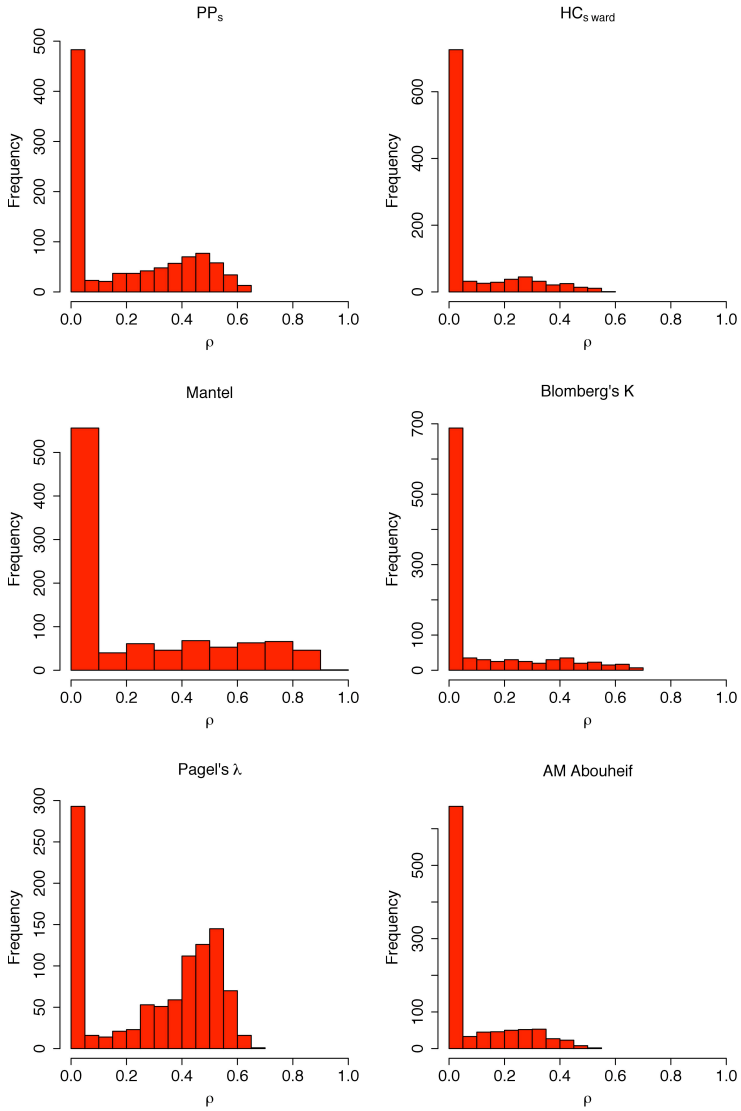


Figure 3.4: Estimated heritability in the entire Rakai dataset. Each plot represents a different statistic, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

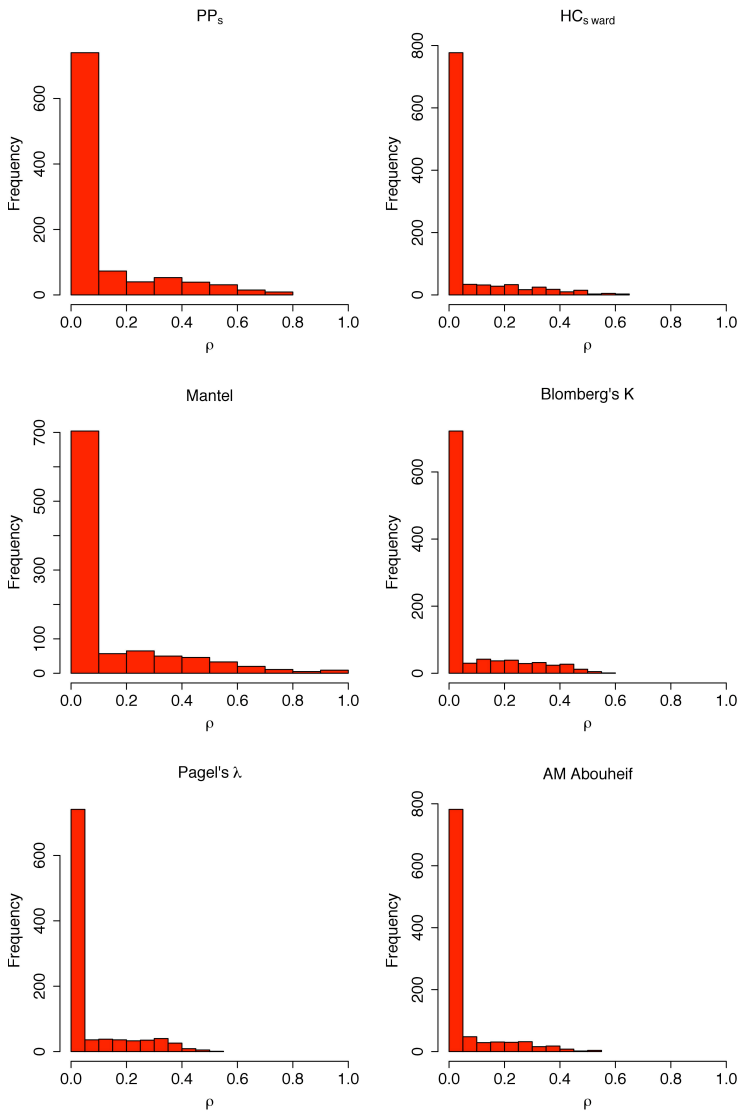


Figure 3.5: Estimated heritability in the subtype A infected individuals in the Rakai dataset. Each plot represents a different statistic, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

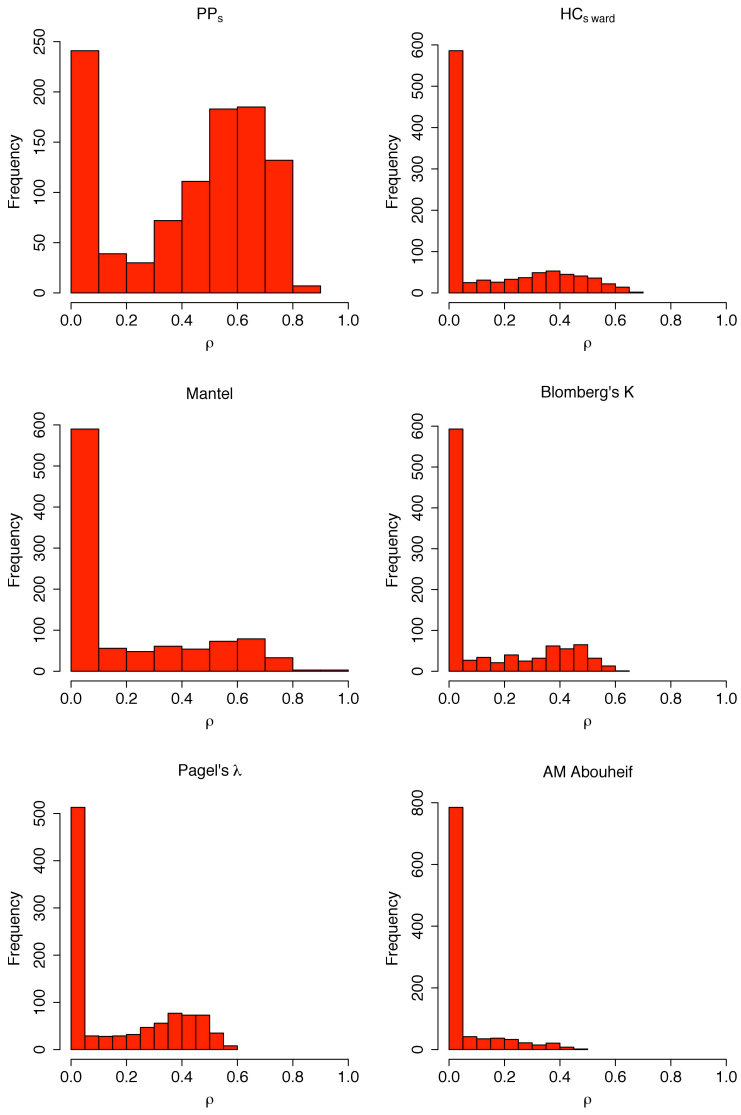


Figure 3.6: Estimated heritability in the subtype D infected individuals in the Rakai dataset. Each plot represents a different statistic, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

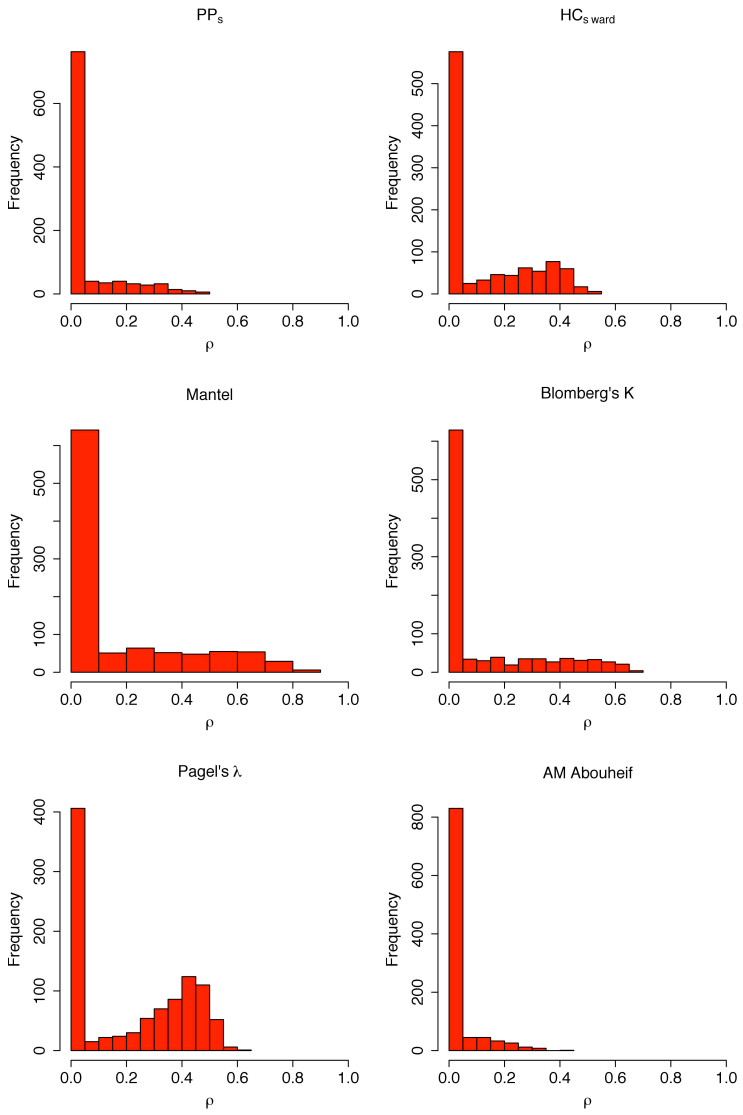


Figure 3.7: Estimated heritability in the entire Swiss dataset. Each plot represents a different statistic, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

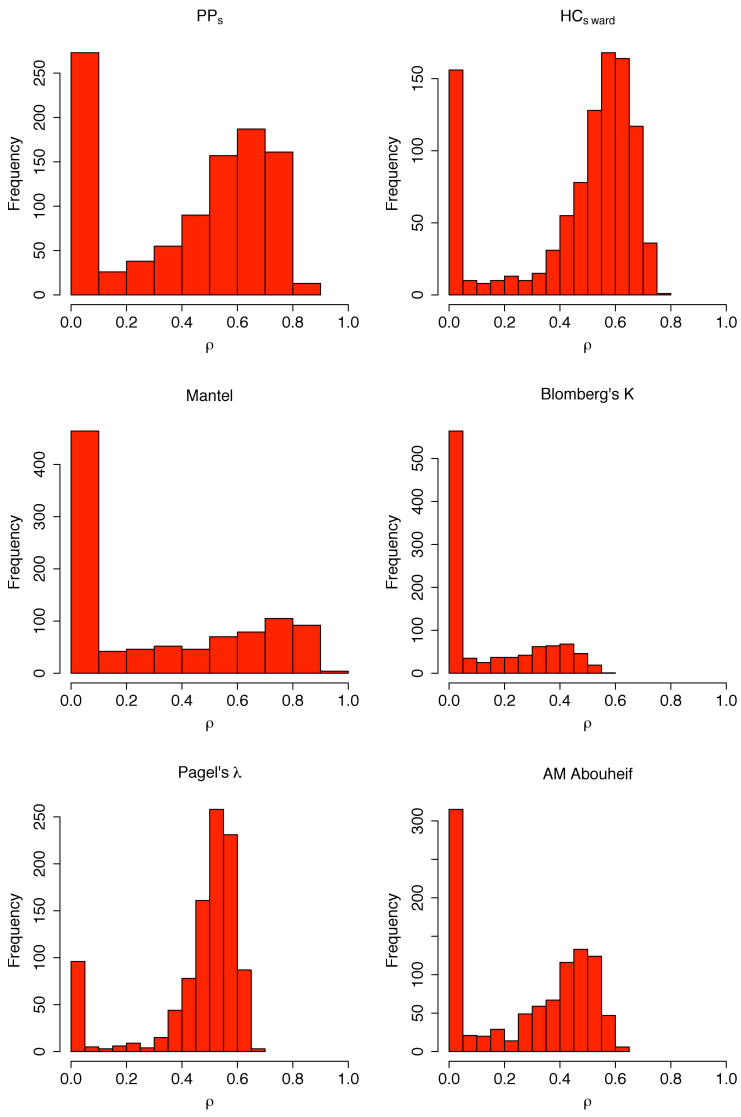


Figure 3.8: Estimated heritability in the MSM individuals with the Strict definition of viral load in the Swiss dataset. Each plot represents a different statistic, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

Table 3.7: Table showing the median ($\hat{\rho}$) and 95% confidence interval values for ρ of each subdivision of the Rakai data, for each of the 6 principle statistics for measuring heritability. The distribution of values of ρ were estimated by ABC and the shape for the entire dataset and subtypes A and D separately are shown in Figures 3.4, 3.5 and 3.6 . The column $p(0)$ shows the proportion of the values in the posterior distribution of ρ which are zero.

	Strong couples			**Strong/weak couples**			**Subtype A**			**Subtype D**			**All**		
	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**
PP_s	0.05	(0.00,0.68)	0.476	0.18	(0.00,0.64)	0.425	0.00	(0.00,0.60)	0.686	0.51	(0.00,0.77)	0.219	0.08	(0.00,0.57)	0.458
HC_s **ward**	0.03	(0.00,0.66)	0.483	0.00	(0.00,0.61)	0.509	0.00	(0.00,0.45)	0.741	0.00	(0.00,0.57)	0.560	0.00	(0.00,0.45)	0.676
Mantel	0.00	(0.00,0.94)	0.528	0.00	(0.00,0.71)	0.597	0.00	(0.00,0.69)	0.643	0.00	(0.00,0.73)	0.533	0.01	(0.00,0.84)	0.494
Garland's K	0.00	(0.00,0.82)	0.542	0.00	(0.00,0.76)	0.559	0.00	(0.00,0.43)	0.691	0.00	(0.00,0.52)	0.557	0.00	(0.00,0.59)	0.657
Pagel's λ	0.05	(0.00,0.66)	0.481	0.15	(0.00,0.66)	0.445	0.00	(0.00,0.38)	0.713	0.02	(0.00,0.52)	0.492	0.38	(0.00,0.59)	0.281
AM Abouheif	0.00	(0.00,0.67)	0.535	0.16	(0.00,0.63)	0.421	0.00	(0.00,0.36)	0.743	0.00	(0.00,0.36)	0.742	0.00	(0.00,0.41)	0.631

Table 3.8: Table showing the median ($\hat{\rho}$) and 95% confidence interval values for ρ of each subdivision of the Swiss data, for each of the 6 principle statistics for measuring heritability. The distribution of values of ρ were estimated by ABC and the shape is shown in Figure 3.8. The column $p(0)$ shows the proportion of the values in the posterior distribution of ρ which are zero.

	MSM Strict			**Strict**			**MSM**			**All**		
	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**	$\hat{\rho}$	**95% CI**	**p(0)**
PP_s	0.51	(0.00,0.78)	0.246	0.00	(0.00,0.43)	0.247	0.46	(0.00,0.74)	0.701	0.00	(0.00,0.36)	0.727
HC_s	0.55	(0.00,0.71)	0.145	0.24	(0.00,0.56)	0.200	0.48	(0.00,0.69)	0.377	0.00	(0.00,0.45)	0.550
Mantel	0.19	(0.00,0.87)	0.419	0.00	(0.00,0.68)	0.433	0.17	(0.00,0.93)	0.576	0.00	(0.00,0.72)	0.571
Garland's K	0.00	(0.00,0.48)	0.533	0.26	(0.00,0.80)	0.601	0.00	(0.00,0.64)	0.399	0.00	(0.00,0.60)	0.612
Pagel's λ	0.52	(0.00,0.63)	0.09	0.28	(0.00,0.55)	0.394	0.25	(0.00,0.52)	0.369	0.26	(0.00,0.53)	0.388
AM Abouheif	0.34	(0.00,0.57)	0.303	0.00	(0.00,0.43)	0.688	0.00	(0.00,0.39)	0.616	0.00	(0.00,0.23)	0.787

3.4.3 Testing the sensitivity of each statistic

The sensitivity for each statistic on a given tree was measured by simulating evolution along the tree with a particular input heritability, and then performing a randomisation test to establish whether the statistic could detect that level of heritability. Results from simulations on the entire Rakai and Swiss phylogenies are shown in Figures 3.9 and 3.10 respectively. The Mantel test appears to be insensitive to even maximum heritability, on either dataset. The PP, HC ward, AM Abouheif and Pagel's λ all perform similarly, requiring $\rho > 0.6$ to consistently detect an effect in the Rakai dataset, while in the Swiss dataset a consistent effect is detected when $\rho > 0.4$. Blomberg's K is less sensitive in either dataset, and the Mantel statistic is insensitive even for values of ρ close to 1. Interestingly, however, when the methods are applied to the tree derived from the Swiss MSM Strict individuals, the K statistic performs comparably with the other methods (Figure 3.11), and this is also the case when applied to the subtype A and subtype D trees from the Rakai data (Figures 3.12 and 3.13).

We also tested an alternative method of simulation which accounted for differences in branch length and allowed multiple generations between two adjacent nodes. This was applied to the Rakai data 3.14 using a clock rate of $1.094*10^{-2}$ substitutions per site per year, estimated from this data using Path-O-Gen [189] which calculates this rate from the phylogeny and tip dates. The generation time of 4.3 years was estimated from the Rakai dataset [103]. The sensitivity appears to be generally poorer than for the simpler method (Figure 3.9).

There were several available variations of both the HC and Abouheif-Moran tests, and throughout this work we refer to the default algorithms ('ward' in the HC method, and 'Abouheif' in the Abouheif-Moran method). We tested the sensitivy of the variations of these methods, which were the 'complete', 'single' and 'average' algorithms in HC (Figure 3.15), and 'oriAbouheif', 'sumDD', 'nNodes' and 'patristic' algorithms, (Figure 3.16). We found that the default methods performed equally well or better than any of the variations.

3.5 Discussion

In this study we developed and evaluated methods for detecting and measuring heritability of SPVL using phylogenetic relationships. We developed two new methods for performing this analysis, the Phylogenetic Pairs (PP) method and the Hierarchical Clustering (HC) method, and tested them, along with other previously used methods on two cohorts from Rakai [104], and Switzerland [4].

Simulations of SPVL evolution on the true phylogenies suggest that none of the methods can detect heritability lower than approximately ρ=0.4, and this

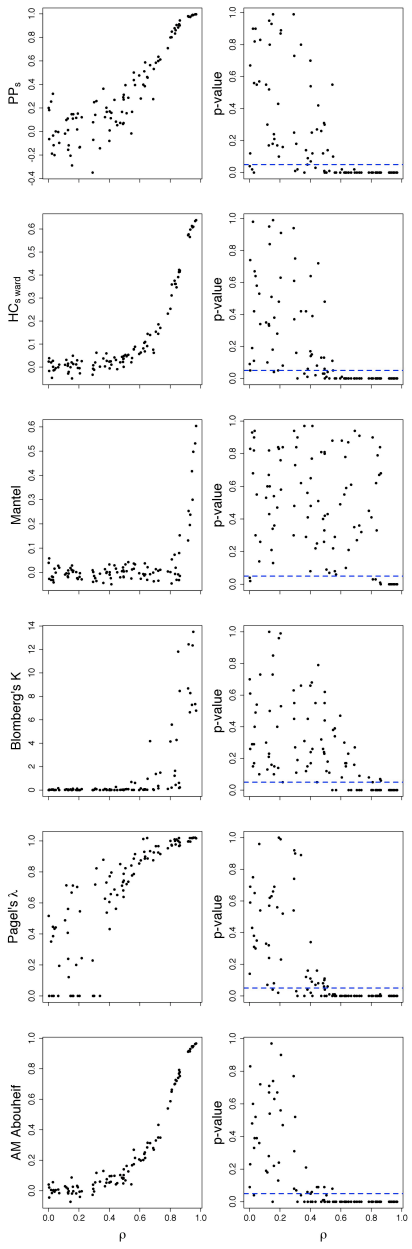


Figure 3.9: Six different statistics under comparison and p-values for each of 100 simulations on the entire Rakai phylogenetic tree. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

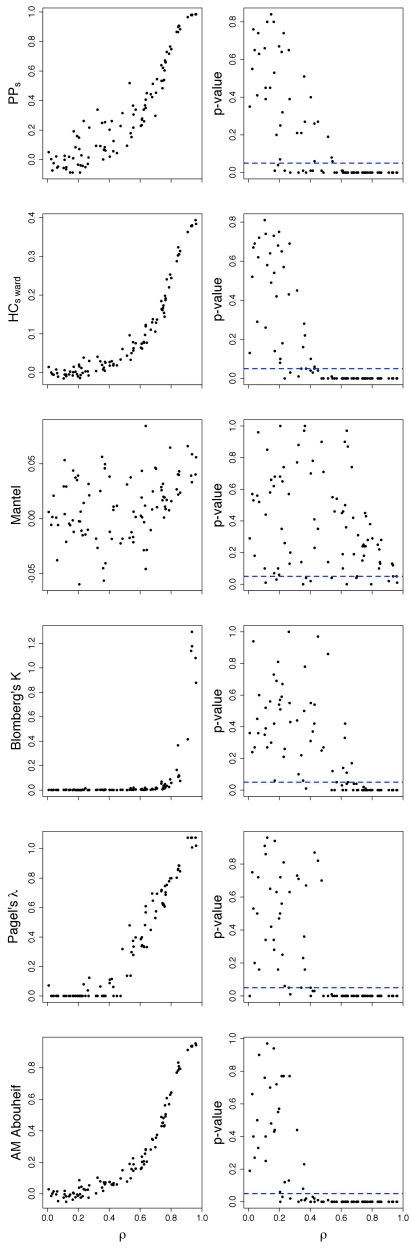


Figure 3.10: Six different statistics under comparison and p-values for each of 100 simulations on the entire Swiss phylogenetic tree. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

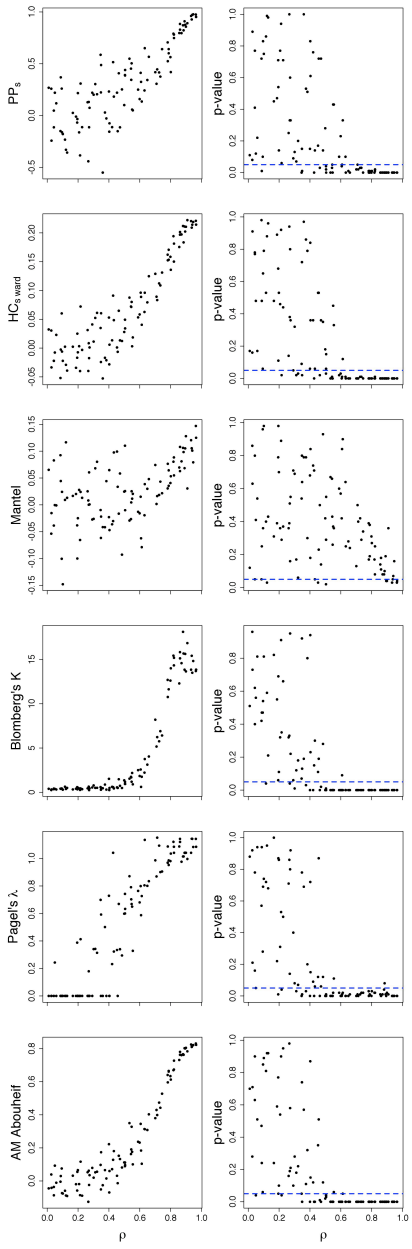


Figure 3.11: Six different statistics under comparison and p-values for each of 100 simulations on the tree derived from only the MSM Strict individuals in the Swiss cohort. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

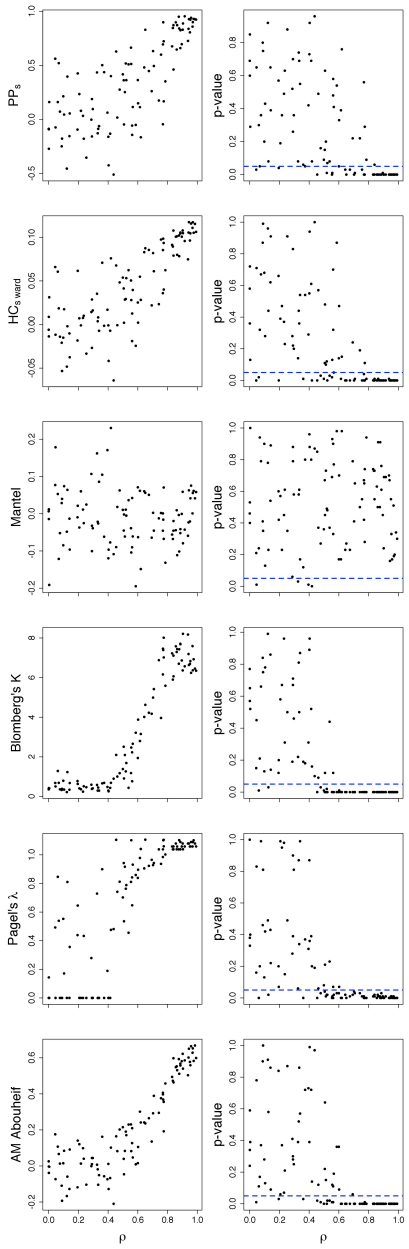


Figure 3.12: Six different statistics under comparison and p-values for each of 100 simulations on the tree constructed from the subtype A sequences in Rakai cohort. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

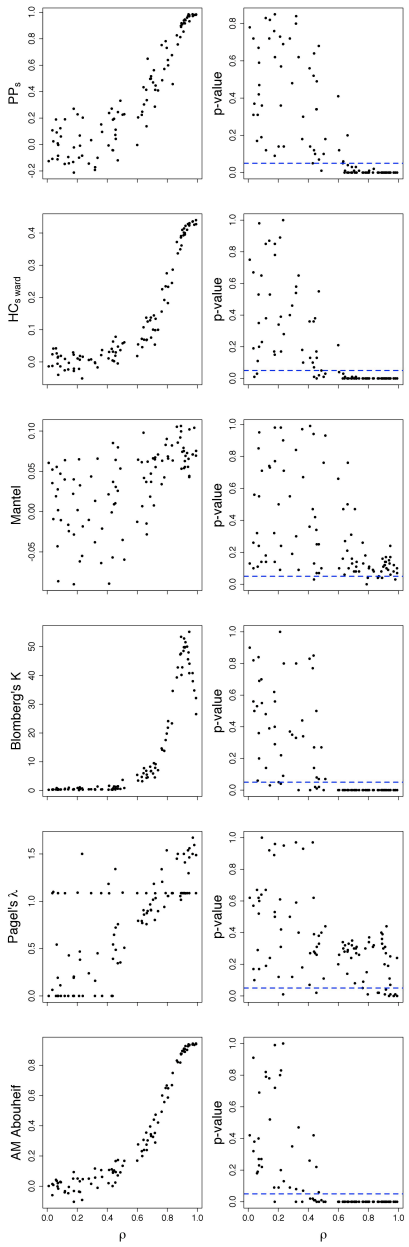


Figure 3.13: Six different statistics under comparison and p-values for each of 100 simulations on the tree constructed from the subtype D sequences in Rakai cohort. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

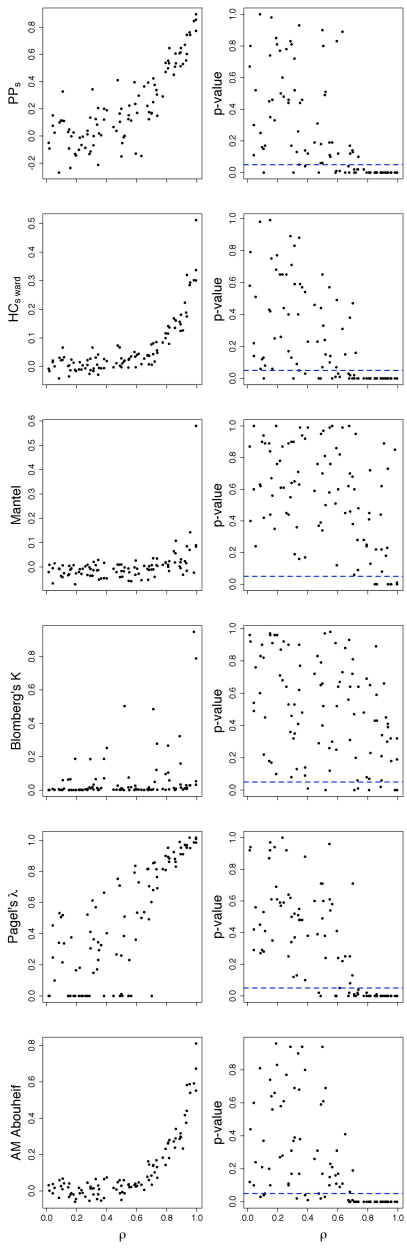


Figure 3.14: Six different statistics under comparison and p-values for each of 100 simulations on the entire Rakai phylogenetic tree. Each simulation is conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

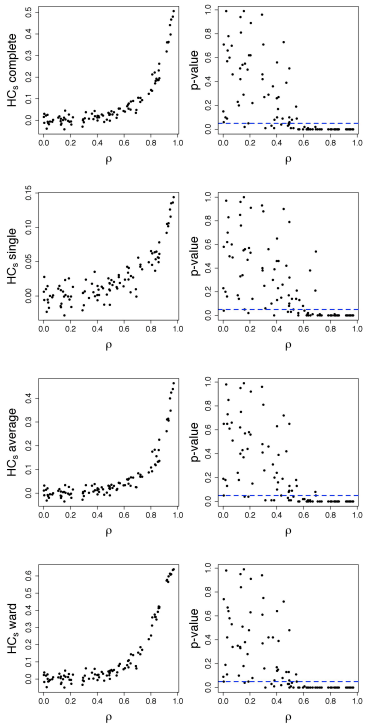


Figure 3.15: Statistics and p-values for each of four different variations of the Hierarchical Clustering method ('complete', 'single', 'average', 'ward') when 100 simulations were conducted on the entire Rakai phylogeny. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

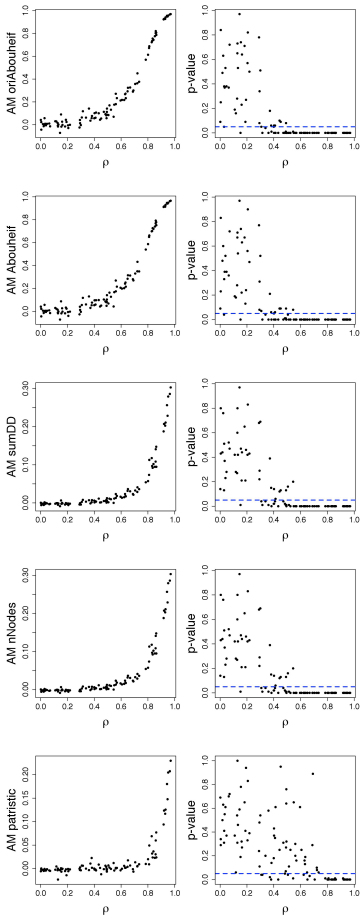


Figure 3.16: Statistics and p-values for each of five different variations of the Abouheif-Moran method ('oriAbouheif', 'Abouheif', 'sumDD', 'nNodes', 'patristic') when 100 simulations were conducted on the entire Rakai phylogeny. Each simulation was conducted with a given input heritability (ρ), and each dot represents a simulation. The left column gives the value of the statistic derived from each simulation, and the right column shows the p-value of that simulated data when the data are randomly permuted. The dotted line is at the p-value of 0.05 to indicate the boundary of formal significance.

threshold is higher in some phylogenies. Since most previous studies which measured heritability have produced estimates lower than this (Table 3.1), it suggests that phylogenetic methods are not adequate to detect heritability in HIV virulence. Modelling work in chapter one has suggested that heritability lower than this is enough to produce a substantial rate of evolution.

3.5.1 Results from the data

Heritability was detected consistently in the MSM Strict subdivision of the Swiss cohort (Table 3.5), supporting the previous study of this data which found significant and high heritability in that subdivision [4]. We also found that several statistics uncovered significant heritability in the Strict and MSM Liberal subdivisions. In contrast, aside from the strong and weakly linked couples subdivision, we did not observe significant heritability in the Rakai data (Table 3.4). Previous work found borderline significance in both [104]. However the small effect in strongly linked couples may be attributed to the very small sample size, approximately half that of the earlier study which also included pairs which shared only one of the two HIV sequence loci, *env* and *gag*. The PP statistic, which selects putative transmission pairs from a larger phylogeny, was the only one to detect an effect in the larger dataset and subtype D. It appears that the level of detectable heritability was similar in subtype D in comparison to the entire Rakai dataset, which it dominates.

The heritability estimates in the MSM Strict subdivision, although variable (Table 3.8), were higher than in any other subdivision in the Swiss cohort, or in the Rakai data (Table 3.7). None of the subdivisions exhibited heritability significantly different from zero by the ABC method. This suggests that the level of heritability in the data is relatively low, but this may also be ascribed to the strength of the hypothesis. The use of the prior on ρ in which 50% of the randomly selected values were $\rho = 0$ requires a substantial amount of heritability to differentiate the data from a zero heritability. This can be observed visually in figures 3.4 to 3.8, in which even though plots where the apparent signal for ρ is positive, there are still a large proportion of zeros in the posterior. This proportion of zeros is also explicitly reported in tables 3.7 and 3.8.

3.5.2 Comparison of statistics

The ability of the different statistics to detect and measure heritability was variable, and dependent on the choice of dataset and subdivision. Simulating data and performing randomization tests allowed us to test their ability to identify known heritability, and also to identify the effect of the phylogeny on the power of a statistic. In the simulations, the PP, HC complete, Pagel's λ and AM Abouheif statistics were comparably sensitive, detecting an effect at greater than approximately 0.4 heritability in the Swiss data, with AM Abouheif

slightly more sensitive (Figures 3.9 to 3.14). While sensitivity simulations provide insight, they are based on simplifying assumptions about evolution along phylogenies, so testing these methods on real data is essential.

Pagel's λ detected heritability in every subdivision of the Swiss data (Table 3.5), but in none of the Rakai subdivisions, where the PP method most effectively detected heritability by the randomisation test (Table 3.4). Although significant heritability was more in evidence in smaller subdivisions of the Swiss data, we did not find any apparent relationship between the power of the methods and the sample sizes studied (Table 3.6), although we did not explore wide ranges of the sample size. Our analysis suggests that the limitations of the phylogenic comparative methods when it comes to detecting low heritability values, are more due to the shape of the phylogeny than the size.

Apart from the PP statistic, all the statistics described here use the entire phylogeny, including topology and branch lengths (although some of the alternative Abouheif Moran statistics ignore the latter). These statistics use all of the available relationship data, and are representative of all levels of relatedness in the transmission network. However, the PP method performed well and arguably better than other methods, especially given its greater computational efficiency (it runs an order of magnitude faster than other methods, data not shown), which strongly recommends its use.

A further advantage of the PP and also the HC methods is the inclusion of cofactors of SPVL: ignoring these may lead to overestimates or underestimates of heritability if they associate or dissociate on the tree, respectively. However, in the absence of a method to include the effect of cofactors in simulations, this aspect of the PP or HC methods cannot be harnessed to measure ρ.

The relative success of the PP method suggest that most of the signal of heritability effectively lies in the tips of the tree, and that very little information lies deeper in the tree, since the PP method ignores this. The other phylogenetic methods which examine deeper relationships in the phylogeny, may thus be more successful for phylogenies that are less star-like than HIV. This contradicts recent unpublished work by Hodcroft et al. [101] who identified heritability in subtype B infections from the UK using pedigree analysis. They found that collapsing poorly supported nodes to polytomies did not greatly affect the estimate, suggesting that the external structure of the phylogeny is less important than the internal. Rigorously identifying which clades or levels of the phylogeny are responsible for heritability would be an interesting direction for future research.

3.5.3 Interpreting heritability

Traditional heritability is defined as h^2, the proportion of variance explained by the viral genotype 3.1. Both Blomberg's K and Pagel's λ have been linked directly with h^2, but they can both exceed one [75, 24] which h^2 can not. In

theory the result from the PP method also represents the amount of variation in SPVL which can be explained by that of another individual carrying genetically similar virus, and may therefore be linked directly to h^2, but this may not be true in practice. The other methods (HC, Mantel and Abouheif-Moran) also provide statistics which are not on the scale of h^2.

We have used the correlation coefficient, ρ, as the measure of heritability in this study as it is a simple and intuitive single measure of heritability and a nearly identical value has been used previously in this context [4]. It has also been used as the measure of heritability in most transmission pair studies [215, 94, 130] and a closely related measure used in the fourth such study [104]. The method of calibration we have used has allowed us to link the value ρ to each of the test statistics and interpret results from the true data in terms of ρ. Simulations in previous work have shown that an input variable closely related to ρ is strongly correlated with Blomberg's K and Pagel's λ [4] but in neither case is the relationship linear.

The correlation coefficient ρ is a simplified and practical measure of heritability. It is the proportion of variance in SPVL which does not change between two generations of infection and is therefore relatively easy to estimate. This differs from the traditional description of heritability, h^2, because the latter describes the proportion of variance in SPVL which is attributable to variance in the VVG at a given point in time. The relationship between these two values can be derived as follows.

Let the SPVLs of an index and a secondary case (V_I and V_S) be defined by the relationships in equations 3.22 and 3.23. The value G_0 represents the quantity of shared virulence genes between the two individuals, G_I and G_S represent their respective and differing genetic components, and E_I and E_S are the different environmental factors, or host effects, influencing the SPVL in each host. The value ϕ is thus defined as the proportion of the genetic variation which is inherited, in that it doesn't mutate between one infection and the next, and is therefore written as in equation 3.24. The individuals' genetic components are independently and identically distributed (distribution 3.25).

$$V_I = \phi G_0 + \sqrt{1-\phi^2} G_I + E_I \tag{3.22}$$

$$V_S = \phi G_0 + \sqrt{1-\phi^2} G_S + E_S \tag{3.23}$$

$$\phi = \frac{{\sigma_G}^2}{{\sigma_G}^2 + {\sigma_M}^2} \tag{3.24}$$

The genetic components are independently and identically distributed (distribution 3.25) with a variance given by that of the genotypes of the population. G_0 is also independently distributed according to the genotypic variance, but centered

on the mean SPVL of the population, $\hat{\mu}$ (distribution 3.26). The environmental components, E_I and E_S are also independently and identically distributed according to σ_E which quantifies the host effect (distribution 3.27).

$$G_I, G_S \sim N(0, {\sigma_G}^2) \tag{3.25}$$

$$G_0 \sim N(\bar{\mu}, {\sigma_G}^2) \tag{3.26}$$

$$E_I, E_S \sim N(0, {\sigma_E}^2) \tag{3.27}$$

By definition, the viral loads are distributed according to the population mean and variance (distribution 3.28), but they are not independent as they share genetic history in G_0. This shared genetic component is proportional to the covariance between the two values (equation 3.29).

$$V_I, V_S \sim N(\bar{\mu}, {\sigma_P}^2) \tag{3.28}$$

$$Cov(V_I, V_S) = \phi^2 Var(G_0) = \phi^2 {\sigma_G}^2 \tag{3.29}$$

The correlation coefficient ρ can then be written as a function of the covariance and variance of V_I and V_S, as in equation 3.30. This leaves the values of ρ and h^2 related by equation 3.31. Given that $0 < \phi < 1$ this implies that the correlation coefficient may underestimate, but never overestimate, the classical definition of heritability.

$$\rho = \frac{Cov(V_I, V_S)}{\sqrt{Var(V_I)Var(V_S)}} = \frac{\phi^2 {\sigma_G}^2}{{\sigma_P}^2} \tag{3.30}$$

$$h^2 = \frac{\rho}{\phi^2} \tag{3.31}$$

Intuitively, h^2 will be higher than ρ because it is a measure of the genotypic variance at an instant in time, so it does not account for ongoing mutation, whereas the correlation coefficient is lower as a result of incorporating this mutational variance. The discrepancy could be ignored in most organisms, where mutation occurs on a much longer timescale than transmission. However HIV can undergo substantial genetic change in a single generation, in spite of the rate of population level evolution being slower than the within-host rate [186].

In chapter one we estimated these two variables separately by fitting to data from the Rakai cohort, but this was done using identified transmission pairs. The maximum likelihood parameter estimates from that study (${\sigma_E}^2$=0.43, ${\sigma_M}^2$=0.01,

$\sigma_P{}^2$=0.61) suggest that the h^2 is 19% higher than ρ. This suggests that the estimates of ρ are considerable underestimates of h^2, and also that the methods tested in this study are insensitive to h^2 <0.47. Unfortunately the same parameter estimates were not available for the Swiss data, but it is safe to assume that the value of ρ underestimates h^2 in all cases.

The value ρ is also directly comparable [71] with the coefficient of determination, R^2, estimated in previous work which used an ANOVA to study transmission pairs [104]. It is also comparable with the analyses of simple linear regressions which report the correlation coefficient between index and secondary infected patients in couples (Table 3.1 [215, 94, 130]).

The simulation method we used is based on a simplistic model of evolution. Most results were based on the assumption that a single branch segment on the tree represents a single transmission, and that mutation of the VVG is the same, regardless of the branch length. This almost certainly underestimates the number of generations separating two tips, so the heritability estimates we achieve are underestimates, which must be born in mind when comparing studies. We address this problem with an alternative method of simulation which allows more generations to occur on longer branches 3.14, but this requires estimates for the generation time and the tree clock rate, both of which are uncertain (the tip dates explained only 4% of the genetic divergence), and to which the heritability estimates are highly sensitive (data not shown).

We also ignore selection in trait evolution. Previous work has identified a relationship between transmission fitness and SPVL [73], which suggests transmitting nodes on the tree are likely to have SPVL values closer to the optimum for transmission fitness. However, it is unclear how to extend this to the mechanism for onward transmission and evolution. Our model of evolution is based on an Ornstein-Uhlenbeck process [220, 65] which artificially maintains the population mean while allowing Brownian motion to take place, so it mimicks balancing rather than directional selection.

We also made the assumption that the infection at the root node was of mean SPVL, whereas an analysis of several studies has identified a general upward trend in the evolution of virulence [96]. However, this is not a significant trend, and without clearer evidence for the dating of the root node it would be misleading to begin simulating with a root node of lower virulence.

3.6 Conclusion

In this study we explored the ability of several statistics to represent heritability in order that it can be detected in simulated and real data and estimated with confidence intervals. The simulations suggested that heritability above 0.4 is consistently detectable. Several of the methods detected heritability in subdivisions of the two datasets where heritability has been previously found. However,

the estimates of heritability and the confidence intervals are variable depending on the statistic chosen. We recommend the Phylogenetic Pair method and Pagel's λ for use in detecting and estimating heritability.

We also demonstrated methods for measuring heritability in a standardised form for comparison between studies. This method gives a distribution for the estimate of the correlation coefficient between an index and a secondary case, ρ, using a method we have called numerical inversion. This value is strongly related to, and provides a lower bound for, traditional heritability, h^2. Confidence intervals were produced, but they exhibited significant differences from zero in no cases.

Chapter 4

Detection and measurement of heritability in HIV-1 virulence in Amsterdam.

4.1 Abstract

Introduction
Heritability of virulence in a pathogen refers to the amount of variation in the severity of infection which can be explained by viral factors, and which is carried by the virus from one infection to the next. Several studies have estimated heritability in SPVL, most by studying transmission pairs as they carry genetically similar virus, but more recently by using phylogenetic information to find related viruses.
Methods
We estimated heritability in a cohort from the Netherlands, for which details of the transmission network were unavailable. We used HIV sequence data to reconstruct the phylogeny and then applied two methods for identifying and measuring heritability, the Phylogenetic Pairs (PP) method and Pagel's λ transformation. Both methods have been previously evaluated as the most suitable for the task of measuring heritability of SPVL using only an HIV phylogeny to reconstruct the transmission network. However, previous work had suggested that neither these nor any other available method can detect heritability of $\rho < 0.4$, where ρ is the correlation coefficient between an index and a secondary case in a transmission pair. We looked for significant heritability in a one-tailed test by randomising the data on the tips. We estimated the value of the heritability by simulating with a range of values of ρ and then observing which output statistics matched that from the true data.

Results
The PP method did not find significant heritability in any subdivision of the dataset, but Pagel's λ found significant heritability when only men who have sex with men (MSM), originating from the Netherlands, with the strict definition of viral load variation, were included (p=0.038). Heritability was estimated at $\rho = 0.34$ (95% confidence interval 0.00 to 0.60).
Conclusion
This suggests that there is heritability of HIV virulence in this cohort, but it is close to the borderline of detectability with these methods.

4.2 Introduction

The virulence of HIV infection is highly variable between individuals. While much of that is due to host factors [64] there is growing evidence that the viral genotype is a source of this variance which is heritable from one infection to the next [215, 4, 94, 104, 130]. Several studies have also explored the evolution of HIV virulence over time. A meta-analysis of 12 such studies identified a significant increase in one virulence marker, baseline CD4 cell count, between 1985 and 2007, and an apparent but insignificant, upward trend in another, SPVL [96].

Quantifying the heritability of virulence is important in understanding HIV pathogenesis, as well as predicting whether and how rapidly the virus may evolve in response to changes in its life history. Such widespread changes may come about with public health interventions, changes in behaviour, or emergence into new populations. A general understanding of the heritability of this trait will shed light on the processes governing this population level evolution.

A number of studies have measured heritability of SPVL, and have produced variable estimates from around 20% to above 50% (reviewed in [158]). It is unclear whether these differences arise from the methods or from the populations themselves. In one study, applying the same method to different subsets of the data has found very different levels of heritability [4]. However, the uncertainty regarding the effect of the choice of method must be addressed.

Heritability is measured by comparing trait similarities between related organisms with trait similarities between unrelated organisms. One of the major challenges in measuring heritability in characteristics of HIV is in determining relatedness. The majority of previous studies [215, 94, 104, 130] have used transmission pairs, where one individual is known to have infected another, to compare viral characteristics. However, in order to identify a transmission pair, all of these studies have required personal information about sexual partnerships, as well as confirming that the transmission occurred within the couple by phylogenetic analysis. These studies have all been relatively small due to the difficulties in acquiring the behavioural data, which is both expensive and intrusive, leading to various sources of bias [67].

Another study successfully adopted a different approach, and used only the phylogeny to reconstruct the transmission network in a Swiss cohort [4]. This approach provides an opportunity to study heritability in more and larger datasets, since phylogenetic data is routinely collected [20] and it is expected that its availability will only increase.

In chapter three we evaluated several statistics which measure phylogenetic signal in SPVL. Phylogenetic signal refers to the extent to which the phylogeny is related to the trait values. This is a measure of heritability which can be estimated using only the phylogeny to reconstruct the relationships between the viruses carried by different infected individuals. These statistics were tested

on previously analysed cohorts from Rakai, Uganda [104] and Switzerland [4]. The results of the reanalysis of Swiss cohort were directly comparable with the original study and reproduced those results. The sensitivities of each statistic were also tested by simulation. We also developed a method for identifying, by simulation, the quantity of true heritability to which these statistics correspond.

Two methods, Pagel's λ [75] and the Phylogenetic Pairs (PP) method (developed in work described in the previous chapter) were identified as the most effective in detecting significant heritability. However, the simulations in that evaluation also implied that a heritability less than 0.4 is undetectable by any of the statistics tested.

In this study these two methods are applied to detecting heritability in a cohort from the Netherlands. Through simulation we also attempt to estimate the quantity of this heritability.

4.3 Methods

4.3.1 Netherlands data

In this study we analysed seroconverters in the Netherlands, for whom appropriate viral load measurements to calculate SPVL were available. All were infected with subtype B virus between 1985 and 2008. These individuals were among 906 whose SPVL was analysed previously and described elsewhere, with those individuals who seroconverted in or after 1996 provided by the ATHENA national observational cohort and those seroconverting before 1996 by the Amsterdam Cohort Studies (ACS) [90]. We have included only those 416 for whom appropriate genetic data were available.

We define SPVL as the mean of all log10 viral load measurements taken between 6 and 24 months after seroconversion, which is the date between the last negative and first positive diagnosis. We used a slightly earlier time window for calculating SPVL than 9 to 27 months as used in the earlier study by Gras et al. [90] on the basis that primary infection has finished by 6 months in a sufficiently high number of cases [103]. Following the definition from a previous analysis [4], we also categorised individuals according to their viral load variability. Individuals fulfilled the Strict criterion if the maximum difference between any pair of viral load measurements within 6-24 months was less than one log. Other characteristics of these individuals which were also used in this analysis are displayed in Table 4.1. The following categories of region of origin were combined for the ANOVA: Netherlands & Western Europe; Latin America & Caribbean; Eastern & Central Europe; South East Asia & Oceania. Missing values for each variable were included as a separate category.

Table 4.1: Characteristics of the Netherlands dataset.

Factor	Category	Number
Sex	Male	394 (95%)
	Female	22 (5%)
Region of origin	Netherlands	332 (80%)
	Western Europe	30 (7%)
	Latin America	12 (3%)
	Central Europe	7 (2%)
	Caribbean	7 (2%)
	South East Asia	5 (1%)
	Sub-Saharan Africa	5 (1%)
	Eastern Europe	4 (1%)
	Oceania	1 (0.2%)
	Unknown	13 (3%)
Risk category	MSM	348 (84%)
	IDU	11 (3%)
	HET	33 (8%)
	Blood transfusion	1 (0.2%)
	Unknown	23 (6%)
Age at seroconversion	15-24	34 (8%)
	25-29	85 (20%)
	30-39	170 (41%)
	40-49	96 (23%)
	50+	31 (7%)
Strict	Yes	246 (59%)
	No	170 (41%)

The phylogeny was reconstructed using a concatenate of the protease and reverse transcriptase components of the *pol* gene, resulting in a sequence of length 2064 nucleotides. We excluded codons strongly associated with drug resistance mutations since most samples were collected in the post-HAART era (after 1997). These were taken from the Stanford drug resistance database [195, 206] in 2009 as follows: in protease, amino acid positions 23, 24, 30, 32, 33, 46, 47, 48, 50, 53, 54, 76, 82, 84, 88, and 90; in reverse transcriptase, amino acid positions 41, 65, 67, 69, 70, 74, 75, 100, 101, 103, 106, 115, 151, 179, 181, 184, 188, 190, 210, 215, 230, and 236.

Four subtype C individuals from the same cohort were used as the outgroup. An additional 4 individuals from the cohort were used to construct the tree, but these were excluded after this point since data on their SPVL was lacking. The phylogenetic relationship was reconstructed using RAxML to search for the tree with the best likelihood score [213]. The substitution model used was

GTR+Γ+I i.e. the General Time Reversible model [216] with the Γ model of rate heterogeneity with 4 distinct rate categories, and a proportion of invariant sites [235]. This model gave the best Akaike Information Criterion score when the sequence data was tested with ModelTest [183]. All tree manipulation thereafter and production of the tree figure was done using the *ape* package [173] in R [217].

4.3.2 Methods for measuring heritability

We used two different methods for measuring heritability following their success in the previous chapter which evaluated several methods, each of which gave a single statistic as a measure of the phylogenetic signal. Details of the two methods, Pagel's λ and the Phylogenetic Pairs (PP) method are described in the previous chapter.

The same method is also used to measure the significance of each of these statistics: the data are randomly permuted 1000 times and the statistics recalculated to give a null distribution of the statistics. The data are also simulated under a range of values of the heritability, ρ, in order to match the true value of the statistic with known input values.

The data are simulated using a method in which the extent of trait evolution between any two nodes on the tree is drawn from the same distribution. An alternative method is also used, as in the previous chapter, in which the rate of trait evolution is proportional to the branch length, and is calculated according to the clock rate of the tree, as calculated using Path-O-Gen [189], and the generation time for MSM, which was taken from previous work on this dataset [22].

4.4 Results

Two tests, Pagel's λ and Phylogenetic Pairs (PP), were applied to detect and measure heritability in each subdivision of the data. In the PP method, developed in chapter three, transmission pairs were identified on the phylogeny (Figure 4.1).

The presence of significant heritability was tested using randomisation (Table 4.2), which found significance only in one subdivision (MSM from the Netherlands with Strict viral loads) and using one statistic (Pagel's λ). No effect was found when the PP method was used. The relationship between the heritability (ρ) and each output statistic was estimate by simulating evolution on the phylogeny, calculating the relevant statistic, and then linking this to the true value using Approximate Bayesian Computation in order to discover the approximate posterior distribution of the true heritability. The likely distributions of ρ in

the true data are shown in histograms (Figures 4.2 and 4.3). These distributions were also used to identify the median value of ρ, as a best estimate of the true heritability (Table 4.2). Confidence intervals on this value could also be identified from this distribution.

An alternative method of simulation which accounted for the different lengths of branches on the phylogeny was also used to estimate values of ρ (Table 4.3). Figure 4.4 shows the comparison between the original and the alternative method. The substitution rate per year (3.62×10^{-3}) was calculated using Path-O-Gen [189] and divided by the number of years per generation ($T_g = 2.1$ years, [22]) to give generations per unit branch length on the tree. The Path-O-Gen package uses the tip dates and the phylogeny to estimate the clock rate on the tree, and in this case the tip dates explained 25% of the genetic divergence in the MSM subdivision. Since the generation time estimate was estimated in a cohort of MSM in the Netherlands, this method could only be applied to the MSM group.

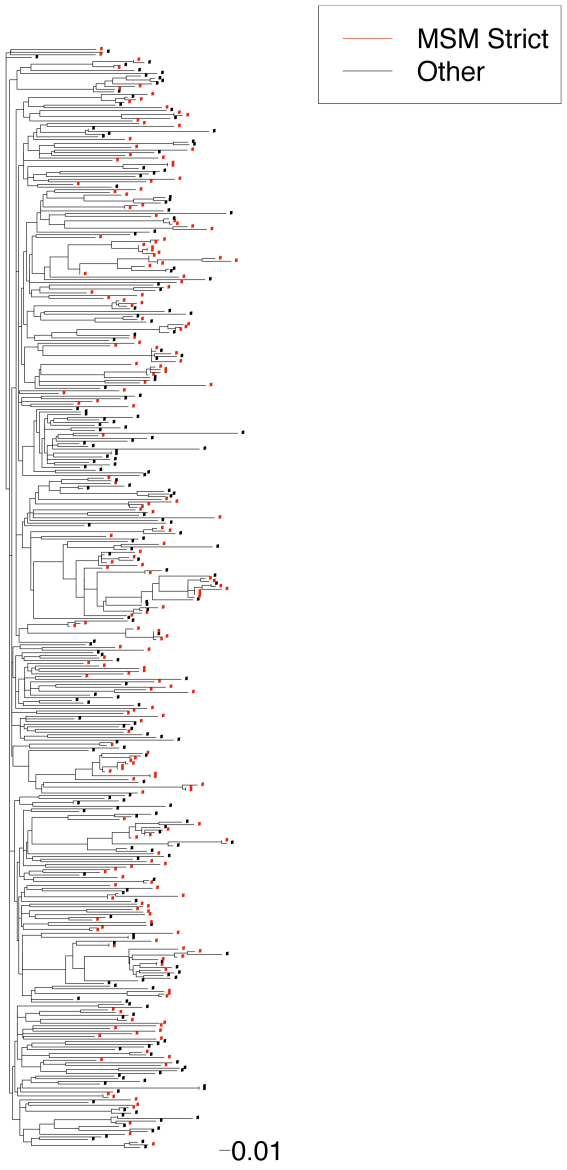


Figure 4.1: The best tree inferred from the entire Netherlands *pol* sequence dataset. Branch lengths separating the phylogenetic pairs are marked in red, with a single pair separated by a large vertical distance marked with red tildas instead.

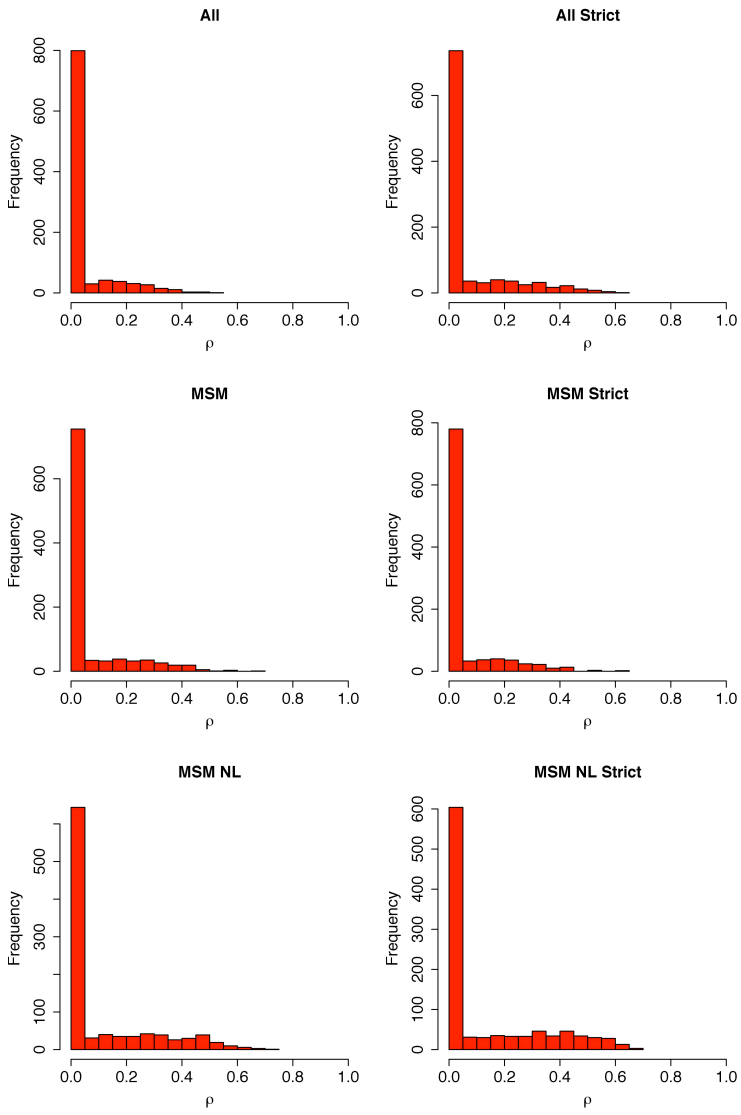


Figure 4.2: Estimating heritability using the Phylogenetic Pairs method. Each plot represents a different subdivision of the dataset, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic (PP_s) which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

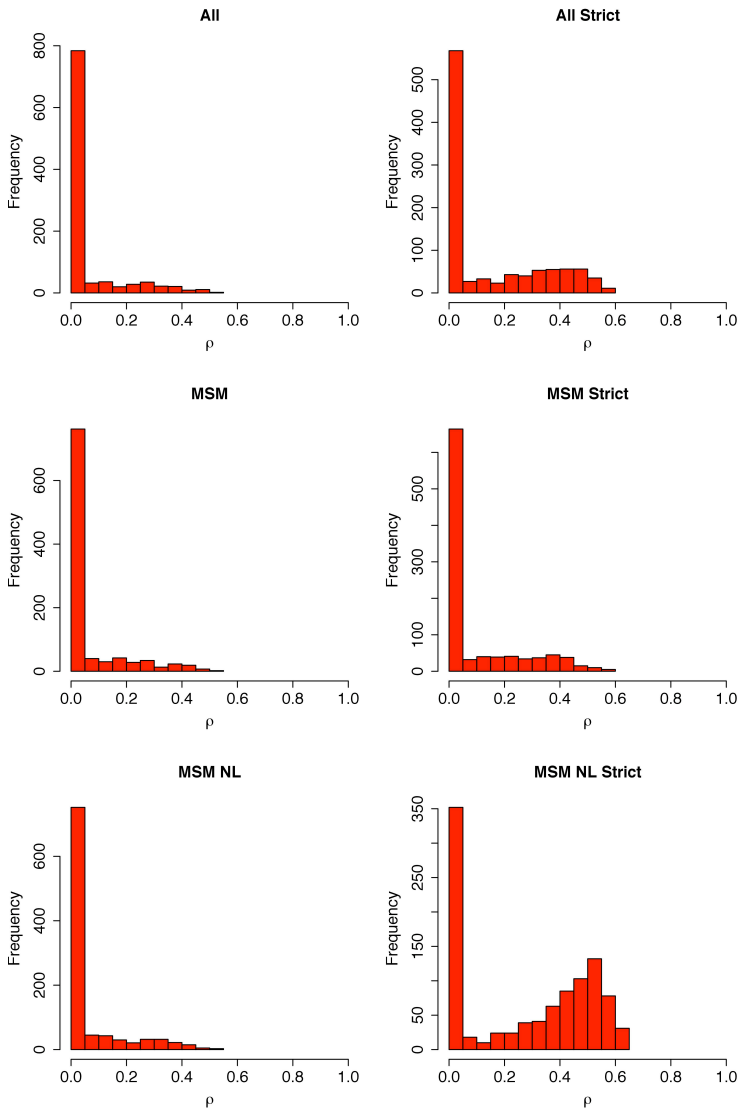


Figure 4.3: Estimating heritability using the Pagel's λ method. Each plot represents a different subdivision of the dataset, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic (λ) which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

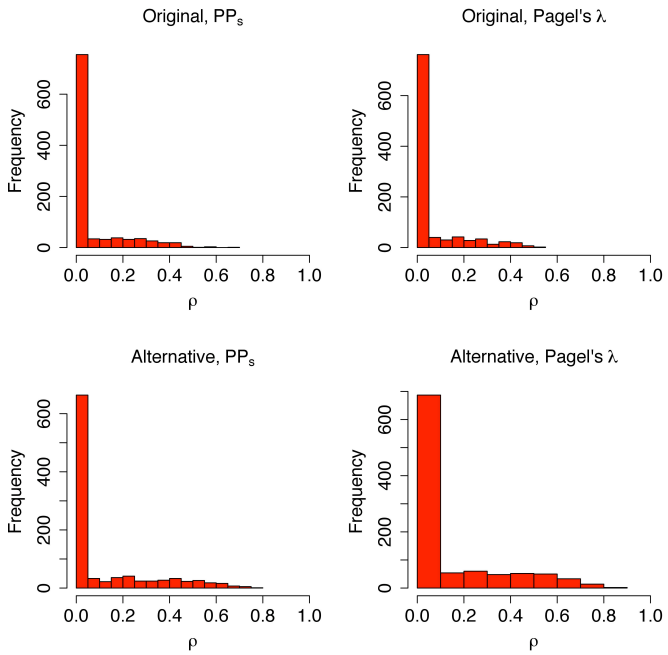


Figure 4.4: Estimating heritability, using an alternative method of simulation which accounts for branch lengths on the phylogeny, in the MSM using Pagel's λ and the PP method. In this plot we compare the distributions resulting from the two models, Original and Alternative. Each plot represents a different subdivision of the dataset, written above the plot. The plot displays the distribution of 1000 values of ρ which result in a simulation and a subsequent test statistic which was sufficiently close to the value of the true statistic. The trial values of ρ were taken from a prior distribution in which 50% of the values were 0 and 50% were drawn from a uniform distribution between 0 and 1.

Table 4.2: Results from significance tests for heritability in the data, and estimated heritability, for each subdivision of the data, and using two different tests, phylogenetic pairs (PP) and Pagel's λ. The sample size for each subdivision is given by n. The true statistic (Z) and p-value (p) are given for each significance test. Heritability was estimated as the median ($\hat{\rho}$) of the approximate Bayesian posterior distribution, of which 95% confidence intervals are also provided. The proportion of values of ρ in the posterior distribution which are equal to zero are given by p(0).

		All	**All strict**	**MSM**	**MSM strict**	**MSM NL**	**MSM NL strict**
Pagel's λ	n	414	246	348	211	286	174
	Z	0.0001	0.0978	0.0001	0.0323	0.0001	0.2628
	p	0.824	0.086	0.827	0.200	0.718	0.038
	$\tilde{\rho}$	0	0	0	0	0	0.34
	95% CI	(0.00,0.39)	(0.00,0.52)	(0.00,0.41)	(0.00,0.47)	(0.00,0.39)	(0.00,0.60)
	p(0)	0.748	0.540	0.721	0.627	0.715	0.338
PP	n^*	164	94	146	86	112	62
	Z	-0.0627	0.2193	0.0763	0.2047	0.0633	0.2277
	p	0.703	0.111	0.272	0.128	0.327	0.130
	$\tilde{\rho}$	0	0	0	0	0	0
	95% CI	(0.00,0.33)	(0.00,0.45)	(0.00,0.41)	(0.00,0.36)	(0.00,0.53)	(0.00,0.57)
	p(0)	0.764	0.707	0.709	0.740	0.614	0.575

* refers to the total number of individuals included in the phylogenetic pairs which are tested.

Table 4.3: Heritability estimates derived from the approximate Bayesian posterior distribution for the MSM subdivision of the data, using both the Original and Alternative method of SPVL simulation. The median ($\hat{\rho}$) is the best estimate of ρ, with 95% confidence intervals provided. The proportion of values of ρ in the posterior distribution which are equal to zero are given in column "p(0)".

		Original	Alternative
Pagel's λ	$\hat{\rho}$	0	0
	95% CI	(0.00,0.41)	(0.00,0.68)
	p(0)	0.017	0.02
PP	$\hat{\rho}$	0	0
	95% CI	(0.00,0.41)	(0.00,0.61)
	p(0)	0.018	0.019

4.5 Discussion

Results from the previous chapter suggested that phylogenetic comparative methods are sensitive enough only to detect a fairly large heritability (ρ >0.4). In this chapter, two different statistics were used independently to detect heritability, and only a single positive result was found with one of the statistics (Pagel's λ) in one out of six subdivisions of the data. This suggests that heritability in this dataset is close to this detection borderline. Indeed the best estimate of heritability which corresponded to the one significant test, was ρ=0.34 which is slightly below the range at which detection would be expected. This level was also within the range of previous estimates [158].

Our exclusion criteria in choosing subdivisions were supported by previous work [63, 4]. We included a further exclusion criterion by also running analyses excluding individuals not originally from the Netherlands, in order to further remove confounding factors associated with nationality.

Unlike in the previous chapter and in previous work [4, 104] there was no general decrease in p-value between the larger and more disparate datasets moving towards the smaller and tighter ones. However, the subdivisions which excluded non-strict individuals showed much lower p-values than those which included these individuals with variable viral loads. This trend can also be observed in the histograms of the heritability distribution derived from Pagel's λ, which are also apparently higher with this exclusion (Figure 4.3), although the same is not true when the PP method is used (Figure 4.2).

Understanding why the strict population seems to exhibit higher heritability

than the population in general may be an important step in both estimating and understanding the mechanisms which cause virulence heritability. There may be reasons for longitudinal viral load variability in certain individuals which are irrelevant to viral biology and therefore the study of the traits heritability, such as unbiased measurement error. However, it is also possible variability implies relevant viral or human characteristics. Simple models fitted to longitudinal viral load data within patients suggests that fluctuations cannot be explained by random fluctuation around a fixed or an exponentially growing level [149]. Fluctuating viral load has been associated with untreated sexually transmitted co-infections [13] and this effect may mask the effect of the viral genotype. Similarly, if the fluctuations are the result of human characteristics, then while these are of interest, their exclusion from certain analyses is justified.

It would also be of interest to investigate why there is more apparent heritability in the Swiss HIV Seroconverters Cohort than in this cohort which has a similar risk group distribution. In the Swiss dataset, the mean SPVL of the strict group is significantly higher than that of the rest (p=0.0004), and the same is true of the MSM over non-MSM (p=0.003) [4]. The same is not true of these two categories in the Netherlands cohort (data not shown). Intuition would suggest that a group exhibiting significantly more virulence than the rest would be more homogeneous and therefore heritability would be harder to identify within that group, but the comparison of these two cohorts suggests otherwise.

The differences in results between the two measures of phylogenetic signal imply that the choice of method is important. The λ statistic has the advantage that it can incorporate all of the phylogenetic information and analyses the entire sample. The PP method analyses only pairs of tips, and most individuals (60%) are excluded from the analysis as they do not form pairs. However, the PP method has the advantage that it is able to include cofactors of SPVL. If cofactors of SPVL cluster together on the phylogeny then the heritability value is likely to be an overestimate if these are not accounted for. Similarly, dissociative clustering may have the reverse effect and lead to underestimates of heritability. The latter effect was observed in the analysis of transmission pairs in the Rakai cohort [104].

4.6 Conclusion

Heritability has generally not been found in these data, although the tests may be insufficiently sensitive. Borderline heritability was found in the most homogeneous subdivision, MSM who originate in the Netherlands with strict viral loads, which supports a previous result on Swiss data. Subdivisions with strict viral loads appear to exhibit higher heritability.

Chapter 5

Final Discussion

In this work I have addressed how the need for transmission between hosts selects for virulence in a pathogen, either by selection of viral genotypes which are most capable of founding an epidemic, or by selection of those which are most capable of transmission during an ongoing epidemic or endemic situation. I have also attempted to quantify the heritability of virulence between infections and to evaluate different methods for achieving this using phylogenies.

In the first half of this thesis I describe the analysis of a model which simulated between-host evolution of SPVL in a transmitting population. Initially a deterministic model was used, as described in chapter one, incorporating previous work relating the SPVL to transmission and prognosis. The model supported this earlier work which hypothesised that viral genotypes which produce intermediate virulence would eventually dominate [73]. The hypothesis was also supported by the model described in chapter two which incorporated stochastic effects and more realistic sexual mixing behaviour.

A deterministic model was used to model the evolution of HIV throughout its history, and it is reasonable to assume that the host population size was large enough that a deterministic model is appropriate. However it could also be argued that a complex population structure such as has been documented in a European phylogeographic study [175] renders a deterministic approximation invalid. Nevertheless it was encouraging that the heritability in SPVL as estimated by this model was very similar to the estimate of heritability by direct analysis of the same dataset [104]. The apparent fit of the modelled SPVL to that of true index and secondary cases supports the models validity and subsequently the assumptions, such as the trade-off hypothesis, on which it is based.

The deterministic model was also based on the assumption that transmission occurred universally through serially monogamous partnerships, and transmission rates were adjusted accordingly with a fixed partner change rate. The stochastic model was therefore a timely improvement as it incorporated not only small

sample sizes, but also heterogeneous sexual behaviour, including long term partnerships but also a high-risk core group and casual transmission, which proved to be very important in determining epidemic outcomes.

The approach to parameter estimation in chapter two also considered a much wider range of scenarios. Although the parameters governing the relationship between SPVL, transmission rate and duration of infection were taken from literature in both models (Fraser, Hollingsworth et al. 2007; Hollingsworth, Anderson et al. 2008), there were numerous other parameters in the stochastic model which governed population dynamics and sexual behaviour. These were allowed to vary over a wide range and the conclusions of the model are robust to this variety. Since the stochastic model was essentially simulating the first emergence of HIV into a population about which we know little, this approach was considered more appropriate than rigorous model fitting. Instead, the model was fit by selecting parameter values which produced a recognisable outcome in terms of prevalence after 100 years.

Uncertainty in the parameters governing the dependencies of transmission rate and duration of infection on SPVL was addressed by bootstrapping in chapter 1, and this demonstrated that the hypothesis is broadly supported over a range of these parameters. The principle model fitting in chapter one, however, did not incorporate this uncertainty, in favour of rigorous fitting to recently collected data with respect to just two parameters, the environmental and mutational variance. These are important values governing between-host selection on virulence, and they provide a link between the within-host dynamics and the outcome of these in terms of transmission. While these sources of variance are a simplification, since they assume that within-host processes introduce only non-directional variance, the model provides an appropriate framework for exploring greater complexity.

In the first part of this thesis I considered a model of evolution in which the host and the environment both play an explicit role, and the viral genotype evolves independently of the host effect, provided it is transmitted. This may be an oversimplification as within-host selection on virulence may be different in different hosts, and this may have an impact on the transmitted viral genotype.

In the second part I considered a different evolutionary model which did not incorporate the host and viral effects separately. This is also a simplification because it is highly plausible that severity may vary according to host factors without this affecting the viral genotype when it is transmitted to the next generation. A model which incorporated within-host selection and dynamics would be able to reconcile these two scenarios and produce a more comprehensive picture of HIV virulence evolution. Further understanding of the within-host processes through experimental and theoretical work is an exciting area of potential study.

Throughout this work it has been assumed that individuals become infected

once. However, superinfection is relatively common (reviewed in [129]), and arguably individuals should not be removed from the susceptible pool once they are infected. However, more complicated and nested models would be required to incorporate phenomena related to superinfection such as increased virulence, susceptibility to superinfection and the possibility of recombination (reviewed in [127]).

Both transmission models also provide a framework for understanding how virulence may adapt to changes in transmission dynamics. This was briefly touched upon in chapter one where the effect of epidemic growth on virulence was illustrated, as it has been elsewhere [18, 73, 25]. Interventions for prevention may affect the average transmission parameters and drive selection in viral virulence. Both models provide a framework for predicting whether such an effect might occur, although the individual-based model from chapter two would allow greater versatility in incorporating targeted, risk-group specific interventions, and would also allow an exploration of whether such an effect could be seen above stochastic fluctuations.

These models may also be used to understand trends in SPVL evolution in recent times. The similarity between the two models suggests that the virulence which is selected by ongoing transmission (chapter one) is very similar to that selected by a founder event (chapter two). The fact that evolution of virulence has been identified in some areas [96] implies that either the founder virus in an evolving population was suboptimal by chance, or it was previously subject to different selective pressures with different optimal virulence. The different trends in SPVL evolution between nations and risk groups provides an opportunity to understand where differences in selective pressure may be responsible for evolutionary patterns. The models also provide a framework for understanding subtype replacement, which may play a major role in virulence evolution, as discussed in chapter one.

A more direct approach to measuring the influence of the viral genotype on SPVL was taken in the second half of this thesis, chapters three and four. As mentioned, several studies have attempted to quantify the extent to which virulence is heritable between infections by comparing SPVL within transmission pairs [215, 94, 104, 130]. The application of phylogenetic comparative analysis to this task [4] has the potential for estimating heritability in a wide range of scenarios where sexual behaviour or contact tracing data are unavailable. This may also be usefully applied to other traits such as drug resistance or immune escape.

Since drug resistance testing is recommended for all HIV infections [99], a large amount of HIV genetic data is and will become available. Techniques such as those discussed in chapters three and four which harness this to identify viral correlates of infection such as virulence may become increasingly useful. However, the results from chapter three suggested that these techniques are not sensitive to low levels of heritability.

The limitations of the techniques for measuring heritability are evident from the study on data from the Netherlands. These results, presented in chapter four, are ambiguous since one of two methods used to estimate heritability found a significant result. There may be substantial heritability in this cohort, sufficient for natural selection by transmission to play a major role in shaping the virulence landscape. A study of virulence markers in the Netherlands found that virulence has increased over time in a dataset [90] containing a large proportion of the individuals from whom we measured heritability. Although there are other possible explanations for this trend, it does provide support for a substantial degree of heritability which has enabled the virulence in this population to evolve.

Further work is required to develop new methods which are more sensitive to lower levels of heritability, and to identify in which populations this type of analysis is appropriate. This and earlier work [4] suggests that it is easier to measure heritability on more homogeneous but not on smaller phylogenies. There are many interesting questions as to how the size, balance, and other measures of the shape of a phylogeny affect the ability to measure heritability. In this study we chose to simulate evolution on phylogenies inferred from real data, but simulating on artificial trees would provide an explicit link between phylogenetic shape and heritability measures. Recent work has attempted to understand how the contact structure of an infection contributes to the shape of the phylogeny [137], and as genetic data and resources for simulation increase this becomes a more fruitful area for study.

This thesis (chapters three and four) and previous work [4] have reported that populations with unvarying viral loads exhibit more heritability, and there may be other such attributes of a population which increase the heritability. Work on the within-host mechanisms which govern the heritability of virulence in HIV (e.g. [144]) may benefit and benefit from these between-host analyses of heritability.

A recent unpublished study by Hodcroft et al. [101] found heritability in SPVL in subtype B-infected individuals from the UK. This study found that the heritability measure was largely unaffected when some of the external structure of the tree was ignored. Our evaluation work (chapter three) found instead that a method which ignored the internal structure of the tree performed equally well as those which use all levels of the tree. Whether this difference results from differences in the dataset or whether this is an artefact of either method requires further investigation. This avenue may deepen understanding of the link between genotypic and phenotypic evolution in HIV and other pathogens.

Several recent studies have addressed population level adaptations in HIV. For example, van Manen et al. identified that human correlates of resistance to HIV have decreased over time, suggesting viral adaptation to immune responses [146]. Pingen et al. review evidence for population level drug resistance evolution [182]. The fixation of adaptations at the population level may reflect a tolerance of the population for within-host adaptations, or alternatively arise from directional

selection by the population against a background of neutral or contrary selection within the host. Studying virulence evolution in HIV provides an opportunity to study these different levels of selection in a trait which is relevant in all pathogens.

This work has not considered the details of the viral genotype which contribute to virulence, but the availability of deep sequencing data, better studies of animal models, and mechanistic understanding of HIV virulence will enable the viral genetic causes of virulence and the evolutionary outcomes of between-host selection to be predicted with finer precision. Kouyos et al. have developed techniques for quantifying the effects of individual alleles and epistatic effects on fitness in HIV-1 [126], and the rising availability of data will increase the possibilities for this approach.

5.1 Conclusion

The interaction between host populations and pathogens produces very complex dynamics which are important and interesting for epidemiologists and evolutionary biologists. The rapid evolution of HIV on the timescale of transmission creates unusual problems and questions which require new techniques to answer them. In this work I have explored the evolution and emergence of HIV in human populations with respect to virulence. I have also tested and applied techniques for measuring viral control of virulence and explored their capacity for analysing viral traits and a growing collection of phylogenetic data.

Bibliography

[1] E. Abouheif. "A method for testing the assumption of phylogenetic independence in comparative data". In: *Evolutionary Ecology Research* 1.8 (1999), pp. 895–909.

[2] L. J. Abu-Raddad, P. Patnaik, and J. G. Kublin. "Dual infection with HIV and malaria fuels the spread of both diseases in sub-Saharan Africa". In: *Science* 314.5805 (2006), pp. 1603–1606.

[3] S. Alizon, F. Luciani, and R. R. Regoes. "Epidemiological and clinical consequences of within-host evolution". In: *Trends Microbiol* 19.1 (2011), pp. 24–32.

[4] S. Alizon et al. "Phylogenetic Approach Reveals That Virus Genotype Largely Determines HIV Set-Point Viral Load". In: *Plos Pathogens* 6.9 (2010), e1001123.

[5] S. Alizon et al. "Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future". In: *Journal of Evolutionary Biology* 22.2 (2009), pp. 245–259.

[6] H. K. Altes, R. M. Ribeiro, and R. J. de Boer. "The race between initial T-helper expansion and virus growth upon HIV infection influences polyclonality of the response and viral set-point". In: *Proceedings of the Royal Society of London Series B-Biological Sciences* 270.1522 (2003), pp. 1349–1358.

[7] M. Altfeld et al. "Influence of HLA-B57 on clinical presentation and viral control during acute HIV-1 infection". In: *AIDS* 17.18 (2003), pp. 2581–2591.

[8] R. M. Anderson and R. M. May. "Coevolution of Hosts and Parasites". In: *Parasitology* 85.Oct (1982), pp. 411–426.

[9] L. Andreoletti et al. "Genetic and phenotypic features of blood and genital viral populations of clinically asymptomatic and antiretroviral-treatment-naive clade A human immunodeficiency virus type 1-infected women". In: *Journal of Clinical Microbiology* 45.6 (2007), pp. 1838–1842.

[10] R. Antia et al. "The role of evolution in the emergence of infectious diseases". In: *Nature* 426.6967 (2003), pp. 658–661.

[11] K. K. Arien et al. “Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time”. In: *AIDS* 19.15 (2005), pp. 1555–64.

[12] L. J. Ashton et al. “HIV-Infection in Recipients of Blood Products from Donors with Known Duration of Infection”. In: *Lancet* 344.8924 (1994), pp. 718–720.

[13] M. C. Atkins et al. “Fluctuations of HIV load in semen of HIV positive patients with newly acquired sexually transmitted diseases”. In: *British Medical Journal* 313.7053 (1996), pp. 341–342.

[14] S. Attia et al. “Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis”. In: *AIDS* 23.11 (2009), pp. 1397–1404.

[15] A. Babiker et al. “Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis”. In: *Lancet* 355.9210 (2000), pp. 1131–1137.

[16] J. M. Baeten et al. “HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads”. In: *J Infect Dis* 195.8 (2007), pp. 1177–80.

[17] J. M. Baeten et al. “Male circumcision and risk of male-to-female HIV-1 transmission: a multinational prospective study in African HIV-1-serodiscordant couples”. In: *AIDS* 24.5 (2010), pp. 737–744.

[18] C. L. Ball, M. A. Gilchrist, and D. Coombs. “Modeling within-host evolution of HIV: Mutation, competition and strain replacement”. In: *Bulletin of Mathematical Biology* 69.7 (2007). 211CW Times Cited:21 Cited References Count:46, pp. 2361–2385.

[19] G. Bello et al. “Plasma viral load threshold for sustaining intrahost HIV type 1 evolution”. In: *AIDS Research and Human Retroviruses* 23.10 (2007), pp. 1242–1250.

[20] D. E. Bennett et al. “The World Health Organization’s global strategy for prevention and assessment of HIV drug resistance”. In: *Antiviral Therapy* 13 (2008), pp. 1–13.

[21] D. Bezemer et al. *Evolution of HIV-1 Set-Point Viral Load within Transmission Networks*. 2012.

[22] D. Bezemer et al. “Transmission networks of HIV-1 among men having sex with men in the Netherlands”. In: *AIDS* 24.2 (2010), pp. 271–282.

[23] N. Bhattarai and J. T. Stapleton. “GB virus C: the good boy virus?” In: *Trends Microbiol* 20.3 (2012), pp. 124–30.

[24] S. P. Blomberg, T. Garland, and A. R. Ives. “Testing for phylogenetic signal in comparative data: Behavioral traits are more labile”. In: *Evolution* 57.4 (2003), pp. 717–745.

[25] B. M. Bolker, A. Nanda, and D. Shah. “Transient virulence of emerging pathogens”. In: *J R Soc Interface* 7.46 (2010), pp. 811–22.

[26] S. Bonhoeffer et al. “Glancing behind virus load variation in HIV-1 infection”. In: *Trends in Microbiology* 11.11 (2003), pp. 499–504.

[27] P. Borrow et al. “Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus”. In: *Nature Medicine* 3.2 (1997), pp. 205–11.

[28] P. Borrow et al. “Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection”. In: *Journal of Virology* 68.9 (1994), pp. 6103–10.

[29] J. M. Brenchley et al. “CD4(+) T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract”. In: *Journal of Experimental Medicine* 200.6 (2004), pp. 749–759.

[30] J. M. Brenchley et al. “Microbial translocation is a cause of systemic immune activation in chronic HIV infection”. In: *Nature Medicine* 12.12 (2006), pp. 1365–1371.

[31] M. Carrington and S. J. O’Brien. “The influence of HLA genotype on AIDS”. In: *Annual Review of Medicine-Selected Topics in the Clinical Sciences* 54 (2003), pp. 535–551.

[32] K. Castro et al. “1993 Revised Classification-System for HIV-Infection and Expanded Surveillance Case Definition for AIDS among Adolescents and Adults”. In: *Laboratory Medicine* 24.5 (1993), pp. 286–294.

[33] M. Centlivre et al. “In HIV-1 pathogenesis the die is cast during primary infection”. In: *AIDS* 21.1 (2007), pp. 1–11.

[34] K. Chalmet et al. “Presence of CXCR4-using HIV-1 in patients with recently diagnosed infection: correlates and evidence for transmission”. In: *Journal of Infectious Diseases* 205.2 (2012), pp. 174–84.

[35] L. Chavez-Galan et al. “Cell Death Mechanisms Induced by Cytotoxic Lymphocytes”. In: *Cellular and Molecular Immunology* 6.1 (2009), pp. 15–25.

[36] Z. W. Chen et al. “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”. In: *Journal of Virology* 71.5 (1997), pp. 3953–3960.

[37] B. Chohan et al. “Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1-V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels”. In: *Journal of Virology* 79.10 (2005), pp. 6528–31.

[38] S. J. Clark et al. “High Titers of Cytopathic Virus in Plasma of Patients with Symptomatic Primary HIV-1 Infection”. In: *New England Journal of Medicine* 324.14 (1991), pp. 954–960.

[39] E. S.; Clopper C.; Pearson. "The use of confidence or fiducial limits illustrated in the case of the binomial". In: *Biometrika 26: 404413. doi:10.1093/biomet/26.4.404.* 26.4 (1934), pp. 404–413.

[40] M. S. Cohen, M. McCauley, and T. R. Gamble. "HIV treatment as prevention and HPTN 052". In: *Current opinion in HIV and AIDS* 7.2 (2012). Cohen, Myron S McCauley, Marybeth Gamble, Theresa R Curr Opin HIV AIDS. 2012 Mar;7(2):99-105., pp. 99–105.

[41] S. R. Cole et al. "Copy-Years Viremia as a Measure of Cumulative Human Immunodeficiency Virus Viral Burden". In: *American Journal of Epidemiology* 171.2 (2010), pp. 198–205.

[42] S. Conroy et al. "Significant Decrease in Subtype D Prevalence in Rakai, Uganda between 1994 and 2002". In: *CROI.* 2009.

[43] D. Coombs et al. "Optimal viral production". In: *Bulletin of Mathematical Biology* 65.6 (2003), pp. 1003–1023.

[44] N. Crum-Cianflone et al. "Is HIV Becoming More Virulent? Initial CD4 Cell Counts among HIV Seroconverters during the Course of the HIV Epidemic: 1985-2007". In: *Clinical Infectious Diseases* 48.9 (2009), pp. 1285–1292.

[45] K. Csillery et al. "Approximate Bayesian Computation (ABC) in practice". In: *Trends in ecology and evolution* 25.7 (2010), pp. 410–8.

[46] E. S. Daar et al. "Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection". In: *N Engl J Med* 324.14 (1991), pp. 961–4.

[47] C. R. Darwin. "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life". In: (1859).

[48] N. J. Deacon et al. "Genomic Structure Of An Attenuated Quasi-species of HIV-1 From A Blood-transfusion Donor And Recipients". In: *Science* 270.5238 (1995), pp. 988–991.

[49] C. A. Derdeyn et al. "Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission". In: *Science* 303.5666 (2004), pp. 2019–2022.

[50] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. "On the Definition and the Computation of the Basic Reproduction Ratio R0 in Models for Infectious-Diseases in Heterogeneous Populations". In: *Journal of Mathematical Biology* 28.4 (1990), pp. 365–382.

[51] C. F. Djoko et al. "Failure to Detect Simian Immunodeficiency Virus Infection in a Large Cameroonian Cohort with High Non-human Primate Exposure". In: *Ecohealth* (2012).

[52] D. Donnell et al. "Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: a prospective cohort analysis". In: *Lancet* 375.9731 (2010), pp. 2092–2098.

[53] C. A. Donnelly et al. “Gender difference in HIV-1 RNA viral loads”. In: *HIV Medicine* 6.3 (2005), pp. 170–178.

[54] M. Dorrucci et al. “Changes over time in post-seroconversion CD4 cell counts in the Italian HIV-seroconversion study: 1985-2002”. In: *AIDS* 19.3 (2005), pp. 331–335.

[55] M. Dorrucci et al. “Temporal trends in postseroconversion CD4 cell count and HIV load: The Concerted Action on Seroconversion to AIDS and Death in Europe Collaboration, 1985-2002”. In: *Journal of Infectious Diseases* 195.4 (2007), pp. 525–534.

[56] S. Dray and A. B. Dufour. “The ade4 package: Implementing the duality diagram for ecologists”. In: *Journal of Statistical Software* 22.4 (2007), pp. 1–20.

[57] C. M. Durand et al. “HIV-1 Gag evolution in recently infected human leukocyte antigen-B*57 patients with low-level viremia”. In: *AIDS* 24.15 (2010), pp. 2405–8.

[58] P. J. Easterbrook et al. “Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy”. In: *Journal of the International AIDS Society* 13 (2010).

[59] J. W. Eaton, T. B. Hallett, and G. P. Garnett. “Concurrent Sexual Partnerships and Primary HIV Infection: A Critical Interaction”. In: *AIDS and Behavior* 15.4 (2011), pp. 687–692.

[60] D. Ebert and J. J. Bull. “Challenging the trade-off model for the evolution of virulence: is virulence management feasible?” In: *Trends in Microbiology* 11.1 (2003), pp. 15–20.

[61] B. H. Edwards et al. “Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma”. In: *Journal of Virology* 76.5 (2002), pp. 2298–305.

[62] M. Eigen. “Selforganization of Matter and Evolution of Biological Macromolecules”. In: *Naturwissenschaften* 58.10 (1971), p. 465.

[63] J. Fellay et al. “A whole-genome association study of major determinants for host control of HIV-1”. In: *Science* 317.5840 (2007), pp. 944–947.

[64] J. Fellay et al. “Common Genetic Variation and the Control of HIV-1 in Humans”. In: *Plos Genetics* 5.12 (2009), pp. –.

[65] J. Felsenstein. “Phylogenies and Quantitative Characters”. In: *Annual Review of Ecology and Systematics* 19 (1988), pp. 445–471.

[66] J. Felsenstein. “Phylogenies and the Comparative Method”. In: *American Naturalist* 125.1 (1985), pp. 1–15.

[67] K. A. Fenton et al. “Measuring sexual behaviour: methodological challenges in survey research”. In: *Sexually Transmitted Infections* 77.2 (2001), pp. 84–92.

[68] M. Fevrier, K. Dorgham, and A. Rebollo. "CD4(+) T Cell Depletion in Human Immunodeficiency Virus (HIV) Infection: Role of Apoptosis". In: *Viruses-Basel* 3.5 (2011), pp. 586–612.

[69] U. S. Fideli et al. "Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in Africa". In: *AIDS Res Hum Retroviruses* 17.10 (2001), pp. 901–10.

[70] RA Fisher. "The correlation between relatives on the supposition of Mendelian inheritance". In: *Transactions of the Royal Society of Edinburgh* 52 (1918), pp. 399–433.

[71] C. Fraser and T. D. Hollingsworth. "Interpretation of correlations in setpoint viral load in transmitting couples". In: *AIDS* 24.16 (2010), pp. 2596–2597.

[72] C. Fraser et al. "The role of antigenic stimulation and cytotoxic T cell activity in regulating the long-term immunopathogenesis of HIV: mechanisms and clinical implications". In: *Proceedings of the Royal Society of London Series B-Biological Sciences* 268.1481 (2001), pp. 2085–2095.

[73] C. Fraser et al. "Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (2007), pp. 17441–17446.

[74] R. P. Freckleton. "The seven deadly sins of comparative analysis". In: *Journal of evolutionary biology* 22.7 (2009). Freckleton, R P Switzerland J Evol Biol. 2009 Jul;22(7):1367-75. Epub 2009 Jun 5., pp. 1367–75.

[75] R. P. Freckleton, P. H. Harvey, and M. Pagel. "Phylogenetic analysis and comparative data: A test and review of evidence". In: *American Naturalist* 160.6 (2002), pp. 712–726.

[76] Y. Gali et al. "Survey of the temporal changes in HIV-1 replicative fitness in the Amsterdam Cohort". In: *Virology* 364.1 (2007), pp. 140–146.

[77] A. P. Galvani. "Epidemiology meets evolutionary ecology". In: *Trends in Ecology and Evolution* 18.3 (2003), pp. 132–139.

[78] S. Gandon et al. "Imperfect vaccines and the evolution of pathogen virulence". In: *Nature* 414.6865 (2001), pp. 751–756.

[79] X. J. Gao et al. "HLA-B alleles associate consistently with HIV heterosexual transmission, viral load, and progression to AIDS, but not susceptibility to infection". In: *AIDS* 24.12 (2010), pp. 1835–1840.

[80] G. P. Garnett and R. M. Anderson. "Sexually transmitted diseases and sexual behavior: Insights from mathematical models". In: *Journal of Infectious Diseases* 174 (1996), S150–S161.

[81] A. M. Geretti. "HIV-1 subtypes: epidemiology and significance for HIV management". In: *Curr Opin Infect Dis* 19.1 (2006), pp. 1–7.

[82] R. B. Geskus et al. "The HIV RNA setpoint theory revisited". In: *Retrovirology* 4 (2007), p. 65.

[83] M. T. P. Gilbert et al. “The emergence of HIV/AIDS in the Americas and beyond”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.47 (2007), pp. 18566–18570.

[84] M. A. Gilchrist, D. Coombs, and A. S. Perelson. “Optimizing within-host viral fitness: infected cell lifespan and virion production rate”. In: *Journal of Theoretical Biology* 229.2 (2004), pp. 281–288.

[85] M. F. Goodman and D. K. Fygenson. “DNA polymerase fidelity: From genetics toward a biochemical understanding”. In: *Genetics* 148.4 (1998), pp. 1475–1482.

[86] P. R. Gorry et al. “Pathogenicity and immunogenicity of attenuated, nef-deleted HIV-1 strains in vivo”. In: *Retrovirology* 4 (2007).

[87] R. Gramacy. “tgp: An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models”. In: *Journal of Statistical Software* 19.9 (2007), pp. 1–46.

[88] R. B. Gramacy and M. Taddy. “Categorical Inputs, Sensitivity Analysis, Optimization and Importance Tempering with tgp Version 2, an R Package for Treed Gaussian Process Models”. In: *Journal of Statistical Software* 33.6 (2010), pp. 1–48.

[89] R. M. Grant et al. “Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men.” In: *New England Journal of Medicine* 363.27 (2010), pp. 2587–2599.

[90] Luuk Gras et al. “Viral load levels measured at set-point have risen over the last decade of the HIV epidemic in the Netherlands”. In: *PLoS One* 4.10 (2009), e7365.

[91] R. H. Gray et al. “Male circumcision and HIV acquisition and transmission: cohort studies in Rakai, Uganda”. In: *AIDS* 14.15 (2000), pp. 2371–2381.

[92] S. Guindon and O. Gascuel. “A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood”. In: *Systematic biology* 52.5 (2003). Guindon, Stephane Gascuel, Olivier England Syst Biol. 2003 Oct;52(5):696-704., pp. 696–704.

[93] R. E. Haaland et al. “Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1”. In: *PLoS Pathog* 5.1 (2009), e1000274.

[94] F. M. Hecht et al. “HIV RNA level in early infection is predicted by viral load in the transmission source”. In: *AIDS* 24.7 (2010), pp. 941–945.

[95] D. R. Henrard et al. “Natural History Of HIV-1 Cell-free Viremia”. In: *Jama-Journal of the American Medical Association* 274.7 (1995), pp. 554–558.

[96] J. T. Herbeck et al. “Is the virulence of HIV changing? A meta-analysis of trends in prognostic markers of HIV disease progression and transmission”. In: *AIDS* 26.2 (2012), pp. 193–205.

[97] F. van Heuverswyn and M. Peeters. "The origins of HIV and implications for the global epidemic". In: *Curr Infect Dis Rep* 9.4 (2007), pp. 338–46.

[98] T. Hinkley et al. "A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase". In: *Nat Genet* 43.5 (2011), pp. 487–9.

[99] M. S. Hirsch et al. "Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel". In: *Clinical Infectious Diseases* 47.2 (2008), pp. 266–285.

[100] D. D. Ho et al. "Rapid Turnover of Plasma Virions and Cd4 Lymphocytes in HIV-1 Infection". In: *Nature* 373.6510 (1995), pp. 123–126.

[101] E. Hodcroft et al. *Population-based Study Indicates Viral Genetic Effect on HIV Virulence Is Small but Significant*. 2012.

[102] R. Hogg et al. "Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies". In: *Lancet* 372.9635 (2008), pp. 293–299.

[103] T. D. Hollingsworth, R. M. Anderson, and C. Fraser. "HIV-1 transmission, by stage of infection". In: *Journal of Infectious Diseases* 198.5 (2008), pp. 687–693.

[104] T. D. Hollingsworth et al. "HIV-1 Transmitting Couples Have Similar Viral Load Set-Points in Rakai, Uganda". In: *Plos Pathogens* 6.5 (2010).

[105] S. E. Holte et al. "Efficient use of longitudinal CD4 counts and viral load measures in survival analysis". In: *Statistics in Medicine* (2012).

[106] Family Health International. *FHI Statement on the FEM-PrEP HIV Prevention Study*. Tech. rep. 2011.

[107] J. P. Ji and L. A. Loeb. "Fidelity of HIV-1 Reverse-Transcriptase Copying a Hypervariable Region of the HIV-1 Env Gene". In: *Virology* 199.2 (1994), pp. 323–330.

[108] B. Jia et al. "Species-Specific Activity of SIV Nef and HIV-1 Vpu in Overcoming Restriction by Tetherin/BST2". In: *Plos Pathogens* 5.5 (2009).

[109] T. Jombart and S. Dray. "adephylo: exploratory analyses for the phylogenetic comparative method." In: *Bioinformatics* 26 (2010), pp. 1907–1909. DOI: 10.1093/bioinformatics/btq292.

[110] K. E. Jones et al. "Global trends in emerging infectious diseases". In: *Nature* 451.7181 (2008), 990–U4.

[111] L. E. Jones and A. S. Perelson. "Transient viremia, plasma viral load, and reservoir replenishment in HIV-Infected patients on antiretroviral therapy". In: *Jaids-Journal of Acquired Immune Deficiency Syndromes* 45.5 (2007), pp. 483–493.

[112] M. A. de Jong and T. B. Geijtenbeek. "Human immunodeficiency virus-1 acquisition in genital mucosa: Langerhans cells as key-players". In: *Journal of Internal Medicine* 265.1 (2009), pp. 18–28.

[113] A. Jung et al. “Recombination - Multiply infected spleen cells in HIV patients”. In: *Nature* 418.6894 (2002), pp. 144–144.

[114] P. J. Kanki et al. “Human immunodeficiency virus type 1 subtypes differ in disease progression”. In: *12th World AIDS Conference*. Vol. 179. Univ Chicago Press, 1998, pp. 68–73.

[115] Q. A. Karim et al. “Effectiveness and Safety of Tenofovir Gel, an Antiretroviral Microbicide, for the Prevention of HIV Infection in Women”. In: *Science* 329.5996 (2010), pp. 1168–1174.

[116] R. A. Katz and A. M. Skalka. “Generation Of Diversity In Retroviruses”. In: *Annual Review of Genetics* 24 (1990), pp. 409–445.

[117] P. Kellam and B. A. Larder. “Retroviral Recombination Can Lead to Linkage of Reverse-Transcriptase Mutations That Confer Increased Zidovudine Resistance”. In: *Journal of Virology* 69.2 (1995), pp. 669–674.

[118] S. W. Kembel et al. “Picante: R tools for integrating phylogenies and ecology”. In: *Bioinformatics* 26.11 (2010), pp. 1463–1464.

[119] A. S. Khan et al. “The reemergence of Ebola hemorrhagic fever, Democratic Republic of the Congo, 1995. Commission de Lutte contre les Epidemies a Kikwit”. In: *Journal of Infectious Diseases* 179 Suppl 1 (1999), S76–86.

[120] N. Kiwanuka et al. “Effect of human immunodeficiency virus type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection”. In: *Journal of Infectious Diseases* 197.5 (2008), pp. 707–713.

[121] N. Kiwanuka et al. “HIV-1 subtypes and differences in heterosexual HIV transmission among HIV-discordant couples in Rakai, Uganda”. In: *AIDS* 23.18 (2009), pp. 2479–84.

[122] N. R. Klatt, G. Silvestri, and V. Hirsch. “Nonpathogenic simian immunodeficiency virus infections”. In: *Cold Spring Harb Perspect Med* 2.1 (2012), a007153.

[123] X. H. Kong et al. “The Human Immunodeficiency Virus Type 1 Envelope Confers Higher Rates of Replicative Fitness to Perinatally Transmitted Viruses than to Nontransmitted Viruses”. In: *Journal of Virology* 82.23 (2008), pp. 11609–11618.

[124] E. L. Korenromp et al. “Clinical prognostic value of RNA viral load and CD4 cell counts during untreated HIV-1 infection–a quantitative review”. In: *PLoS One* 4.6 (2009), e5950.

[125] R. A. Koup et al. “Temporal Association of Cellular Immune-Responses with the Initial Control of Viremia in Primary Human-Immunodeficiency-Virus Type-1 Syndrome”. In: *Journal of Virology* 68.7 (1994), pp. 4650–4655.

[126] R. D. Kouyos et al. “Assessing predicted HIV-1 replicative capacity in a clinical setting”. In: *PLoS Pathog* 7.11 (2011), e1002321.

[127] C. S. Kraft et al. “Timing and source of subtype-C HIV-1 superinfection in the newly infected partner of Zambian couples with disparate viruses”. In: *Retrovirology* 9.1 (2012), p. 22.

[128] A. Kulkarni, R. Paranjape, and M. Thakar. “Expansion of defective NK cells in early HIV type 1C infection: a consequence of reduced CD161 expression”. In: *AIDS Res Hum Retroviruses* 28.1 (2012), pp. 100–5.

[129] A. C. van der Kuyl and M. Cornelissen. “Identifying HIV-1 dual infections”. In: *Retrovirology* 4 (2007).

[130] A. C. van der Kuyl et al. “HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients”. In: *AIDS* 24.10 (2010), pp. 1607–1608.

[131] L. Lamport. *LaTeX: a Document Preparation System*. 2nd ed. Illustrations by Duane Bibby. Reading, Ma.: Addison-Wesley Publishing Company, 1994.

[132] S. E. Langford, J. Ananworanich, and D. A. Cooper. “Predictors of disease progression in HIV infection: a review”. In: *AIDS Res Ther* 4 (2007), p. 11.

[133] A. Le Tortorec and S. J. D. Neil. “Antagonism to and Intracellular Sequestration of Human Tetherin by the Human Immunodeficiency Virus Type 2 Envelope Glycoprotein”. In: *Journal of Virology* 83.22 (2009), pp. 11966–11978.

[134] B. Ledergerber et al. “The Swiss HIV Cohort Study - Rationale, Organization and Selected Base-Line Characteristics”. In: *Sozial-Und Praventivmedizin* 39.6 (1994), pp. 387–394.

[135] P. Lemey, A. Rambaut, and O. G. Pybus. “HIV evolutionary dynamics within and among hosts”. In: *AIDS Reviews* 8.3 (2006), pp. 125–140.

[136] A. J. Leslie et al. “HIV evolution: CTL escape mutation and reversion after transmission”. In: *Nature Medicine* 10.3 (2004), pp. 282–289.

[137] G. E. Leventhal et al. “Inferring epidemic contact structure from phylogenetic trees”. In: *Plos Computational Biology* 8.3 (2012), e1002413.

[138] B. R. Levin and J. J. Bull. “Short-sighted evolution and the virulence of pathogenic microorganisms”. In: *Trends in Microbiology* 2.3 (1994), pp. 76–81.

[139] D. N. Levy et al. “Dynamics of HIV-1 recombination in its natural target cells”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.12 (2004), pp. 4204–4209.

[140] J. D. Lifson et al. “The extent of early viral replication is a critical determinant of the natural history of simian immunodeficiency virus infection”. In: *Journal of Virology* 71.12 (1997), pp. 9508–14.

[141] J. R. Lingappa et al. “Estimating the Impact of Plasma HIV-1 RNA Reductions on Heterosexual HIV-1 Transmission Risk”. In: *PLoS One* 5.9 (2010), pp. –.

[142] J. R. Lingappa et al. “Genomewide association study for determinants of HIV-1 acquisition and viral set point in HIV-1 serodiscordant couples with quantified virus exposure”. In: *PLoS One* 6.12 (2011), e28632.

[143] K. H. Luo et al. “The Role of Initial AIDS-Defining Illness in Survival Following AIDS”. In: *AIDS* 9.1 (1995), pp. 57–63.

[144] C. Lythgoe K.A.; Fraser. “New Insights into the Evolutionary Rate of HIV-1 at the Within-Host and Epidemiological Levels”. In: *Proc. Roy. Soc. B.* (2012).

[145] Z. M. Ma et al. “High specific infectivity of plasma virus from the pre-ramp-up and ramp-up stages of acute simian immunodeficiency virus infection”. In: *Journal of Virology* 83.7 (2009), pp. 3288–97.

[146] D. van Manen et al. “Rising HIV-1 viral load set-point at a population level coincides with a fading impact of host genetic factors on HIV-1 control”. In: *AIDS* (2011).

[147] N. Mantel. “The detection of disease clustering and a generalized regression approach”. In: *Cancer Res* 27.2 (1967), pp. 209–20.

[148] D. M. Margolis. “Eradication Therapies for HIV Infection: Time to Begin Again”. In: *AIDS Research and Human Retroviruses* 27.4 (2011), pp. 347–353.

[149] J. Masel et al. “Fluctuations in HIV-1 viral load are correlated to CD4(+) T-lymphocyte count during the natural course of infection”. In: *Journal of Acquired Immune Deficiency Syndromes* 23.5 (2000), pp. 375–379.

[150] J. J. Mattapallil et al. “Massive infection and loss of memory CD4(+) T cells in multiple tissues during acute SIV infection”. In: *Nature* 434.7037 (2005), pp. 1093–1097.

[151] R. M. May and R. M. Anderson. “Epidemiology and Genetics in the Coevolution of Parasites and Hosts”. In: *Proceedings of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences* 390.1798 (1983), pp. 219–219.

[152] J. W. Mellors et al. “Prognosis in HIV-1 infection predicted by the quantity of virus in plasma”. In: *Science* 272.5265 (1996), pp. 1167–1170.

[153] E. Mills et al. “Male circumcision for the prevention of heterosexually acquired HIV infection: a meta-analysis of randomized trials involving 11,050 men”. In: *HIV Medicine* 9.6 (2008), pp. 332–335.

[154] E. J. Mills et al. “Life Expectancy of Persons Receiving Combination Antiretroviral Therapy in Low-Income Countries: A Cohort Analysis From Uganda”. In: *Annals of Internal Medicine* 155.4 (2011), 209–U41.

[155] F. Mittelbach et al. *The LaTeX Companion*. Addison-Wesley Series on Tools and Techniques for Computer Typesetting. Boston: Addison-Wesley Professional, 2004.

[156] T. Miura et al. "HLA-B57/B*5801 Human Immunodeficiency Virus Type 1 Elite Controllers Select for Rare Gag Variants Associated with Reduced Viral Replication Capacity and Strong Cytotoxic T-Lymphocyte Recognition (vol 83, pg 2743, 2009)". In: *Journal of Virology* 83.11 (2009), pp. 5961–5961.

[157] M. Morris and M. Kretzschmar. "Concurrent partnerships and the spread of HIV". In: *AIDS* 11.5 (1997), pp. 641–648.

[158] V. Müller, C. Fraser, and J. T. Herbeck. "A Strong Case for Viral Genetic Factors in HIV Virulence". In: *Viruses-Basel* 3.3 (2011), pp. 204–216.

[159] V. Müller, A. F. Maree, and R. J. De Boer. "Small variations in multiple parameters account for wide variations in HIV-1 set-points: a novel modelling approach". In: *Proc Biol Sci* 268.1464 (2001), pp. 235–42.

[160] V. Müller et al. "Increasing clinical virulence in two decades of the Italian HIV epidemic". In: *Plos Pathogens* 5.5 (2009), e1000454.

[161] V. Müller et al. "Stable virulence levels in the HIV epidemic of Switzerland over two decades". In: *AIDS* 20.6 (2006), pp. 889–894.

[162] A. Munoz et al. "Long-Term Survivors with HIV-1 Infection - Incubation Period and Longitudinal Patterns of Cd4(+) Lymphocytes". In: *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 8.5 (1995), pp. 496–505.

[163] O. T. Ng et al. "Increased rate of CD4+ T-cell decline and faster time to antiretroviral therapy in HIV-1 subtype CRF01_AE infected seroconverters in Singapore". In: *PLoS One* 6.1 (2011), e15738.

[164] V. Novitsky et al. "Extended high viremics: a substantial fraction of individuals maintain high plasma viral RNA levels after acute HIV-1 subtype C infection". In: *AIDS* 25.12 (2011), pp. 1515–22.

[165] M. A. Nowak et al. "Antigenic Diversity Thresholds and the Development of AIDS". In: *Science* 254.5034 (1991), pp. 963–969.

[166] T. van Opijnen et al. "Adaptation of HIV-1 Depends on the Host-Cell Environment". In: *PLoS One* 2.3 (2007).

[167] H. A. Orr. "Fitness and its role in evolutionary genetics". In: *Nature Reviews Genetics* 10.8 (2009), pp. 531–539.

[168] N. S. Padian et al. "HIV prevention transformed: the new prevention research agenda". In: *Lancet* 378.9787 (2011), pp. 269–278.

[169] M. Pagel. "Inferring the historical patterns of biological evolution". In: *Nature* 401.6756 (1999), pp. 877–84.

[170] M. Paiardini et al. "Lessons Learned from the Natural Hosts of HIV-Related Viruses". In: *Annual Review of Medicine* 60 (2009), pp. 485–495.

[171] I. Pandrea et al. "Paucity of CD4(+)CCR5(+) T cells is a typical feature of natural SIV hosts". In: *Blood* 109.3 (2007), pp. 1069–1076.

[172] G. Pantaleo et al. “The qualitative nature of the primary immune response to HIV infection is a prognosticator of disease progression independent of the initial level of plasma viremia”. In: *Proc Natl Acad Sci U S A* 94.1 (1997), pp. 254–8.

[173] E. Paradis, J. Claude, and K. Strimmer. “APE: Analyses of Phylogenetics and Evolution in R language”. In: *Bioinformatics* 20.2 (2004), pp. 289–290.

[174] D. Paraskevis et al. “Increasing prevalence of HIV-1 subtype a in Greece: Estimating epidemic history and origin”. In: *Journal of Infectious Diseases* 196.8 (2007), pp. 1167–1176.

[175] D. Paraskevis et al. “Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach”. In: *Retrovirology* 6 (2009), pp. –.

[176] M. Parera et al. “HIV-1 protease catalytic efficiency effects caused by random single amino acid substitutions”. In: *Molecular Biology and Evolution* 24.2 (2007), pp. 382–387.

[177] O. Patashnik. “Using BibTeX”. Documentation for general BibTeX users. 1988.

[178] S. Pavoine et al. “Testing for phylogenetic signal in phenotypic traits: New matrices of phylogenetic proximities”. In: *Theoretical Population Biology* 73.1 (2008), pp. 79–91.

[179] K. M. Pepin et al. “Identifying genetic markers of adaptation for surveillance of viral host jumps”. In: *Nat Rev Microbiol* 8.11 (2010), pp. 802–13.

[180] S. Philpott et al. “Human immunodeficiency virus type 1 genomic RNA sequences in the female genital tract and blood: Compartmentalization and intrapatient recombination”. In: *Journal of Virology* 79.1 (2005), pp. 353–363.

[181] M. Piatak et al. “High-Levels of Plasma Virus during All Stages of HIV-1 Infection, Including Asymptomatic Infection - Measurement by Quantitative Competitive Pcr (Qc-Pcr)”. In: *Journal of Cellular Biochemistry* (1993), pp. 21–21.

[182] M. Pingen et al. “Evolutionary pathways of transmitted drug-resistant HIV-1”. In: *Journal of Antimicrobial Chemotherapy* 66.7 (2011), pp. 1467–1480.

[183] D. Posada and K. A. Crandall. “MODELTEST: testing the model of DNA substitution”. In: *Bioinformatics* 14.9 (1998), pp. 817–818.

[184] K. A. Powers et al. “The role of acute and early HIV infection in the spread of HIV and implications for transmission prevention strategies in Lilongwe, Malawi: a modelling study”. In: *Lancet* 378.9787 (2011), pp. 256–268.

[185] A. Prendergast et al. “HIV-1 infection is characterized by profound depletion of CD161(+) Th17 cells and gradual decline in regulatory T cells”. In: *AIDS* 24.4 (2010), pp. 491–502.

[186] O. G. Pybus and A. Rambaut. “Evolutionary analysis of the dynamics of viral infectious disease”. In: *Nature Reviews Genetics* 10.8 (2009), pp. 540–550.

[187] T. C. Quinn et al. “Viral load and heterosexual transmission of human immunodeficiency virus type 1”. In: *New England Journal of Medicine* 342.13 (2000), pp. 921–929.

[188] S. Rainwater et al. “No evidence for rapid subtype C spread within an epidemic in which multiple subtypes and intersubtype recombinants circulate”. In: *AIDS Res Hum Retroviruses* 21.12 (2005), pp. 1060–5.

[189] A. Rambaut. “Path-O-Gen: temporal signal investigation tool.” In: (2010).

[190] J. M. Read and M. J. Keeling. “Disease evolution across a range of spatio-temporal scales”. In: *Theoretical Population Biology* 70.2 (2006), pp. 201–213.

[191] R. R. Regoes et al. “Roles of target cells and virus-specific cellular immunity in primary simian immunodeficiency virus infection”. In: *Journal of Virology* 78.9 (2004), pp. 4866–4875.

[192] S. Rerks-Ngarm et al. “Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand”. In: *New England Journal of Medicine* 361.23 (2009), pp. 2209–2220.

[193] L. J. Revell. *phylosig*. 2011.

[194] S. J. Reynolds et al. “HIV-1 transmission among HIV-1 discordant couples before and after the introduction of antiretroviral therapy”. In: *AIDS* 25.4 (2011), pp. 473–477.

[195] S. Y. Rhee et al. “Human immunodeficiency virus reverse transcriptase and protease sequence database”. In: *Nucleic Acids Research* 31.1 (2003), pp. 298–303.

[196] B. A. Richardson et al. “Comparison of human immunodeficiency virus type 1 viral loads in Kenyan women, men, and infants during primary and early infection”. In: *Journal of Virology* 77.12 (2003), pp. 7120–7123.

[197] D. Richman et al. “Human-Immunodeficiency-Virus Type-1 Mutants Resistant to Nonnucleoside Inhibitors of Reverse-Transcriptase Arise in Tissue-Culture”. In: *Proceedings of the National Academy of Sciences of the United States of America* 88.24 (1991), pp. 11241–11245.

[198] C. M. Rousseau et al. “HLA class I-driven evolution of human immunodeficiency virus type 1 subtype C proteome: Immune escape and viral load”. In: *Journal of Virology* 82.13 (2008), pp. 6434–6446.

[199] E. Saathoff et al. “Viral and Host Factors Associated With the HIV-1 Viral Load Setpoint in Adults From Mbeya Region, Tanzania”. In: *Jaids-Journal of Acquired Immune Deficiency Syndromes* 54.3 (2010), pp. 324–331.

[200] M. Sagar et al. “Selection of HIV Variants with Signature Genotypic Characteristics during Heterosexual Transmission”. In: *Journal of Infectious Diseases* 199.4 (2009), pp. 580–589.

[201] J. F. Salazar-Gonzalez et al. “Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection”. In: *Journal of Experimental Medicine* 206.6 (2009), pp. 1273–1289.

[202] M. L. Santiago et al. “SIVcpz in wild chimpanzees”. In: *Science* 295.5554 (2002), pp. 465–465.

[203] D. Sauter et al. “Tetherin-Driven Adaptation of Vpu and Nef Function and the Evolution of Pandemic and Nonpandemic HIV-1 Strains”. In: *Cell Host and Microbe* 6.5 (2009), pp. 409–421.

[204] F. Schoeni-Affolter et al. “Cohort Profile: The Swiss HIV Cohort Study”. In: *International Journal of Epidemiology* 39.5 (2010), pp. 1179–1189.

[205] R. Schuurman et al. “Rapid Changes in Human-Immunodeficiency-Virus Type-1 Rna Load and Appearance of Drug-Resistant Virus Populations in Persons Treated with Lamivudine (3tc)”. In: *Journal of Infectious Diseases* 171.6 (1995), pp. 1411–1419.

[206] R. W. Shafer. “Rationale and uses of a public HIV drug-resistance database”. In: *Journal of Infectious Diseases* 194 Suppl 1 (2006), S51–8.

[207] R. Shankarappa et al. “Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection”. In: *Journal of Virology* 73.12 (1999), pp. 10489–10502.

[208] R. L. Shapiro et al. “Antiretroviral Regimens in Pregnancy and Breast-Feeding in Botswana”. In: *New England Journal of Medicine* 362.24 (2010), pp. 2282–2294.

[209] P. M. Sharp and B. H. Hahn. “Origins of HIV and the AIDS Pandemic”. In: *Cold Spring Harb Perspect Med* 1.1 (2011), a006841.

[210] G. Shirreff et al. “Transmission Selects for HIV-1 Strains of Intermediate Virulence: A Modelling Approach”. In: *Plos Computational Biology* 7.10 (2011).

[211] G. Silvestri et al. “Divergent host responses during primary simian immunodeficiency virus SIVsm infection of natural sooty mangabey and nonnatural rhesus macaque hosts”. In: *Journal of Virology* 79.7 (2005), pp. 4043–54.

[212] B. Soubeyrand and S. A. Plotkin. “Microbial evolution - Antitoxin vaccines and pathogen virulence”. In: *Nature* 417.6889 (2002), pp. 609–610.

[213] A. Stamatakis. "RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models". In: *Bioinformatics* 22.21 (2006), pp. 2688–2690.

[214] H. Streeck et al. "Human Immunodeficiency Virus Type 1-Specific CD8(+) T-Cell Responses during Primary Infection Are Major Determinants of the Viral Set Point and Loss of CD4(+) T Cells". In: *Journal of Virology* 83.15 (2009), pp. 7641–7648.

[215] J. M. Tang et al. "HLA allele sharing and HIV type 1 viremia in seroconverting Zambians with known transmitting partners". In: *AIDS Research and Human Retroviruses* 20.1 (2004), pp. 19–25.

[216] S. Tavaré. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences". In: *American Mathematical Society: Lectures on Mathematics in the Life Sciences*. Vol. 17. Amer Mathematical Society, 1986, pp. 57–86. ISBN: 0821811673. URL: `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/0821811673`.

[217] R Development Core Team. *R: A language and environment for statistical computing*. 2011.

[218] D. M. Tebit and E. J. Arts. "Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease". In: *Lancet Infectious Diseases* 11.1 (2011), pp. 45–56.

[219] R. M. Troyer et al. "Changes in human immunodeficiency virus type 1 fitness and genetic diversity during disease progression". In: *Journal of Virology* 79.14 (2005), pp. 9006–9018.

[220] L. S. Uhlenbeck G.E.; Ornstein. "On the Theory of the Brownian Motion". In: *Phys. Rev.* 36 (1930), 823841.

[221] UNAIDS. *World AIDS Day Report*. Tech. rep. 2011.

[222] L. V. Wain et al. "Adaptation of HIV-1 to its human host". In: *Molecular Biology and Evolution* 24.8 (2007), pp. 1853–1860.

[223] J. W. Ward et al. "The Natural-History of Transfusion-Associated Infection with Human Immunodeficiency Virus - Factors Influencing the Rat of Progression to Disease". In: *New England Journal of Medicine* 321.14 (1989), pp. 947–952.

[224] M. J. Wawer et al. "A randomized, community trial of intensive sexually transmitted disease control for AIDS prevention, Rakai, Uganda". In: *AIDS* 12.10 (1998), pp. 1211–1225.

[225] M. J. Wawer et al. "Control of sexually transmitted diseases for AIDS prevention in Uganda: a randomised community trial". In: *Lancet* 353.9152 (1999), pp. 525–535.

[226] M. J. Wawer et al. "Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda". In: *Journal of Infectious Diseases* 191.9 (2005), pp. 1403–1409.

[227] X. Wei et al. “Viral dynamics in human immunodeficiency virus type 1 infection”. In: *Nature* 373.6510 (1995), pp. 117–22.

[228] P. J. White et al. “Epidemiology of rabbit haemorrhagic disease virus in the United Kingdom: evidence for seasonal transmission by both virulent and avirulent modes of infection”. In: *Epidemiology and Infection* 132.3 (2004), pp. 555–567.

[229] D. Wodarz and M. A. Nowak. “Specific therapy regimes could lead to long-term immunological control of HIV”. In: *Proc Natl Acad Sci U S A* 96.25 (1999), pp. 14464–9.

[230] F. de Wolf et al. “AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: Markers with reciprocal predictive value over time after seroconversion”. In: *AIDS* 11.15 (1997), pp. 1799–1806.

[231] N. D. Wolfe et al. “Naturally acquired simian retrovirus infections in central African hunters”. In: *Lancet* 363.9413 (2004), pp. 932–937.

[232] T. F. W. Wolfs et al. “Naturally-Occurring Mutations within HIV-1 V3 Genomic Rna Lead to Antigenic Variation Dependent on a Single Amino-Acid Substitution”. In: *Virology* 185.1 (1991), pp. 195–205.

[233] M. Worobey et al. “Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960”. In: *Nature* 455.7213 (2008), 661–U57.

[234] O. O. Yang et al. “Suppression of human immunodeficiency virus type 1 replication by CD8(+) cells: Evidence for HLA class I-restricted triggering of cytolytic and noncytolytic mechanisms”. In: *Journal of Virology* 71.4 (1997), pp. 3120–3128.

[235] Z. H. Yang. “Among-site rate variation and its impact on phylogenetic analyses”. In: *Trends in Ecology and Evolution* 11.9 (1996), pp. 367–372.

[236] S. Yerly et al. “Transmission of antiretroviral-drug-resistant HIV-1 variants”. In: *Lancet* 354.9180 (1999), pp. 729–733.

[237] T. F. Zhu et al. “Genotypic and Phenotypic Characterization of HIV-1 In Patients With Primary Infection”. In: *Science* 261.5125 (1993), pp. 1179–1181.