Viral fitness, genomic recombination, and immune evasion: Mechanisms of HIV evolution across scales

Harriet Longley

Keble College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

January 2025

Abstract

The within-host evolution of HIV is characterised by rapid selection of immune-evasion mutations, pervasive recombination, and vast fitness landscapes, making it one of the fastest evolving organisms. The enormous evolutionary capacity of the virus has challenged therapeutic development, with a cure remaining elusive and drug resistance continues to be a concern. In contrast, between-host evolution shows evidence of neutral evolution, with multiple subtypes coexisting globally. Selection pressures and fitness trade-offs can vary or conflict across within and between-host scales, and the impact of within-host selection on the virus circulating at a population level is not fully understood. In this thesis, I investigate the evolutionary forces driving viral dynamics, focusing on how selection and recombination shape viral populations within and between hosts. Using whole-genome deep sequencing data from hundreds of longitudinally sampled transmission pairs in sub-Saharan Africa, I explore several aspects of viral evolution.

First, I examine virulence evolution as a classic example of the conflict between within-host and between-host selection. Through modelling viral load as determined by the number of weakly deleterious mutations across many segregating sites, I demonstrate how between-host selection for transmission fitness can overcome short-term evolutionary pressures within-host as a result of the balance between mutation and selection pressure. I then investigate recombination, a major source of viral genetic diversity, discussing challenges in accurately inferring recombination rates in dynamic, rapidly evolving viral populations. I identify consistent recombination hot and cold spots across the genome that persist across subtypes and sequencing platforms, corresponding with previously identified inter-subtype patterns. I then quantify the evolutionary rate at both the within and between-host scales across the genome, proposing that transient, high-frequency mutations (“toggling” mutations) explain discrepancies in rate estimation across the two scales. Finally, I show that CTL escape mutations dominate early viral evolution and undergo selection and reversion during transmission, linking these dynamics to within-host toggling at the amino acid level.

This work reveals general principles of viral evolution that persist across hosts, subtypes, and scales. The findings have important implications for understanding the long-term impact of immune pressures, the balance of selection forces, and methodological approaches to studying rapidly evolving viruses.

Table of Contents

Abstract *2*

List of figures *7*

List of tables *8*

Acknowledgements *9*

Declaration *11*

Papers for Publication *12*

1 Introduction *13*

1.1 Motivation and outline **13**

1.2 HIV epidemic background **15**

1.3 HIV Biology **15**

1.3.1 The HIV genome 16

1.3.2 HIV reservoir 17

1.3.3 Stages of infection and pathogenesis 18

1.3.4 Immune response 19

1.3.5 Compartmentalisation 20

1.4 Within-Host evolution of HIV **21**

1.4.1 Sources of variation 21

1.4.2 Viral population dynamics 23

1.4.3 Viral quasispecies framework 23

1.4.4 Selection pressure 25

1.4.5 Fitness landscape 26

1.4.6 The within-host evolutionary rate 27

1.5 Between-host evolution of HIV **28**

1.5.1 The evolutionary rate mismatch 29

1.6 Thesis Outline and Key Findings **30**

1.7 Significance **31**

2 Methods and Data *33*

2.1 Dataset: Partners in Prevention Studies **33**

2.1.1 HLA dataset 35

2.1.2 Analysis datasets 36
2.1.3 Estimated date of seroconversion 37
2.1.4 Set-point viral loads 37
2.1.5 Subtypes 37
2.1.6 Data availability 37

3 Attenuation of HIV severity by slightly deleterious mutations can explain the long-term trajectory of virulence evolution. 39

3.1 Abstract: 39

3.2 Introduction 39

3.3 Methods 42

3.3.1 Within-host dynamics 43
3.3.2 Viral load 44
3.3.3 Duration of chronic infection 45
3.3.4 Infectivity profile 45
3.3.5 Between-host model 47
3.3.6 Host heterogeneity 48
3.3.7 Heritability 48
3.3.8 Viral load analysis in study cohort 49

3.4 Results 49

3.4.1 Many mutations with low fitness cost results in intermediate virulence evolving within individuals 50
3.4.2 Having many mutations with low fitness cost slows the tempo of within-host evolution 51
3.4.3 Many mutations with low fitness cost leads to between-host diversity in viral loads 53
3.4.4 Many mutations with low fitness cost results in viral load evolving to intermediate levels between-hosts 55
3.4.5 Viral loads are similar within transmission pairs 56
3.4.6 Support for model predictions in serodifferent studies dataset 58

3.5 Discussion 60

4 HIV within-host recombination across the genome 64

4.1 Abstract 64

4.2 Introduction 64

4.3 Methods 69

4.3.1 Study cohort 69

4.3.2 Linkage Disequilibrium Calculation 69
4.3.3 Recombination Analysis 70
4.3.4 Viral Load Matching 71
4.3.5 Recombination estimates by sliding windows 72
4.3.6 Estimation of correlation for window-specific rates 72
4.3.7 Simulations 73
4.4 Results 73
4.4.1 Dataset characteristics 73
4.4.2 Measured recombination saturates across long genomic distances and timescales 74
4.4.3 Recombination rate varies by viral load stage of infection 77
4.4.4 Substantial variation in rates of recombination between and within genes 79
4.4.5 Subtype-specific patterns of recombination 82
4.4.6 Recombination patterns supported in Illumina pipeline sequences 84
4.4.7 Recombinants selected for in *pol* but against in *env* 86
4.5 Discussion 87
5 The evolutionary rate of HIV at the within and between host scale 91
5.1 Abstract: 91
5.2 Introduction 91
5.3 Methods 96
5.3.1 Study Population 96
5.3.2 Sequence processing and filtering 97
5.3.3 Evolutionary Analyses 98
5.3.4 BEAST Analysis 100
5.3.5 Between-host dataset and analysis 102
5.4 Results 103
5.4.1 Genome-wide patterns of genetic divergence within and between host 103
5.4.2 Evolutionary rate analysis with BEAST 107
5.4.3 Tempo of evolution differs across methodologies and scales 114
5.4.4 Elevated within-host rates across shorter time scales 118
5.4.5 Extensive within-host toggling for synonymous and non-synonymous mutations 120
5.5 Discussion 123
6 Direct evidence of CTL escape and reversion across transmission Pairs 128
6.1 Abstract 128

6.2 Introduction 128

6.3 Methods 132

6.3.1 Study datasets 132

6.3.2 Subtype-specific consensus 134

6.3.3 Identifying candidate CTL-escape codon positions 134

6.3.4 Estimating the strength of selection 135

6.4 Results 136

6.4.1 High HLA diversity among transmission pairs 136

6.4.2 CTL Escape occurs at all stages of infection with substantial variation in selection strength across escapes 137

6.4.3 Shared Selection Patterns Among HLA-Matched Individuals 140

6.4.4 Evidence of shifting selection pressures between and within hosts 142

6.4.5 Broad distribution of fitness costs of transmitted escape mutations 144

6.4.6 More frequent and costlier escape associated with protective HLA alleles 146

6.4.7 CTL escape is the dominant force of evolution 148

6.5 Discussion 150

7 Discussion 157

7.1 Summary of findings 157

7.2 Key themes 159

7.2.1 Multi-scale selection 159

7.2.2 The evolutionary rate mismatch: within and between host toggling 160

7.3 Strengths and Limitations 162

7.4 Future Outlook 164

7.5 Concluding remarks 164

8 Bibliography 165

9 Appendix 190

9.1 Chapter 3 Supplementary material 190

9.2 Chapter 4 Supplementary material 194

9.3 Chapter 5 Supplementary material 195

9.4 Chapter 6 Supplementary material 196

List of figures

FIGURE 1.1 THE STRUCTURE OF THE HIV GENOME ORDERED BY READING FRAME 17
FIGURE 2.1 DATASET SELECTION CRITERIA FOR THE ANALYSES OF EACH CHAPTER 36
FIGURE 3.1 WITHIN-HOST MODEL FITNESS AND EQUILIBRIUM VIRAL LOAD 51
FIGURE 3.2 WITHIN HOST VIRAL LOAD TRAJECTORIES 52
FIGURE 3.3 BETWEEN-HOST OUTCOMES AT ENDEMIC EQUILIBRIUM 54
FIGURE 3.4 BETWEEN-HOST OUTCOMES AT EQUILIBRIUM FOR AN INCREASINGLY HETEROGENOUS HOST POPULATION. 55
FIGURE 3.5 AVERAGE SPVL OVER TIME IN SIMULATED EPIDEMICS. 56
FIGURE 3.6 HERITABILITY ESTIMATES. 58
FIGURE 3.7 VIRAL LOAD PATTERNS IN TRANSMISSION PAIRS ENROLLED IN SERODIFFERENT STUDIES. 60
FIGURE 4.1 RECOMBINATION RATE DECREASES WITH TIME-SCALED DISTANCE. 76
FIGURE 4.2 AVERAGE RECOMBINATION RATE VARIES BY VIRAL LOAD AND STAGE OF INFECTION. 78
FIGURE 4.3 RECOMBINATION RATE /SITE/GENERATION FOR OVERLAPPING SLIDING WINDOWS OF LENGTH 500BPS, MOVING IN INCREMENTS OF 10BPS. 81
FIGURE 4.4 VARIATION IN RECOMBINATION RATES BY SUBTYPE. 83
FIGURE 4.5 FINDINGS ARE SUPPORTED BY ILLUMINA DATASET. 85
FIGURE 4.6 SUBSTANTIAL DIFFERENCES IN RECOMBINATION RATE IN ENV BY CODON POSITION. 86
FIGURE 5.1 DIVERGENCE RATES ACROSS THE HIV GENOME REVEAL DISTINCT PATTERNS BETWEEN WITHIN-HOST AND BETWEEN-HOST EVOLUTION. 105
FIGURE 5.2 CORRELATION BETWEEN WINDOWS-SPECIFIC RATES AT THE WITHIN AND BETWEEN HOST SCALES. 107
FIGURE 5.3 WITHIN-HOST EVOLUTIONARY RATES INFERRED BY BEAST BY GENE. 109
FIGURE 5.4 EVOLUTIONARY RATE VARIATION ACROSS HIV CODON POSITIONS DIFFERS BETWEEN GENES. 110
FIGURE 5.5 RATIOS BETWEEN BRANCH-SPECIFIC EVOLUTIONARY RATES. 111
FIGURE 5.6 EVOLUTIONARY RATES ON INTERNAL BRANCHES ARE CORRELATED BETWEEN VIRAL GENES WITHIN INDIVIDUALS. 113
FIGURE 5.7 WITHIN-HOST HIV EVOLUTIONARY RATES DECREASE WITH HIGHER VIRAL LOADS ACROSS MAJOR GENES. 114
FIGURE 5.8 BEAST WITHIN-HOST RATES SYSTEMATICALLY GREATER THAN DIVERGENCE RATES. 115
FIGURE 5.9 COMPARING EVOLUTIONARY RATES ACROSS METHODS AND SCALES 117
FIGURE 5.10 WITHIN-HOST DIVERGENCE AND EVOLUTIONARY RATES DECEASE OVER LONGER TIME SCALES 119

FIGURE 5.11 TRAJECTORY OF MINOR VARIANTS AT GENOMIC SITES THAT “TOGGLE” OR SWEEP OVER TIME 121
FIGURE 5.12 PERSISTENCE OF MUTATIONS ABOVE FREQUENCY THRESHOLDS 123
FIGURE 6.1 SUMMARY OF HLA GROUPS A, B AND C IN THE STUDY COHORT. 136
FIGURE 6.2 CTL ESCAPE IN RECIPIENT AND SOURCE INDIVIDUALS 138
FIGURE 6.3 RECURRENT ESCAPE AMONG INDIVIDUALS. 142
FIGURE 6.4 CHANGE IN SELECTION PRESSURE BETWEEN AND WITHIN HOST. 143
FIGURE 6.5 THE DISTRIBUTION OF FITNESS COSTS OF TRANSMITTED ESCAPE MUTATIONS 146
FIGURE 6.6 FASTER AND MORE FREQUENT ESCAPE IN INFECTIONS OF INDIVIDUALS WITH PROTECTIVE HLA-ALLELES. 148
FIGURE 6.7 TOTAL CONTRIBUTION OF CTL EVOLUTION COMPARED TO “BACKGROUND” EVOLUTION AT THE END OF ~ 1 YEAR OF INFECTION 150
FIGURE 9.1 THE FREQUENCY DISTRIBUTION OF THE VIRAL POPULATION AT EQUILIBRIA. 190
FIGURE 9.2 THE TOTAL PREVALENCE AT THE ENDEMIC STEADY STATE. 191
FIGURE 9.3 BETWEEN-HOST OUTCOMES AT EQUILIBRIUM FOR AN INCREASINGLY HETEROGENOUS HOST POPULATION FOR ALL MODELS 192
FIGURE 9.4 REGRESSION SLOPES OF SIMULATED VIRAL LOAD TRAJECTORIES WITH NOISE 193
FIGURE 9.5 SUBSTANTIAL DIFFERENCES IN RECOMBINATION RATE IN ENV BY CODON POSITION FOR SITES AT WHICH ALLELE FREQUENCY IS STABLE. 194
FIGURE 9.6 TRAJECTORIES OF TOGGLING MUTATIONS FOR ALL INDIVIDUALS. 195
FIGURE 9.7 INDIVIDUALS WITH SIMILAR HLA PROFILES DRIVE THE SAME ESCAPE MUTATION MORE OFTEN THAN BY CHANCE 196
FIGURE 9.8 WITHIN-HOST TRAJECTORIES OF TOGGLING MUTATIONS IN SOURCE INDIVIDUALS 197

List of tables

TABLE 3.1 VARIABLE AND PARAMETERS DESCRIPTIONS FOR THE WITHIN-HOST AND BETWEEN HOST MODELS. 46
TABLE 4.1 THE RECOMBINATION RATE /SITE/GENERATION BY GENE OR PROTEIN DOMAIN. 80
TABLE 4.2 THE RECOMBINATION RATE PER SITE PER GENERATION FOR HOT AND COLD SPOTS IDENTIFIED IN UNIQUE RECOMBINANT FORMS AND CIRCULATING RECOMBINANT FORMS SEEN IN AFRICA 82
TABLE 5.1 EVOLUTIONARY RATE MISMATCH BY SCALE AND METHODOLOGY. 118
TABLE 9.1 POSITIONS OF BETWEEN-HOST TOGGLING SITES. 198

Acknowledgements

First, I would like to thank my incredible supervisor, Katrina Lythgoe, whose kind support and guidance during the last three and a half years has been invaluable and I will always be so grateful for. Even with a growing research group and responsibilities, Katrina has continued to make time each week for our discussions, which always left me feeling excited and confident in my research. Thank you also to my secondary supervisor Christophe Fraser for providing insight, suggestions and support. Thank you to the current and past members of the Lythgoe, Fraser and Bonsall groups, who each have an incredible enthusiasm for their research. Their supportive and generous feedback has been hugely helpful in developing new ideas, and I will miss sharing an office with such lovely people. In particular, thank you to Lele Zhao whose work on HIV transmission in the Partners datasets built some of the foundational thinking behind this thesis, and who helped me when I began working with viral sequencing data in 2021 as a clueless maths graduate. And also thank you to Chris Wymant for your generous and insightful feedback on my modelling of virulence evolution. I am grateful to my external examiner, Roland Regeos, and internal examiner, Aris Katzourakis, for their careful reading of this thesis and for the constructive discussion during my viva examination. Their insightful questions and suggestions have contributed to strengthening this work.

I am deeply grateful to the ICRC at the University of Washington for conducting the HIV serodiscordant couples research studies, to the PANGEA consortium for their amazing work in sequencing the samples, and to the study participants who made this research possible. I also gratefully acknowledge the support of the Health Data Science Centre of Doctoral Training, and all the wonderful people I had the opportunity to meet through the programme.

Thank you to my family, most importantly my mum and dad who have always been so selflessly supportive of me. Thank you for solving my problems and for constantly believing in my abilities - I would not have completed this thesis without you. And apologies for almost a decade of student financial support, I'll pay you back one day.

Thank you to my siblings, Lizzie and Tim, for always making me laugh and keeping me humble. And for all my extended family – the Rainfords and the Peters – for their love, support and cocktails, and for being an incredible group of people that I am lucky to have been born into.

Finally, thank you to Lewis for being my biggest cheerleader and best friend since the first days of lectures in Uni Hall Bath campus back in 2016. He brings joy to my every day and is the best thing to have happened to me (yet).

This thesis is dedicated to Ben Peters

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Harriet Longley

2025

Papers for Publication

The work presented in **chapter three** considers the hypothesis that viral load is determined by the number of weakly deleterious mutations, and as a proof of principle I model this process with a nested modelling framework. The theoretical component of this work is under revision at *PLoS* Computational Biology and is available on BioRXV.

The work presented in **chapter four** describes the patterns of within-host recombination with whole genome long and short read data sequenced by the PANGEA consortium and sampled as part of the Partners studies described in chapter 2. This work published at *Virus Evolution* and can be found at https://doi.org/10.1093/ve/veaf052.

The work presented in **chapters five** and **six** considers the role of mutations toggling over time, both at the within-host and between-host scale, with the latter linked to immune-escape pressure varying across host. These processes are linked to the observed evolutionary rate mismatch across scales. Key aspects of this work across the two chapters are to be submitted for publication at *PLoS Biology* and will be available on BioRXV.

1 Introduction

1.1 Motivation and outline

Human Immunodeficiency Virus 1 (HIV-1, hereafter referred to as HIV) remains one of the most significant global public health challenges of our time, affecting approximately 39 million people worldwide with 1.3 million new infections occurring annually (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2024). Despite major advances in our understanding of the epidemiology and biology of HIV, as well as the development of antiretroviral therapy (ART), neither an effective vaccine nor cure are currently available. The elimination of HIV has yet to be realised in large part due to the extraordinary capacity of the virus to rapidly evolve in the face of selection pressure, namely from the cellular and humoral immune responses, as well as therapeutic interventions (Rambaut *et al.*, 2004; Deeks *et al.*, 2015). Additionally, the balance and potential conflict between within-host selection, which favours short-term survival and extensive replication in the host, and between-host selection, which influences transmission fitness, is not fully understood (Lythgoe and Fraser, 2012; Leitner, 2018).

This thesis explores the determinants and dynamics of within-host evolution and examines how these processes contribute to population-level viral evolution, with particular focus on the rate at which these adaptive processes unfold across different scales. Understanding these dynamics holds significant implications not only for the development of more effective HIV therapeutics and epidemiological control strategies, but also provides broader insights into the co-evolutionary dynamics of chronic viral infections, for which HIV serves as a model organism.

The first results chapter of this thesis examines a classic example of the tension between within-host and between-host viral evolution: the evolution of virulence. Using a multi-scale model to capture viral load dynamics, I explore a fundamental principle of evolutionary biology—the conflict of selective pressures at different biological scales. Subsequent chapters move beyond this theoretical framework to a data-driven investigation of the factors shaping HIV evolution within individual hosts and the broader implications for population-level evolution.

In the next chapter, I quantify the role of recombination across the HIV genome with unprecedented granularity and examine how selection pressures shape the impact of

recombination on viral evolution. I also explore how factors such as read length and the time intervals between sampling influence the accuracy of recombination measurements, conducting analyses on both long-read and short-read deep-sequenced datasets.

In the third results chapter, I build on these findings and assess the rate of HIV evolution across different scales and compare methodologies for quantifying the tempo of evolution, with a particular attention on understanding why measured evolutionary rates are markedly slower at the between-host level.

In the final results chapter, I present compelling evidence supporting the leading hypothesis for difference in evolutionary rates across scales — namely, the reversion of host-specific adaptations after transmission to a new host due to the innate fitness cost of immune escape.

This work leverages a unique and remarkable dataset comprising of hundreds of HIV transmission pairs, providing novel insights into the dynamics of viral evolution. For each individual who seroconverted over the course of the study, the viral population from early into infection was sequenced, typically a matter of weeks post seroconversion, and then multiple times for the next one or two years. This early sampling allows us to capture a snapshot of the virus before it undergoes extensive adaptation within the host. I also analyse data generated by two different next-generation sequencing approaches, enabling comparison of results using long-read and short-read data. In addition to the viral sequences, human leukocyte antigen (HLA) data is available for a subset of the individuals included in the study, allowing us to directly link evolutionary dynamics to changes in host selection pressure, specifically the CTL response. Together, this dataset offers a powerful combination of breadth and depth, enabling detailed investigations of viral recombination, selection, and adaptation within-host.

Collectively, this thesis offers a comprehensive investigation into the multi-scale dynamics of HIV evolution, providing deeper insight into the complex interaction between within-host processes and between-host constraints. Before delving into the specific analyses, the following sections provide an overview of HIV's epidemiology, biology, and evolutionary dynamics. These foundational concepts are critical for understanding the processes that underpin HIV's rapid adaptation on both scales.

1.2 HIV epidemic background

HIV is assumed to have first spread from non-human primates to humans during the early 20th century (Keele *et al.*, 2006; Faria *et al.*, 2014). Phylogenetic studies have shown that HIV Group M – the most widespread globally - emerged in West Africa in the 1920s, where circulation continued unnoticed until the 1980s, and in 1983 HIV was determined as the causative agent of the AIDS (auto-immune deficiency syndrome) pandemic (Barré-Sinoussi *et al.*, 1983). Over the following decades, HIV has diversified into numerous co-circulating subtypes and recombinant forms which are typically geographically distinct and exhibit substantial genetic diversity. Globally, almost half of all infections are subtype C (47%), followed by B (12%) and A (10%), as well as circulating recombinant forms CRF02_AG and CRF01_AE (5%) (Hemelaar *et al.*, 2019). Given the historical dominance of subtype B in Western Europe and North America, many early studies and clinical trials focused on subtype B infections, leading to potential gaps in our molecular understanding of other subtypes (Bbosa, Kaleebu and Ssemwanga, 2019; Igiraneza *et al.*, 2024).

Breakthroughs in antiretroviral therapy (ART) since the late 1980s have transformed HIV from a fatal infection to a chronic condition with near normal life expectancy for those for whom therapy is effective. The advent of combined ART (cART) in the mid-1990s proved more effective against drug resistance than single-drug therapy (Gulick *et al.*, 1997), and pre-exposure prophylaxis (PrEP) has dramatically reduced infection risk in at-risk communities (Grant *et al.*, 2010). However, ART regimes must be strictly adhered to be effective, and it is reported that approximately a quarter of people living with HIV have not had access to treatment (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2024). Over four decades on from the first identification of the virus, HIV remains a leading cause of morbidity and mortality worldwide, with the burden of disease greatest in sub-Saharan Africa (Hemelaar *et al.*, 2019; Carter *et al.*, 2024).

1.3 HIV Biology

Understanding the biology of HIV infection and replication within the host is fundamental to studying within-host evolutionary processes, as stages such as cell entry, replication, and immune evasion provide opportunities for fitness differences between viral genomes to emerge and be selected for.

HIV primarily infects immune cells that are crucial for maintaining the immune system's functionality, particularly activated CD4+ T cells, although other immune cells can also be targeted, such as macrophages and microglial cells (Chan and Kim, 1998). The infection begins when HIV's surface glycoprotein, GP120, binds to the CD4 receptor on the surface of these target cells (Kwong *et al.*, 1998). This binding triggers a conformational change in GP120—a structural rearrangement of the protein—that allows it to interact with one of two co-receptors: CCR5 or CXCR4 (Bleul *et al.*, 1996; Dragic *et al.*, 1996; Checkley, Luttge and Freed, 2011). CCR5 is predominantly used during the early stages of infection, whereas CXCR4 is more commonly associated with later stages (Connor *et al.*, 1997). This interaction pulls the viral and cell membranes closer together.

Following this, another viral protein, GP41, undergoes a structural change that facilitates the fusion of the viral envelope with the host cell membrane (Caillat *et al.*, 2021). This fusion enables the viral core—the innermost part of the virus, which contains genetic material and associated proteins—to enter the cell. Inside the cell, the core disassembles, releasing the viral RNA and enzymes into the cytoplasm. The viral RNA is then reverse transcribed into DNA, which integrates into the host genome (Aiken and Rousso, 2021). Once integrated, HIV begins replicating within the host cell, producing new virions that are released to infect additional immune cells (Roberts, Bebenek and Kunkel, 1988). The reverse transcriptase enzyme lacks proofreading activity, resulting to an error rate estimated to be of the order 10^{-5} mutations per site per replication cycle, equivalent to one error on the genome for every new virus particle (Roberts, Bebenek and Kunkel, 1988; Perelson *et al.*, 1996).

1.3.1 The HIV genome

The HIV genome consists of approximately 9.7 kb of single-stranded, positive-sense RNA, with each virion carrying two copies, making it pseudodiploid. It contains nine genes that encode around 15 proteins, most of which overlap in reading frames (Turner and Summers, 1999) (Figure 1.1). The three largest genes – GAG, POL, ENV - are common to retroviruses, and they provide the basic structural and enzymatic machinery. GAG encodes the structural proteins that form the viral matrix, capsid and nucleocapsid, which together enable crucial viral functions, such as budding from infected cells and RNA packaging (Klingler *et al.*, 2020). The POL gene codes for three enzymes that are essential for viral replication – reverse transcriptase, integrase and

protease – and mutations within the gene can be lethal to the virus, which is why the gene is relatively conserved from an evolutionary perspective, and why it is a target for anti-retroviral drugs (Hill, Tachedjian and Mak, 2005). The envelope gene, ENV, encodes the two envelope glycoproteins GP120 and GP41, which are critical for cell entry, as previously described. While the functional regions needed for receptor binding are relatively conserved, the exposed surface loops that are visible to antibodies vary significantly to evade recognition (Wei *et al.*, 2003; West *et al.*, 2014), making ENV incredibly diverse, with 20-35% sequence divergence across subtypes (Lynch *et al.*, 2009).

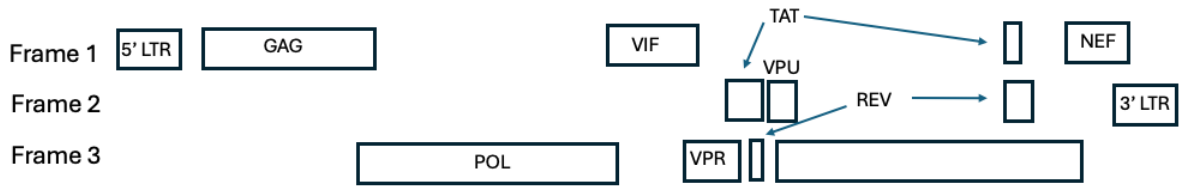

Figure 1.1 The structure of the HIV genome ordered by reading frame

In addition to the structural gene, there are two regulatory genes (TAT and REV) which control viral gene expression, and four accessory genes (VIF, VPR, VPU and NEF) that help the virus to evade the host immune response and enhance cell-to-cell transmission (Emerman and Malim, 1998).

1.3.2 HIV reservoir

A critical challenge in HIV treatment is the establishment of a latent reservoir, primarily in memory CD4+ T cells (Chen *et al.*, 2022). This reservoir forms rapidly, with studies in Simian Immunodeficiency Virus (SIV) models demonstrating establishment within days of initial infection (Whitney *et al.*, 2014). Once integrated into these long-lived memory cells, the viral genome can persist indefinitely in a dormant state due to homeostatic proliferation (self-renewal) of these cells (Chomont *et al.*, 2009; Buzon *et al.*, 2014). While ART effectively prevents the infection of new cells by blocking viral replication, it cannot eliminate virus already integrated into these reservoir cells. Consequently, if treatment is interrupted, these latent proviruses can reactivate, leading to rapid viral rebound and resumption of active infection. There are several unknowns and open questions on the reservoir, such as what causes viral reactivation

from the latency state (Dahl, Josefsson and Palmer, 2010). From an evolutionary perspective, the reservoir essentially serves as a time-capsule for the virus and may offer a way for the virus to bypass transmission costs associated with escape from the immune response (Lythgoe *et al.*, 2017).

1.3.3 Stages of infection and pathogenesis

An untreated HIV infection progresses through three distinct phases. The acute phase develops within 2-4 weeks of infection, typified by rapid viral replication leading to peak plasma viral loads and a sharp decline in CD4+ T cells (Clark *et al.*, 1991). Around 6 months post-infection, the immune response has better developed, and viral loads stabilise at what is known as the set-point viral load (spVL), marking the beginning of the chronic phase (Koup *et al.*, 1994). This chronic phase typically lasts 8-10 years without treatment, though this duration varies considerably between individuals, from 2 to over 15 years (Moss *et al.*, 1988). During this period, there is ongoing immune activation, continuous viral evolution and immune escape, accompanied by a gradual decline in CD4+ T cells. Finally, when CD4+ T cells become severely depleted, the infection progresses to AIDS, characterised by increasing viral loads, a severely compromised immune system, and increased susceptibility to opportunistic infections and cancers.

The spVL is a strong predictor of disease progression, with lower viral loads associated with slower progression to AIDS (Mellors *et al.*, 1996). The spVL varies substantially between individuals, up to 6 orders of magnitude (Fraser *et al.*, 2007), and has been shown to be a heritable trait across transmission pairs, suggesting viral genetic factors influence this variation (Hollingsworth *et al.*, 2010; Bonhoeffer, Fraser and Leventhal, 2015; Blanquart *et al.*, 2016; Bertels *et al.*, 2018). While the exact mechanisms driving spVL variation are not fully understood, insights into host factors have come from studying 'elite controllers'—rare individuals who maintain very low viral loads without treatment, specifically those carrying particular protective HLAs such as B*57 (Borrell *et al.*, 2021).

Naively, it might be expected that viral loads would increase over time as the virus variants with greatest replicative capacity are selected, yet between-host selection of variants causing infection with intermediate viral loads that maximise transmission potential has been shown to dominate (Fraser *et al.*, 2007; Lythgoe, Pellis and Fraser,

2013). In chapter 3, I review the conflict between short-sighted evolution at the within-host scale with the between-host level selection of maximising transmission, and the pose a mechanism that can explain the multiple conundrums of HIV virulence evolution.

1.3.4 Immune response

HIV infections trigger robust but ultimately inadequate immune responses from the host (Deeks and Walker, 2004; McMichael *et al.*, 2010). The response involves both the cellular and humoral immune mechanisms, and an arms race ensues between viral evolution and immune adaptation. While the immune system continuously develops new responses, a high mutation rate and evolutionary strategies for evasion allow the virus to stay one step ahead, leading to eventual immune system deterioration in untreated infections.

The cytotoxic T lymphocyte (CTL) response, mediated by CD8+ T cells, plays an important role in early viral control and the establishment of the spVL. CTLs detect and eliminate infected cells by recognising short viral peptides, known as epitopes, that are presented by HLA class I molecules on the cell surface (McMichael and Rowland-Jones, 2001). The genes encoding HLA molecules are highly diverse across human populations, and consequently, the viral epitopes presented to CTLs differ between individuals (Carlson and Brumme, 2008; Carlson *et al.*, 2012; Kløverpris, Leslie and Goulder, 2016; Li *et al.*, 2023). The virus has evolved multiple escape strategies to evade CTL recognition, primarily through mutations in these epitopes that prevent HLA binding or CTL recognition while maintaining viral function (Allen *et al.*, 2005; Carlson *et al.*, 2012). Eventually, after long-term antigenic stimulation, T cells become exhausted, losing their cytotoxic function and cytokine production capacity, further compromising immune control of the virus (Day *et al.*, 2006; Trautmann *et al.*, 2006).

The humoral immune response – specifically antibodies – is an active area of research for therapeutics, particularly for preventing new infections. Early into infection, antibodies typically target the GP41 protein, however they are largely unsuccessful in preventing cell infection (Tomaras *et al.*, 2008; Goonetilleke *et al.*, 2009; McMichael *et al.*, 2010). During chronic infection, neutralising antibodies (NAbs) appear which an inhibit cell entry, however efficacy is limited (Poignard *et al.*, 1999). More potent are

broadly neutralising antibodies (bNAbs) which can develop several years into infection and target conserved genome regions (Burton and Hangartner, 2016). However, bNAbs develop in only a small number of people (10-30%) and often arise too late in natural infection to be effective for viral control (Doria-Rose *et al.*, 2009; Sather *et al.*, 2009; Simek *et al.*, 2009). Despite these challenges, bNAbs have shown promise as a basis for therapeutic interventions, and ongoing research aims to harness their potential for vaccine and antibody-based therapies (Bonsignori *et al.*, 2017; Caskey, 2020).

1.3.5 Compartmentalisation

HIV exhibits a metapopulation structure within hosts, where viral populations in different anatomical compartments evolve semi-independently while maintaining limited genetic exchange (Chun *et al.*, 1997; Finzi *et al.*, 1997). These compartments include blood, lymphoid tissues, the central nervous system, genital tract, and other tissues, each presenting distinct evolutionary challenges. The selective pressures vary between compartments, for example due to differences in target cell populations and immune responses (Kimata, Rice and Wang, 2016; Sengupta and Siliciano, 2018). A well-studied example is the blood-brain barrier that creates a unique environment within the central nervous system, leading to viral variants that have adapted to infect different types of cells using different entry mechanisms (Ash, Al-Harthi and Schneider, 2021). A metapopulation structure has significant clinical implications as it can create viral reservoirs that persist despite therapy, enable independent evolution of drug resistance in different tissues, and complicate sampling and monitoring of viral populations (Sengupta and Siliciano, 2018) .

In this thesis, viral populations were sampled and sequenced from blood plasma. While blood represents a well-mixed and clinically significant compartment that reflects infection dynamics, this sampling approach cannot capture the full complexity of viral evolution. This point is particularly relevant in the context of transmission and relating within-host and between-host evolution, where the transmitted virus typically originates from the genital tract and may carry genetic signatures distinct from those observed in blood plasma (Lythgoe and Fraser, 2012).

1.4 Within-Host evolution of HIV

1.4.1 Sources of variation

Over the course of an untreated infection, HIV evolves incredibly rapidly, with diversity levels reaching up to 5-10% and 10% divergence from the founder variant within a few years of infection (Shankarappa *et al.*, 1999). As previously described, the mutation rate of the virus is exceptionally high at approximately 10^{-5} per site per generation (Mansky, 1996b), such that mutations are generated in almost every replication round. With early viral loads on the order of 10^{6} (Lyles *et al.*, 2000; Richardson *et al.*, 2003) and a short generation time of between 1-2 days (Perelson *et al.*, 1996), the pool of viral variants upon which selection can act is vast.

Genetic recombination is a key mechanism by which the virus is able to prevent the accumulation of deleterious mutations, as well as combining beneficial mutations onto a single genome. Recombination occurs during reverse transcription when the enzyme switches between two different viral RNA templates present in a multiply-infected cell (Zhuang *et al.*, 2002; Onafuwa *et al.*, 2003; Chen *et al.*, 2009), resulting in a daughter virus with a mosaic genome that contains genetic material from both parental viruses. The rate of recombination varies across studies, however there is strong evidence that the recombination rate is of the order or exceeds the mutations rate, which makes it one of the highest of all organisms (Neher and Leitner, 2010; Batorsky *et al.*, 2011; Grant *et al.*, 2020; Romero and Feder, 2024). On the other hand, substantial recombination can hinder adaptation through the breaking of beneficial epistatic interactions.

HIV recombination observed *in vivo* is a composite outcome of multiple distinct biological processes operating at different scales. At the molecular level, the primary mechanism is template switching during reverse transcription—a process estimated to occur between 2 to 20 times per genome per replication cycle (Jetz et al 2000, Jung et al 2002). This mechanism depends critically on co-packaging, wherein two genetically distinct RNA genomes are encapsidated within a single virion. Co-packaging, in turn, requires prior cellular co-infection by genetically distinct viruses, though the efficiency of subsequent template switching is also influenced by the degree of sequence homology between genomes.

Crucially, the recombination rates reported in clinical studies represent *effective* rates that integrate both the frequency of recombination events and the influence of selection. Recombinant genomes may be favoured by positive selection when they produce advantageous combinations or purged by negative selection if recombination disrupts co-adapted genetic elements. These selection pressures can significantly distort the observed frequencies of recombinant forms, potentially obscuring the underlying rate at which recombination events actually occur. The different processes that contribute to the measured recombination rate *in-vivo* are further discussed in chapter 4.

In the case of a dual infection where an individual is infected by two distinct viral variants, recombination can generate novel viral forms. This process operates both within and between HIV subtypes, contributing significantly to global viral diversity through the emergence of circulating and unique recombinant forms (CRFs and URFs) (Blackard, Cohen and Mayer, 2002). The success and spread of these recombinant variants demonstrate their epidemiological significance and suggests recombination can enhance transmission fitness, ultimately influencing HIV evolution at the population level.

Extensive recombination has implications more broadly for evolutionary inferences, as many phylogenetic approaches assume each sequence has a single evolutionary history. The presence of recombination can lead to misleading results, such as artificially inflated branch lengths or false signals of natural selection (Schierup and Hein, 2000). In molecular evolution studies, researchers can address this problem by identifying and excluding recombinant sequences from their analyses. However, this approach fails at the within-host level, where sequence diversity is too low for standard recombination detection methods to work effectively, and even at the between-host level removing all recombinants is challenging. To address this, recombination-aware phylogenetic approaches have been developed, such as methods designed for bacterial genomes (Didelot and Wilson 2015) and software for Ancestral Recombination Graph (ARG) inference (Hubisz and Siepel 2020). However, established approaches for handling recombination in within-host viral phylogenies of rapidly evolving viruses like HIV remain lacking.

In chapter 4, I characterise recombination across the genome using a sliding window approach by quantifying the rate of linkage-disequilibrium decay over time and

distance, and highlight variation by viral load, stage of infection, subtype, genome position and codon position. My analysis provides a comprehensive view of recombination dynamics in this system.

1.4.2 Viral population dynamics

The relative importance of selection versus genetic drift in driving within-host evolution remains debated. Despite vast population sizes, the effective population size, *Ne*, is likely substantially smaller, meaning drift may play an important role in determining which variants reach high frequency and shape the viral population. Estimates of the effective population size range from 10^3 (Brown, 1997; Nijhuis *et al.*, 1998), where stochastic effects would dominate, up to 10^6, where the dynamics are largely deterministic (Coffin, 1995; Rouzine and Coffin, 1999). However, low estimates of *Ne* are partly due to the assumption of neutrality despite evidence of selection, e.g. at immune escape sites (Leitner, 2018). Selection leads to low *Ne* estimates because of the increase in variance in reproductive success among viral lineages. Both purifying and diversifying selection affect *Ne,* but through different mechanisms. Purifying selection reduces genetic diversity through background selection, eliminating deleterious mutations and linked neutral variation. Diversifying selection at immune escape sites creates selective sweeps that temporarily reduce genome-wide diversity, mimicking demographic bottlenecks.

Overall, the evidence favours positive selection as the major driving force for within-host evolution, supported by consistent patterns of immune escape across hosts, phylogenetic signals of positive selection, and the conservation of important functional constraints (Rambaut *et al.*, 2004). However, genetic drift becomes an important force during population bottlenecks, such as at transmission, after acute infection and at the beginning of antiretroviral therapy (Gutiérrez, Michalakis and Blanc, 2012; Joseph *et al.*, 2015).

1.4.3 Viral quasispecies framework

Within-host viral populations can be described as a quasispecies, defined as a dynamic and diverse cloud of closely related genetic variants (Wilke, 2005). This concept, initially developed for self-replicating molecules by Eigen and Schuster (1979), has become particularly influential in understanding RNA virus evolution, including HIV. Unlike population genetic approaches that traditionally characterise

viral populations through consensus sequences, the quasispecies framework emphasises that RNA viruses exist as mutant spectra or clouds—collections of genetically related variants that arise due to error-prone replication and collectively contribute to the characteristics and fitness of the viral population.

Quasispecies theory is based on several key assumptions that shape its conceptual and mathematical foundation (Sardanyes et al 2024). First, it assumes high mutation rates near an “error threshold,” where frequent mutations generate a diverse cloud of related genotypes rather than a single dominant sequence, and the error threshold represents a critical mutation rate beyond which selection can no longer maintain the population's genetic information, leading to fitness collapse and genetic meltdown. Second, it posits that natural selection acts on this entire mutant spectrum—referred to as the quasispecies—as a collective unit, rather than on individual genomes. A related assumption is mutational coupling, which suggests that variants interact through mechanisms such as complementation or interference, meaning their evolutionary fate is interdependent. Classical models also assume infinite population sizes, enabling the stable maintenance of genetic diversity—an assumption often violated in real-world scenarios like transmission bottlenecks. Finally, these models typically employ simple, often single-peak fitness landscapes that may not reflect the complex, rugged adaptive landscapes viruses encounter. Together, these assumptions give quasispecies theory its explanatory power, however they can also limit the applicability to real-life systems.

The relevance of quasispecies theory in virus evolution remains controversial. Critics argue that many of the phenomena attributed to quasispecies dynamics—such as genetic diversity, mutation–selection balance, and adaptation—can be adequately explained by classical population genetics without invoking additional mechanisms (Holmes 2010). One major criticism described by Holmes is that the mutation rates required for genuine quasispecies behaviour, particularly the dominance of a mutant cloud around a "master sequence," may be unrealistically high even for RNA viruses. Another is that selection operates primarily on individual genetic variants, not on the population as a whole, undermining the idea of the quasispecies as a collective unit of selection. Furthermore, critics contend that the theory is often applied too loosely, with the term “quasispecies” used to describe any virus population with genetic diversity,

regardless of whether the underlying assumptions—such as infinite population size, high mutation rates, and mutational coupling—are met.

Despite these concerns, quasispecies theory remains a useful framework for studying rapidly mutating viruses like HIV when the limitations are clearly described. When applied with awareness of the assumptions, the framework can offer insights beyond standard population genetics, such as for the study of collective behaviour or when modelling allele frequencies is not feasible, as is the case in chapter 3.

The details of the quasispecies theory and mathematical formulation are further defined in chapter 3, and the motivation for applying this controversial theory is outlined.

1.4.4 Selection pressure

The specific nature of HLA-restricted CTL recognition means that viral adaptation patterns are largely determined by each individual's HLA alleles, creating distinct evolutionary trajectories in different hosts (Goulder and Walker, 2012). However, the extent of CTL-driven evolution across individuals selection remains debated. While some studies report rapid selection of escape mutations (Goonetilleke *et al.*, 2009), distinguishing these mutations from transmitted/founder (TF) virus effects is challenging. Although earlier work suggested that over 50% of fixed mutations outside of the envelope in early infection represent CTL escape (Allen *et al.*, 2005), studies accounting for founder effects found that a third of individuals showed no evidence of CTL escape during the first two years of infection (Roberts *et al.*, 2015).

The T242N mutation in the TW10 epitope of GAG is a well-studied example of a predictable escape pattern, consistently emerging during acute infection in individuals carrying the protective HLA-B*57/58:01 allele (Leslie *et al.*, 2004; Crawford, *et al.*, 2007; Carlson and Brumme, 2008). These predictable evolutionary pathways, shaped by host HLA profiles, suggest that HLA diversity within populations may influence the evolution at a population level, as adaptation to common HLA alleles leaves lasting imprints on circulating viruses (Moore *et al.*, 2002; Carlson *et al.*, 2008; Kawashima *et al.*, 2009).

Escape mutations typically carry fitness costs and are expected to revert when transmitted to hosts with different HLA profiles (Friedrich *et al.*, 2004; Leslie *et al.*, 2004; Li *et al.*, 2007; Davenport *et al.*, 2008). While experimental and *in vivo* evidence

supports this reversion pattern, most studies have focused on a small set of HLA alleles, particularly those defined as protective due to their association with long-term viraemic control and which have been shown to target conserved epitopes (Crawford *et al.*, 2007). This leaves broader questions about the prevalence of escape, associated fitness costs, and patterns of host-specific adaptation for a broader range of host genetic backgrounds.

The fitness costs of immune escape can be mitigated by compensatory mutations that restore viral replicative capacity, for example the H219Q mutation is reported to compensate fitness costs of the T242N mutation (Brockman *et al.*, 2007; Crawford *et al.*, 2007; Liu *et al.*, 2014). When these compensated escape variants are transmitted, they may persist even in hosts with different HLA profiles, potentially obscuring the historical relationships between specific HLA alleles and viral mutations. This compensation process has important implications for viral evolution: while individual escape mutations might be costly, the accumulation of compensatory mutations can lead to their maintenance in the viral population even in the absence of the selecting HLA allele.

The envelope protein also faces strong antibody pressure on the virus surface, leading to multiple evasion strategies including modifications to N-linked glycosylation sites and changes in variable loop regions (V1-V5) within GP120 (Wei *et al.*, 2003; Derdeyn *et al.*, 2004). While these adaptations help the virus escape antibody recognition within a host, their impact on transmission fitness remains complex. Evidence suggests that some escape mechanisms, particularly an increase in the number of glycosylation sites (glycan density), may reduce transmission efficiency, creating a trade-off between within-host survival and between-host spread (Chohan *et al.*, 2005).

1.4.5 Fitness landscape

With longitudinal sequencing of within-host viral populations it is possible to measure changes in frequency of mutations over time and quantify their fitness cost. The vast majority of mutations that are generated are deleterious, with a proportion that are neutral and a small minority that are beneficial to either viral replication or in immune evasion. Synonymous mutations - nucleotide substitutions that do not alter the amino acid sequence – are typically assumed to be neutral, however a study quantifying fitness costs of synonymous mutations in the C2-V5 regions of ENV found an

unexpectedly high fraction of these mutations to be deleterious, particularly those in RNA stems that flank the variable loops (Zanini and Neher, 2013).

An in-depth study of whole genome deep sequencing data of longitudinal samples of HIV infection characterised the innate fitness cost of mutations for every site along the genome (Zanini *et al.*, 2017). The study found half of all non-synonymous mutations carried a large fitness cost (defined by the researchers as greater than 10 percent frequency change per day), while the majority of synonymous mutations carried small fitness costs (less than 1% frequency change per day), with the exception of those in important RNA structures and regulatory regions. Most importantly, the fitness costs identified within-host were consistent across the 10 individuals included in the study, implying that the landscape of fitness costs is a universal property of the virus, while mutations assisting in immune evasion vary in their cost and benefits across hosts. The concept of a "universal fitness landscape" has been further supported by the consistency in regions of high and low divergence both across individuals and when measured at the within and between-host scales (Zanini *et al.*, 2015), and sites that are diverse within-host have been shown to be diverse between-hosts, suggesting sites are repeatedly under selection (Illingworth *et al.*, 2020).

1.4.6 The within-host evolutionary rate

HIV evolves at an extraordinary rate over the course of an infection, with estimates ranging from 10^{-3} to 10^{-2} substitutions per site per year (Lemey, Rambaut and Pybus, 2006; Lemey *et al.*, 2007; Alizon and Fraser, 2013; Novitsky *et al.*, 2013; Zanini *et al.*, 2015; Raghwani *et al.*, 2018; Druelle and Neher, 2023), and rate estimates can vary significantly across individuals. Proposed drivers of this heterogeneity include viral load, strength of the immune response, and replication dynamics (Lemey *et al.*, 2007; Piantadosi *et al.*, 2009). Notably, the observation that evolutionary rates correlate across different genes within an individual suggests that viral factors may systematically shape the tempo of evolution across the entire genome. Rates have also been observed to lower over the course of infection, due to reduced immune response, changes in population size and the availability of target cells, however the studies demonstrating this trend are based on analysis of ENV sequences only (Shankarappa *et al.*, 1999).

Comparisons of evolutionary rate estimates across studies are challenging due to differences in methodologies, including sequencing technologies, sampling frequency, and the specific genomic regions analysed. Furthermore, the choice of evolutionary model can significantly impact rate estimation, particularly whether the model accounts for multiple substitutions at the same site (Tay, Kocher and Duchene, 2024).

1.5 Between-host evolution of HIV

Within-host evolution is primarily driven by positive selection, whereas between-host evolution is a comparatively more neutral process. Phylogenetically, within-host tree structures tend to be ladder-like, reflecting the ongoing selection of beneficial variants (Rambaut *et al.*, 2004; Lemey, Rambaut and Pybus, 2006). In contrast, between-host evolution is characterised by the coexistence of multiple lineages. The rate of evolution varies significantly both within and between subtypes, influenced by epidemiological factors such as transmission rates and broader determinants like social behaviour, HLA diversity within host populations, and access to ART (Nasir *et al.*, 2021).

A critical aspect of between-host evolution is the transmission bottleneck. During sexual transmission, it has been demonstrated that a single viral variant typically establishes a new infection (Carlson *et al.*, 2014; Joseph *et al.*, 2015), however the bottleneck is wider for other modes of transmission (Bar *et al.*, 2010). The bottleneck drastically reduces genetic diversity and the effective population size, amplifying the role of genetic drift in shaping evolutionary dynamics. Evidence suggests that evolutionary rates differ among at-risk groups, which has been linked to sex ratios within risk groups and differences in viral loads between men and women (Vrancken *et al.*, 2015). The time between infection and onward transmission also influences the evolutionary rate, with transmission earlier into infection lowering the overall rate, and the evolutionary rate therefore slows down when the epidemic rate increases. (Maljkovic Berry *et al.*, 2009)

In the case of sexual transmission, the narrowness of the transmission bottleneck is attributed to strong selection pressures within both the source host and the recipient. These pressures include physical barriers, such as the mucus layer in the female genital tract (Haaland *et al.*, 2009), and immunological barriers, including co-receptor usage and the density of N-linked glycosylation sites as previously described.

1.5.1 The evolutionary rate mismatch

Studies of within and between host evolution show that the rate of evolution is 2-5 times faster at the within-host level compared to the between-host level, and this holds true for both synonymous and non-synonymous mutations (Lemey, Rambaut and Pybus, 2006; Alizon and Fraser, 2013; Raghwani *et al.*, 2018). The selection at the point of transmission is assumed to contribute to slowing the accumulation of mutations at a population level. Certain viral phenotypes, such as co-receptor usage, are linked with higher probability of transmission. Specifically, while CCR5-tropic viruses dominate during transmission and early infection, CXCR4-tropic variants are preferred during chronic infection. Reactivation of ancestral virus in the viral reservoir with greater similarity to the T/F virus provides a mechanism by which virus with a transmission advantage continues to circulate. This process is called “store and retrieve” and is supported by the finding that lineages with slower rates are transmitted, and evidence that the founder virus in a recipient is more genetically similar to “older” source virus (Lythgoe and Fraser, 2012; Redd *et al.*, 2012; Lythgoe *et al.*, 2017).

Other hypotheses include the so-called “escape and revert” mechanism (also known as “adapt and revert”), where mutations are advantageous to the virus in one individual in evading the immune response but are no longer beneficial upon transmission to a new host environment and “revert” (Leslie *et al.*, 2004; Herbeck *et al.*, 2006; Zanini *et al.*, 2015; Illingworth *et al.*, 2020). Under this hypothesis, we expect the subtype-specific consensus sequence to represent a universally fit viral state in the absence of immune pressure. Over longer timescales, evolution towards these subtype-specific consensus sequences is observed, contributing to a slowing of evolutionary rates at the population level (Druelle and Neher, 2023).

An additional complication is that estimates of evolutionary rates of viral pathogens have been shown to decay over longer timescales of measurement, known as the time-dependent rate phenomenon (Ho *et al.*, 2011; Duchêne, Holmes and Ho, 2014; Aiewsakun and Katzourakis, 2016). Extensive site saturation helps explain this relationship, and in HIV specifically it has been demonstrated that when divergence is measured from a more phylogenetically distant reference sequence, the measured rate of divergence is slower over time at both the within and between host scales (Druelle and Neher, 2023). As a corollary, sampling frequency could significantly impact rate estimation - dense temporal sampling within hosts may detect short-lived

variants that inflate within-host rates, likely contributing to the observed rate mismatch. In chapter five I review the hypotheses for the evolutionary rate mismatch in greater detail, and in chapter six I examine the “escape and revert” hypothesis and the supporting literature in-depth, as well as provide new supporting evidence.

1.6 Thesis Outline and Key Findings

In chapter three, I adapt a multi-scale mathematical model to explore our hypothesis that viral load is determined by the number of weakly deleterious mutations. The evolution of virulence, for which viral load is a proxy, is an excellent case study in understanding evolution across scales, as the short-term evolutionary effect of selection for high replicative capacity conflicts with the long-term fitness of the virus that is maximised when transmission rates are high. I apply a nested model framework, where a within-host quasispecies model is nested in a between-host susceptible-infected (SI) model, and viral load is determined at any moment in time by the composition of the quasispecies, where I assume that the higher the number of deleterious mutations carried by population of viral variants, the lower the associated viral load due to fitness costs.

I compare outcomes for an increasing number of evolving positions in the viral genome (sites), where the higher the number of sites, the smaller the associated fitness effect of an additional mutation. This model reveals that when viral load is determined by a sufficiently large number of sites, each with small individual effects, a mutation-selection balance emerges that constrains viral loads. This finding helps resolve several paradoxes in HIV virulence evolution: the stability of the spVL despite ongoing genetic change, the heritability of viral load across transmission pairs, and the selection for intermediate viral loads at the population level. Using the longitudinal dataset analysed in subsequent chapters, I provide empirical support for our model's predictions by demonstrating a regression-to-the-mean in viral load over time.

Following from chapter three where mutation alone generates genetic diversity, I next introduce the role of recombination in chapter four. I apply a method that measures recombination by quantifying the relationship between linkage decay between two sites and the product of the time between sampling points and genetic distance. With longitudinal samples from over 300 individuals, representing multiple subtypes and both early and late infection, I identify hot and cold spots across the genome by taking

a sliding window approach. I demonstrate that variation in recombination rate is largely consistent across subtype, sequencing platform and at the between-host level. Additionally, I show that rates vary significantly between codon position, suggesting the interconnectedness between recombination and selection. Finally, by considering both our sequencing data and a simulated dataset I highlight the sensitivity of the rate estimation to read length and sampling density.

In chapter five, building on the finding that recombination breaks down linkage beyond 500 base pairs (bps), I infer evolutionary rates in 500bp windows across the genome during the first 12-18 months of infection, using longitudinal sequence data from 89 individuals with three or more sampling timepoints. By tracking sequence divergence from the founder virus across the genome, I show that patterns of high and low evolutionary rates mirror those observed at the between-host level. I compare two methodological approaches: direct measurement of sequence divergence and Bayesian evolutionary analysis (BEAST). The latter, which accounts for the full evolutionary process, yields higher rate estimates, likely due to capturing "toggling" mutations that transiently rise in frequency. I demonstrate that evolutionary rates correlate across genes within hosts but decrease when measured over longer time periods. This analysis highlights how methodological choices influence rate estimation and suggests that toggling mutations contribute substantially to elevated measured within-host evolutionary rates.

Finally, I explore the “escape and revert” hypothesis in explaining the evolutionary rate mismatch in chapter 6. I bring together longitudinal sequence samples of 62 transmission pairs, corresponding HLA genotype data for both source and recipient, and known literature of epitope variants and their corresponding HLA restrictions. I identify likely escape mutations, showing that the large majority of escapes are unique to the individual. I find that protective HLA alleles lead to faster and more costly escape, and I provide examples of the escape and revert process in action across transmission pairs.

1.7 Significance

This work advances our understanding of HIV evolution in several important ways. By leveraging a unique dataset of HIV transmission pairs with early infection sampling, comprehensive sequencing data, and HLA information, I provide unprecedented

insights into viral evolutionary dynamics across different biological scales. My findings have both fundamental and practical implications.

At a fundamental level, this work clarifies the complex relationship between within-host and between-host evolution. Through analysis of virulence evolution, recombination patterns, and evolutionary rates, I demonstrate how selection pressures at different scales shape viral adaptation. Our multi-scale modelling demonstrates how mutation-selection balance can explain conundrums of the evolution of HIV virulence, while our empirical analyses illuminate the role of recombination in shaping genetic diversity and adaptation.

The findings have significant implications for HIV treatment and prevention strategies. My analysis of how read length and sampling frequency affect measurements of recombination and evolutionary rates helps establish the limitations and requirements for accurate viral surveillance, which is particularly relevant as sequencing technologies continue to evolve. By tracking the escape-and-revert cycle in transmission pairs, I identify specific epitope variants that revert upon transmission, suggesting these regions carry substantial fitness costs. These highly constrained epitopes may represent promising targets for CTL-based vaccines or therapeutics, as viral escape from immune pressure targeting these regions appears particularly costly.

More broadly, this research contributes to fundamental principles in evolutionary biology, particularly regarding how selective pressures at different biological scales interact and sometimes conflict. The insights gained from studying HIV's rapid evolution may help understand similar dynamics in other host-pathogen systems, especially those involving chronic infections and strong immune selection.

2 Methods and Data

2.1 Dataset: Partners in Prevention Studies

All within-host data analysed in this thesis was collected as part of one of three studies of serodifferent heterosexual couples in East and Southern Africa between 2005 and 2013. The studies were run by the International Clinical Research Center (ICRC) at the University of Washington, who contributed samples to the PANGEA consortium, who performed the sequencing. As well as viral samples, metadata capturing clinical, behavioural and demographic data was also collected. None of the individuals were on antiviral treatment at enrolment, with treatment initiated according to treatment guidelines at the time. In my analyses, no samples taken after the commencement of treatment were included. All participants gave written informed consent for the storage of samples and future HIV genetic studies. The study was registered with ClinicalTrials.gov and approved by the University of Washington Human Subjects Review Committee, as well as the ethics review committees at any institutions collaborating with individual study sites.

Three studies provided data for this analysis. In all studies, HIV-positive individuals were followed quarterly with plasma sampling. The **Partners in Prevention HSV-HIV transmission study** (2004-2008) (Lingappa *et al.*, 2009; Celum *et al.*, 2010) was a clinical trial of HSV-2 suppressive therapy for HIV prevention, and which also tested HIV-negative participants quarterly. The **Partners PrEP study** (2008-2013) (Mujugira *et al.*, 2011; Baeten *et al.*, 2012) evaluated preexposure prophylaxis efficacy, and tested HIV-negative individuals monthly, switching to quarterly sampling if they seroconverted. The **Couples Observational Study** (COS) (2008-2010) (Lingappa *et al.*, 2011) followed serodifferent couples quarterly for up to 12 months.

Across the three studies, a total of 9,042 couples were longitudinally followed for 1-3 years, with collection of clinical, behaviour and demographic information as well as samples including blood plasma. Across the three cohorts, HIV sequencing was carried out on plasma from 2,648 couples sampled to include both partners in HIV seroconverting pairs (linked by plasma HIV sequence [see below]), and People Living With HIV (PLWH) in non-transmitting serodifferent couples. Hereafter, individuals who were seropositive at the commencement of the study are termed ‘source’ individuals (N=2,648), irrespective of whether they transmitted the virus to their partner, and

seroconverters are termed ‘recipients’ (N=278). Recipients were sampled early in their infection, usually between 3 to 18 months post-infection; however, the time since infection for source individuals is not known. Among couples where HIV seroconversion was identified, targeted regions of plasma HIV genome sequence (regions of GAG and ENV for Partners in Prevention HSV/HIV Transmission Study and COS, while regions of GAG and POL were sequenced in the Partners PrEP study) were compared between source and recipient to determine epidemiological linkage. Sample preparation, sequencing and processing. Sequencing of regions to determine linkage was performed at the University of Washington as a part of the original studies run by ICRC, and are separate to the sequences analysed in this thesis.

Plasma samples were prepared, sequenced and processed by colleagues at the University of Oxford, initially with Illumina sequencing which produces short reads (~250bps), and later samples were re-sequenced (or sequenced for the first time) on the PacBio sequencing platform, which generated longer reads (up to ~ 1500bps).

Samples were processed using the veSEQ-HIV protocol (Bonsall *et al.*, 2020) which has been validated for drug resistance testing on the Illumina platform (Jenkins *et al.*, 2023). Briefly, total RNA was extracted from plasma, and DNA sequencing libraries were prepared with the SMARTer Stranded Total RNA-Seq – Pico-Input Mammalian Kit v2. Unfragmented RNA was reverse transcribed using adapter-linked random hexamers, followed by conversion into double-stranded dual-indexed DNA libraries with 12 PCR cycles. Libraries were pooled and size-fractionated with AMPure-XP beads, adjusting the Polyethylene glycol (PEG) ratios to produce a low molecular weight (LMW) and high molecular weight (HMW) fraction for sequencing. For Illumina sequencing the cutoffs for LMW and HMW material was 0.5 and 0.68 PEG:pool *(v:v)* , and for PacBio cut offs were 0.68 and 0.8 PEG : sample pool *(v:v).* The HMW and LMW fractions were re-pooled at a molar ratio of 1:9 for subsequent HIV enrichment. Custom HIV-specific biotinylated probes captured HIV DNA fragments using the IDT hybridisation wash kit. Following capture, libraries were amplified for sequencing with 12 PCR cycles, ensuring sufficient yield while minimizing amplification bias. Where stated, captured pooled libraries were sequenced on Illumina NovaSeq using the P2 500 cycle flow cell. For PacBio sequencing, captured material was ligated to SMRT-bell PacBio adapters and loaded onto a Sequel II SMRT Cell (8M) for sequencing on a PacBio Sequel IIe instrument.

Illumina reads were processed with Kraken to remove human and bacterial reads, followed by trimming with Trimmomatic. Contigs were assembled using SPAdes and metaSPAdes and clustered with cd-hit-est. Reads were mapped to sample-specific references, and consensus sequences were generated using Shiver (Wymant, Blanquart, *et al.*, 2018). For samples without assembled contigs, reads were mapped to the closest matching genome from the Los Alamos HIV database (http://www.hiv.lanl.gov). The same pipeline was adapted for PacBio HiFi reads, which were demultiplexed using PacBio LIMA, and a dummy reverse complement-read was synthesized to mimic paired-end data for shiver. In total, 576 samples were sequenced with the PacBio pipeline for the purposes of this study, with additional samples previously sequences for previous studies.

I did not perform sample preparation, sequencing or processing, and this work was performed by colleagues at the University of Oxford. For the sequence analysis aspects of this thesis, I performed analysis on the data outputted by the Shiver pipeline, where the Shiver processing was also performed by colleagues at the University of Oxford.

2.1.1 HLA dataset

HLA genotypes were generated at the Northwest Genomics Center at the University of Washington. DNA samples were purified from whole blood samples and genotyping was performed using Human OmniExpress BeadiChips. Genotypes were generated for individuals enrolled in the PrEP study. HLA genotyping was performed by researchers at the University of Washington and four-digit HLA A B and C were kindly provided to me for analysis in chapter 6.

2.1.2 Analysis datasets

For each chapter of this thesis, I analyse different subsets of the data due to the data requirements for the research questions. The data subset of each chapter is described in the following flowchart, where the broadest filtering steps are outlined, with more detailed filtering steps outlined in the methods section of each chapter, such as removal of dual infected individuals, time since infection and read depth. In each chapter, metadata is also incorporated, including sex, subtype and viral load.

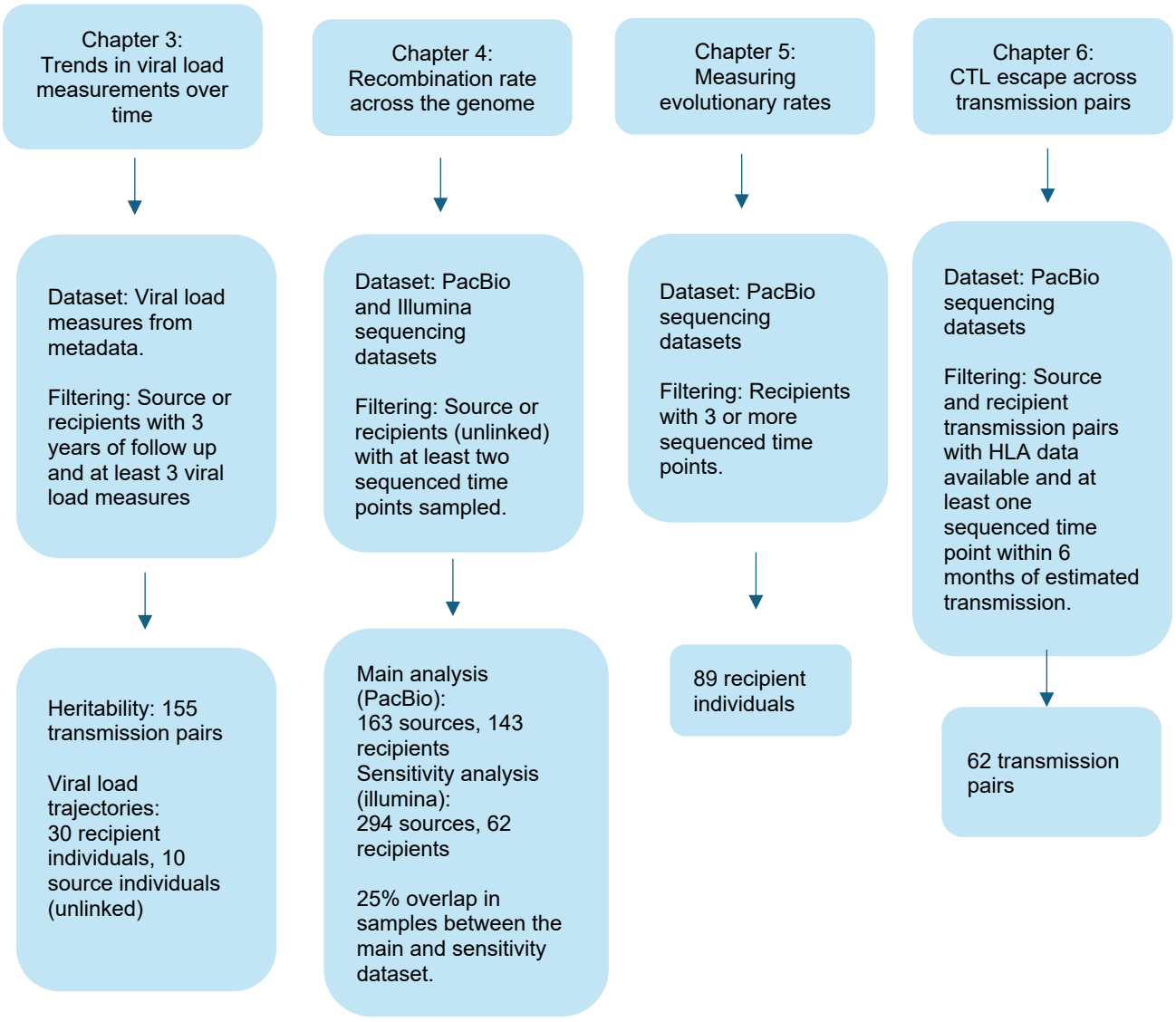

Figure 2.1 Dataset selection criteria for the analyses of each chapter

2.1.3 Estimated date of seroconversion

The estimated date of seroconversion for all recipients was determined as the calendar midpoint between the last HIV-negative test and the first HIV-positive test. While this approach carries a considerable degree of uncertainty, alternative methods, such as those based on genetic diversity, yield a similar level of imprecision, and Fiebig staging data were unavailable. Notably, in the case of the PrEP study, which accounts for more than half of the individuals, HIV testing was conducted monthly rather than quarterly, thereby reducing the margin of error in the seroconversion date estimation.

2.1.4 Set-point viral loads

The SPVL was measured by taking all viral load measurements, excluding any measurements taken within the first 6 months of the estimated date of seroconversion. If the earliest measure was an order of magnitude greater than the median value across all samples, the infection was assumed to be in the acute stage and that measurement was removed from the overall estimate.

2.1.5 Subtypes

The HIV-1 subtype was determined by the subtype of the best reference selected by Shiver during the sequence processing stage. The set of references is that provided in the Phyloscanner software and includes recombinant forms, however novel or rare recombinant forms are not included, leading to potential misclassification. If multiple subtypes were assigned to the same infection across multiple sampling points, the infection was classified as a recombinant form. The most common subtype in the complete dataset was A1, with over half of infections determined to be A1 subtype, followed by D at approximately 18% and C at approximately 10%. The remainder of the dataset was composed of mainly circulating recombinant forms.

2.1.6 Data availability

The data presented and analysed in this thesis were collected from HIV sero-discordant couples in sub-Saharan Africa, presenting significant ethical and legal challenges in data sharing due to the continued stigma surrounding HIV infection and the criminalisation of HIV transmission in certain countries. The sensitive nature of the

data requires careful consideration of their potential impact on participants, and access to study data must be in accordance with participant consent at the time of the study.

Advances in next-generation sequencing technology now enable precise determination of transmission networks, which further complicates the public sharing of such sensitive information. As a result, it is not feasible to deposit the full dataset in open-access databases to protect the privacy and safety of the research participants.

However, the metadata and raw sequencing data can be obtained for verification purposes. Researchers may apply to access the data by becoming accredited members of the PANGEA consortium. Detailed information about the application process can be found at the consortium website https://www.pangea-hiv.org/join-us. Upon attaining access, researchers can apply the specific processing and filtering steps used in the analysis of each chapter.

3 Attenuation of HIV severity by slightly deleterious mutations can explain the long-term trajectory of virulence evolution.

3.1 Abstract:

HIV-1 is a well-studied example of a pathogen that has evolved an intermediate level of virulence that maximises transmission. For a trait to evolve it must be heritable, and although viral load—a proxy for disease severity—has been shown to be a heritable trait, it is surprising that specific heritable viral factors remain mostly elusive. Rapid within-host evolution is also expected to diminish heritability. We hypothesised that rather than a small number of mutations of large effect determining viral load, the number of slightly deleterious mutations could be key. As a proof of principle, I explored how viral load is expected to evolve within and between hosts using a nested modelling approach that links within-host evolution with epidemiological outcomes. For mutations of sufficiently small effect, a mutation-selection balance is gradually reached during infection, resulting in slow changes in viral load despite rapid rates of genomic evolution. By incorporating host heterogeneity, I generated realistic population distributions of viral loads and estimates of heritability. The existence of many slightly deleterious mutations provides a mechanism that can help to explain why viral loads change slowly during infection, broad distributions of viral loads among individuals, and why searches for viral factors that determine viral load have had limited success.

3.2 Introduction

HIV-1 has become a prominent example of a pathogen that has evolved towards intermediate virulence as a result of the trade-off between virulence—defined here as disease severity—and transmission (Fraser *et al.*, 2007). The trade-off hypothesis (Anderson and May, 1982; Ewald, 1983) is based on the idea that the duration of an infection and the virulence of the infecting pathogen are inversely correlated. If the relationship between virulence and the number of expected onward transmissions saturates, the transmission fitness of the pathogen will be maximised at a finite level of virulence (Alizon *et al.*, 2009). If in addition virulence is heritable, meaning that differences in virulence between individuals is partly determined by genetic differences

in the infecting pathogen, the pathogen is expected to evolve towards the level of virulence that maximises the number of onward transmissions.

Demonstrating the existence of evolutionary trade-offs of virulence for pathogens in real-world systems is notoriously difficult (Alizon *et al.*, 2009; Cressler *et al.*, 2016), and HIV-1 is one of the few systems where a transmission-virulence trade-off has been shown (Fraser *et al.*, 2007). During untreated chronic infection, HIV-1 viral loads typically remain at a relatively steady level, known as the set-point viral load (spVL). This is a striking observation given that spVLs are extremely heterogeneous when measured among individuals, varying by over four orders of magnitude (Mellors *et al.*, 1996). Because untreated individuals with higher viral loads tend to progress to AIDS and death sooner than those with lower viral loads (Mellors *et al.*, 1996; De Wolf *et al.*, 1997), spVL is the most commonly used proxy for HIV-1 virulence. In addition, viral load has been shown to be correlated with infectiousness, with high viral load individuals much more likely to transmit the virus per contact (Quinn *et al.*, 2000; Fideli *et al.*, 2001). The relationship between the expected number of onward transmissions and the duration of infection was shown to saturate (Fraser *et al.*, 2007), and moreover that observed viral loads cluster around the values that are expected to maximise transmission, thus supporting the trade-off hypothesis. A follow up study found further support for the trade-off hypothesis from a large HIV-1 prospective cohort and argued that the attenuation of viral loads in the cohort over two decades was the result of between-host adaptation towards maximising transmission potential (Blanquart *et al.*, 2016).

For a trait to evolve under natural selection it must be heritable. Although the study of heritability has its roots in animal breeding, the same concept can also be applied to pathogen virulence, i.e. for measuring the extent to which severity of infection is determined by the genotype of the infecting pathogen. Multiple studies have provided support for HIV-1 spVL being a heritable trait, with analyses proposing estimates of broad sense heritability ranging between 20-30% (Alizon *et al.*, 2010; Hollingsworth *et al.*, 2010; Fraser *et al.*, 2014; Bonhoeffer, Fraser and Leventhal, 2015; Blanquart *et al.*, 2017; Mitov and Stadler, 2018). Lower estimates have been proposed (8%, Hodcroft et al 2014), however a 2017 study (Blanquart *et al.*, 2017) argued that these lower estimates resulted from the inappropriate assumption of a Brownian motion

model for spVL evolution. When the same dataset was reanalysed using an Ornstein-Uhlenbeck model, the resulting estimates fell within the 20–30% range.

The source of the heritability of HIV-1 spVL remains for the most part unexplained. A genome-wide association study (GWAS) failed to identify viral factors of statistical significance (Bartha *et al.*, 2013), however this study was only powered to detect effects of 4% or greater of heritability, suggesting that such factors have small individual effect. More recently, a GWAS study had four hits of viral mutations linked to viral load (Gabrielaite *et al.*, 2021), however the narrow-sense heritability explained by these mutations is substantially lower than the estimates of broad-sense heritability.

Host-factors have also been attributed to variation in viral loads (McLaren *et al.*, 2015); however, a 2017 study found the fraction of variation explained by human genetic factors to be relatively low at 8.4% once viral genetic diversity is accounted for (Bartha *et al.*, 2017). Some human polymorphisms, such as the 32 base pair deletion in the CCR5 gene, have been found to significantly reduce HIV virulence (Liu *et al.*, 1996; Martin, 1998). Virus-host interactions also likely affect virulence, for example the changes in immune escape pressure following transmission to a new host environment (Van Dorp, Van Boven and De Boer, 2014). There is also some evidence that a proportion of heritability can be explained by accounting for similarity in HLA profiles within transmission pairs, with adaptation to HLA alleles in the source increasing viral load, leading to transmitted adaptation and consequently higher viral loads in the recipient (Carlson *et al.*, 2016).

Another difficulty in understanding the heritability of HIV-1 spVL is rapid within-host evolution and often long delays between infection and onward transmission. Rapid diversification and evolution is likely to reduce heritability as the viral genotype an individual is infected with is expected to differ from the genotype that they go on to transmit (Lythgoe and Fraser, 2012; Fraser *et al.*, 2014). For example, we may expect viral variants with high replicative capacity to emerge and quickly sweep through the virus population, and if replicative capacity affects disease factors such as virulence, then the heritability of these factors will be reduced.

Viral replicative capacity has been shown to be associated with viral load (Kouyos *et al.*, 2011; Prince *et al.*, 2012; Selhorst *et al.*, 2017). Even fairly small differences in the fitness of virus variants are expected to result in rapid within-host evolution,

notwithstanding complications of whether the virus is pre-adapted to a recipient's HLA allele type (Payne *et al.*, 2014; Carlson *et al.*, 2016). Hence if there were a strong link between replicative capacity and viral load, we would expect viral loads to increase greatly during infection, low estimates of heritability, and the evolution of high virulence even at a cost to overall transmission ("short-sighted evolution") (Lythgoe, Pellis and Fraser, 2013; Van Dorp, Van Boven and De Boer, 2014). Yet viral loads only change modestly during chronic infection (Henrard, 1995; Troyer *et al.*, 2005), heritability of viral load is remarkably high, and ever-increasing viral loads has not been a feature of the HIV-1 epidemic. Possible mechanisms for how virulence has evolved under the constraints of both within and between host selection include a complex fitness landscape, and the preferential transmission of ancestral strains (Fraser *et al.*, 2014). Underpinning any mechanism are the elusive heritable viral factors.

I propose that rather than being controlled by few high-impact mutations (as usually targeted by whole-genome association studies), viral load is largely determined by the number of slightly deleterious mutations a genome has. In such a scenario a mutation-selection balance, in which mutations continually arise through mutation and are slowly lost through purifying selection, is gradually approached during the course of infection. Because this process is inherently slow, it can reconcile seemingly incompatible aspects of the within- and between-host processes, namely the stability of viral load during chronic infection, the high heritability of viral load, and the selection of intermediate viral loads at the population level. Moreover, it may explain why HIV-1 virulence factors have been so hard to identify. As a proof of concept, I considered the impact of many slightly deleterious mutations on the within and between-host evolutionary dynamics of the virus within treatment-naïve individuals using a nested modelling approach. Finally, I tested a prediction of our model on real viral load trajectories over three years of untreated infection, finding evolution towards an intermediate viral load within-host over the course of an infection.

3.3 Methods

I modelled the evolving population of viral genotypes during infection using a quasispecies framework, with mutations occurring at a frequency of $\mu = 3 \times 10^{-5}$ per site per generation (Mansky, 1996a) at *m* segregating sites. I assume each segregating site in the viral genome has two potential alleles, wild-type or slightly deleterious, with 'virus types' defined by the number of deleterious mutations in the

viral genome, regardless of the position of those mutations, and hence viruses with different genotypes can have the same virus type. Individuals are assumed to be initially infected by a single virus type, which defines the infection type: an individual who was infected with virus type *j* will be of infection type *j.* After infection, the dual forces of a high mutation rate and weak selection result in the emergence of a diverse within-host viral population and the maintenance of deleterious mutations, with the population reaching mutation-selection balance. At time *t* during a type *j* infection, a virus type *i* has a frequency $x_{ij}(t)$ that is determined by the quasispecies equation (see next sub-section). I assumed that the viral load of an infection at time *t* is determined by the number of deleterious mutations in the within-host viral population at time *t*. For the epidemiological (between-host) modelling, I use a nested modelling framework from Lythgoe *et al*. (Lythgoe, Pellis and Fraser, 2013) in which the within-host dynamics are 'nested' within a between-host epidemiological model, which in turn is based upon the theory of multi-type epidemic models (Diekmann and Heesterbeek, 2000), and the course of an individual's infection is entirely determined by the within-host model.

3.3.1 Within-host dynamics

I assumed the viral genome has a maximum of *m* segregating sites, and a virus type is defined by the number of deleterious mutations, *j,* regardless of the position of the mutation. The generation time of an infected cell was assumed to be one day, and each newly infected cell can acquire or lose at most one mutation relative to the infecting virion. The *m x m* reproduction matrix $Q = q_{ij}$ describes the probability that a progeny of virus type *j* is of virus type *i*, where q_{ij} is only non-zero when *i* and *j* differ by 0 or 1. I considered the competing virus types within a host as a quasispecies for which the dynamics are governed by the below quasispecies equation, and the relative fitness of virus type *j* is a multiplicative function of the number of deleterious mutations: $A_j = (1-s)^j$ where *s* is the selection coefficient and $A_0 = 1$. Here, viral fitness is synonymous with the replicative capacity of the virus type. The change over time in the relative frequencies of the virus types can be described by the quasispecies equation (Wilke, 2005) :

$$\frac{\boldsymbol{dx}}{\boldsymbol{dt}} = \boldsymbol{Wx} - \bar{\boldsymbol{w}}\boldsymbol{x}$$

where $W = w_{ij} = q_{ij}A_j$ and the term $\bar{w} = \Sigma_i^m \Sigma_j^m w_{ij} x_{ij}(t)$ bounds the sum of the frequencies to one. The solution at equilibrium, $\tilde{x}$, can be found analytically as the dominant eigenvector of *W* (Wilke, 2005). The pre-equilibrium analytical solution is computationally expensive to calculate when *m* is large, therefore I solved the system numerically with the *deSolve* package in R v. 4.0.2.

The higher the fixed number of segregating sites in the model, the smaller the relative fitness cost of one additional mutation. I explored a range of fitness costs, from 10^{-2} per mutation when 10 sites are segregating to 5×10^{-5} per mutation when 250 sites are segregating, spanning strong selection to close to neutral effects approaching the mutation rate (Zanini *et al.*, 2017). I capped the maximum number of segregating sites, *m*, at 250 because the matrices needed to track all possible combinations of viral variants become too large to handle - they grow exponentially with each additional site added to the model - however the range of values I considered effectively captures the key dynamics of the model.

3.3.2 Viral load

There are three distinct stages of an HIV infection. The first and third stages of infection, termed acute and late infection respectively, are characterised by high viral loads. The second stage is chronic infection, when viral loads are relatively stable yet vary significantly between individuals. I assumed that during chronic infection the viral load at a given time is determined by the current composition of the viral population. As a consequence, the viral load can change as the viral population evolves: the fewer the number of deleterious mutations the higher the viral load. At any moment in time the viral population is likely to be comprised of multiple types, so I first defined the contribution of each viral type, V_i, to the viral load during chronic infection:

$$V_i = V_0 - \lambda_m i$$

where $V_0 = 7 \log_{10}$(viral copies per ml), *i* is the number of mutations, and λ_m is the reduction in viral load due to an additional mutation, which is in turn determined by the maximum number of mutations, *m*. The maximum possible viral load, $7 \log_{10}$(viral copies per ml), was chosen to match the highest viral loads typically observed (Fraser *et al.*, 2014). The range was further verified in our dataset of spVLs.

When I assumed a large number of mutations, I assumed that the fitness cost of a single mutation is small, and therefore the reduction in viral load given one additional

mutation is also small. Conversely, if there are few mutations of large effect, I assumed viral load will reduce significantly when a deleterious mutation appears. The value of λ_m is chosen such that the difference in the viral load between the fittest and weakest virus is $5\log_{10}$(viral copies per ml) to correspond with a realistic range of viral loads, and so $\lambda_m = \frac{5}{m}$. In determining viral load in this way, there is a linear relationship between the viral load - $\log_{10}$(viral copies per ml) - of a variant and the relative fitness of the variant. I then determined the realised viral load to be the mean of these contributions, so that viral load of a type *j* infection at time *t* into infection is given by:

$$V_j(t) = \Sigma_{i=0}^{m} x_{ij}(t) V_i$$

3.3.3 Duration of chronic infection

I used the previously parameterised decreasing Hill function for the duration of chronic infection as a function of spVL (Fraser *et al.*, 2007). To account for the change in viral load over time in the calculation of the duration of infection, I took a weighted average of the viral load over the maximum possible duration of chronic infection, determined to be 20.4 years by the Hill function. The weight of $V_j(t)$ in the calculation of $\bar{V}_j$ is inversely proportion to *t,* such that the viral load close to the start of chronic infection contributes more to the calculation than the viral load several years into infection. The weighted average viral load is then used as an input to the decreasing hill function to provide a chronic infection duration, termed T_j. To attribute a spVL, V_j, to an infection we take the arithmetic mean viral load over T_j.

3.3.4 Infectivity profile

In order to determine how evolution proceeds at the between-host level, we need to know not only the frequency of different virus types during infection, but also their probability of transmission. I assumed a single virus type is transmitted during a transmission event, and the hazard for transmitting virus type *i* in a type *j* infection at time *t* is:

$$\beta_{ij}(t) = x_{ij}(t)\gamma_j(t)\,, \qquad t < T_j$$

$$\beta_{ij}(t) = 0 \qquad \text{Otherwise}$$

Where $\gamma_j(t)$ is the infectiousness of a virus type *j* infection at time *t* and T_j is the duration of a virus type *j* infection. For all infections I assume the acute phase has a

duration of 0.25 years and an infectiousness of 2.76 onward infections/year, and the late phase has a duration of 0.75 years and an infectiousness of 0.76 onward infections/year (Fraser *et al.*, 2007). During chronic infection, the infectiousness $\gamma_j(t)$ is described by an increasing Hill function of viral load, *V(t),* at time *t* with parameter values based upon the optimal values detailed in (Fraser *et al.*, 2007).

Table 3.1 Variable and parameters descriptions for the within-host and between host models.

Parameters	Definition	Value
$\boldsymbol{\mu}$	Mutation rate	3×10^{-5} (Mansky, 1996a)
$\boldsymbol{m}$	The number of segregating sites, and therefore the maximum number of deleterious mutations.	10, 50, 100, 150, 200, 250
$\boldsymbol{s_m}$	The relative fitness cost of an additional mutation ranging from $\boldsymbol{s_{10}} = 10^{-2}$ to $\boldsymbol{s_{250}} = 5 \times 10^{-5}$ to capture a realistic distribution of selection coefficients. The values between m=10 and m=250 were determined by fitting a straight line between s_{10} and $\boldsymbol{s_{250}}$ on a log-linear scale to scale selection strength with the number of segregating sites.	$s_{10} = 10^{-2}$ $s_{50} = 10^{-2.4}$ $s_{100} = 10^{-2.9}$ $s_{150} = 10^{-3.4}$ $s_{200} = 10^{-3.8}$ $s_{250} = 10^{-4.3}$
$\boldsymbol{A_j}$	Replication rate of a type *j* virus variant.	$A_j = (1-s)^j$
$\boldsymbol{\lambda_m}$	The cost of an additional mutation on the associated viral load. It is given by $\lambda_m = \frac{5}{m}$ The value is determined such that the minimum viral load is 2 $\log_{10}$(viral copies per ml).	$\lambda_{10} = 0.5$ $\lambda_{50} = 0.1$ $\lambda_{100} = 0.05$ $\lambda_{150} = 0.033$ $\lambda_{200} = 0.025$ $\lambda_{250} = 0.02$
$\boldsymbol{T_j}$	The duration of a type-*j* infection.	Variable
$\boldsymbol{\gamma_j(t)}$	Infectivity of a type-*j* infection at time *t* into infection.	Variable and a function of $V_j(t)$. Parameterised in [1].
$\boldsymbol{V_i^s}$	The viral load associated with a type-*i* virus for a given fitness cost.	$7 \log_{10}(\text{copies per ml}) - \lambda_s i$
$\boldsymbol{B}$	Rate at which individuals enter the susceptible population.	200 per year (Lythgoe, Pellis and Fraser, 2013)
$\boldsymbol{D}$	Natural mortality rate.	0.02 per year (Lythgoe, Pellis and Fraser, 2013)
$\boldsymbol{h}$	Host type.	1-50
$\boldsymbol{e}$	Maximum absolute host effect on viral load on $\log_{10}$ scale	0.1, 0.25, 0.5, 0.75, 1
$\boldsymbol{E_h}$	Additive host effect size on $\log_{10}$ viral load.	Discretely uniformly distributed over the range [-e, e] in the host population.
Variables		
$\boldsymbol{x_{ij}(t)}$	The frequency of a type-*i* virus at time *t* of a type-*j* infection.	
$\boldsymbol{V_j(t)}$	Viral load at time t of a type-*j* infection	
$\boldsymbol{\beta_{ij}(t)}$	The infectivity profile i.e. the rate at which a type-*i* virus is transmitted at time t of a type-*j* infection.	
$\boldsymbol{K}$	The next generation matrix.	
$\boldsymbol{I_i(t)}$, *I**	The number of infected individuals infected with a type-*i* virus at time *t* into the epidemic ($I_i(t)$) and at the endemic equilibrium (*I**).	Initially 1

*S(t), S**	The number of susceptible individuals at time *t* into the epidemic (*S(t)*) and at the endemic equilibrium (*S**).	Initially 10,000
*N(t), N**	The number of humans hosts at time *t* into the epidemic (*N(t),*) and at the endemic equilibrium (*N**).	Initially 10,000
h^2	Broad-sense heritability estimate.	

3.3.5 Between-host model

I used an SI model with demography to model between-host transmission, where I assumed a natural death rate *D* and that individuals enter the susceptible pool of individuals at a rate of *B*. The force of infection of virus type *i* at time τ of an infection founded by virus type *j* is defined as $\beta_{ij}(\tau)e^{-D\tau}$ when $\tau \leq T_j$ and 0 otherwise. The between-host dynamics are modelled by the renewal equation, as described in Lythgoe *et al.* (Lythgoe, Pellis and Fraser, 2013). Specifically, incidence is calculated by the renewal equation which follows the logic that incidence at time *t* is the integral of the past incidences weighted by how much the individuals previously infected would still be transmitting. The between-host dynamics at time *t* since the beginning of the epidemic are therefore described as follows:

$$I_i(t) = \frac{S(t)}{N(t)} \sum_{j=1}^{m} \int_0^{T_j} \beta_{ij} I_j(t-\tau) e^{-D\tau} \, d\tau$$

$$S(t) = N(t) - \sum_{i=1}^{m} \int_0^{T_j} I_j(t-\tau) e^{-D\tau} \, d\tau$$

$$\frac{dN(t)}{dt} = B - DN(t) - \sum_{i=1}^{m} I_i(t-T_i) e^{-DT_i}$$

Where *N* is the total population size, *S* is the number of susceptible individuals and $\boldsymbol{I_i}$ is the incidence of infections founded by a type *i* virus. All individuals have the same natural death hazard, *D*, and infected individuals also have an infinite death hazard at the moment their infection ends, which is pre-determined by their infection type.

The epidemic begins with an incidence of 1. The solutions were determined numerically using the basic forward Euler method. I ran 3 simulations of the between-host dynamics, where the initial circulating virus type and therefore average viral load

differed for each simulation. Epidemiological theory shows that the next generation matrix *K*, with *ij*th element $k_{ij} = \int_0^{T_j} \beta_{ij}(\tau) e^{-\mu\tau} d\tau$, has a unique and real dominant eigenvalue that gives the value of the basic reproduction number R_0, and the associated eigenvector describes the population structure of the virus types at equilibrium (Lythgoe, Pellis and Fraser, 2013), which can be normalised to give the proportions of each virus type at their equilibrium state, I_j^*. The transmission potential of an infection by a type *j* virus is defined as the number of expected onward transmissions during infection and is determined from the next generation matrix as $p = \Sigma_i k_{ij}$, where the term $k_{ij}$is the expected number of transmissions of a type *i* virus during an infection of a type *j* virus.

3.3.6 Host heterogeneity

As well as considering a population of genetically identical host individuals, I also examined the effect of host heterogeneity. The host effect is quantified by an additive effect to the viral load, such that for a type *j* infection in a host type *h*, the viral load at time *t* is given by $V_j^h(t) = V_j(t) + E_h$. The parameter *E* is discretely uniformly distributed, $E \sim \mathcal{U}(-e, e)$ and there are 50 possible host types. As a result, the probability of transmitting virus type *j* at time *t* differs by host type, however the within-host dynamics are identical for all infections. I considered a range of values of *e* from 0.1 to 1 in order to quantify the impact of host specific viral load effects on model outcomes. I present the analytical equilibrium solution for populations with host-heterogeneity as it was not feasible to derive pre-equilibrium solutions numerically due to computational constraints.

3.3.7 Heritability

To estimate broad-sense heritability, h^2 (heritability hereafter for brevity), we use the classic parent-offspring regression method (Wright, 1920). To apply the method to a simulated infection population, I generated 1000 samples of 500 transmission pairs, where the source infection type was sampled based upon the population distribution of infection types at the endemic steady state. The probability of a recipient having an infection of a type *i* virus from a source infected with type *j* virus was described by the vector of transmission potentials, k_{ij}, normalised to sum to 1. For each set of transmission pairs, an estimate of h^2 was given by the regression coefficient of a simple linear regression with recipient spVL as the outcome. The overall estimate of

heritability is calculated by the average h^2 across the sets of transmission pairs. For a non-homogeneous host population, the host type of the source and recipient was also sampled from a uniform distribution of host types.

I calculated the broad-sense heritability in spVL in our dataset of transmission pairs by performing a linear regression between the log-transformed source spVL and log-transformed recipient spVL. The spVL was calculated for every individual as described in Chapter 2.

3.3.8 Viral load analysis in study cohort

From the main analysis dataset of serodifferent couples, I identified all source and recipients who were followed up for over 1,000 days with at least 3 viral load measurements. Due to accuracy in measuring low viral loads, any values lower than 1000 viral copies per ml were removed. I first performed a within-host regression for the change in log-transformed viral load over time. With the output of each regression, I then performed a between-host analysis to assess whether there was an association between initial viral load and the direction of change. The slope of the within-host regression was regressed against the initial viral load measure of the individual.

As a regression towards the mean can be observed as a result of noise, I performed a correlation test. I randomly sampled 100 viral load measures to represent 100 distinct infections, and added realistic noise to two or three additional measurements, with sampling times spread evenly between 100 and 1000 days, and then performed the same analysis as on the observed data (within-host slope regressed against within-host first measure). Noise for each measurement was randomly sampled from a normal distribution with a mean of 0 and a standard deviation of 0.3, to correspond with reported error in viral load measures (Bonsall *et al.*, 2020). Log-transformed viral load measures were sampled from a normal distribution to match our observed spVL distribution, with an average of 4.5 $\log_{10}$(viral copies per ml), and standard deviation of 1. I repeated this simulation 1000 times to generate a distribution of regression slopes to compare with our observed measure.

3.4 Results

To summarise our approach, I modelled the within-host dynamics with a quasispecies model that describes the changing frequencies of competing virus types within a host virus population, where each virus type is defined by its number of deleterious mutations. The relative fitness of a virus type is a function of the number of deleterious mutations, and the conflicting forces of a high mutation rate and purifying selection results in a mutation-selection balance of deleterious mutations if the fitness cost is sufficiently low. The viral load at time *t* is determined by the average fitness of the virus population at time *t,* where high fitness induces a high viral load. The within-host model is nested in a between-host model that can describe the population-level distributions of viral loads and the amount of heritability for different parameter choices.

3.4.1 Many mutations with low fitness cost results in intermediate virulence evolving within individuals

I determined the within-host equilibrium solution for a range of values for the number of segregating sites, *m*, and calculated the corresponding viral load (Figure 3.1). We find that for a relatively low number of sites (fewer than approximately 100), selection of the fittest virus types dominates, and viral load is at the maximum possible value. As the number of sites increases, the selection cost decreases, and we approach a mutation-selection balance. Consequently, the number of deleterious mutations in the population is maintained and the viral load is reduced. Eventually, the cost of an additional mutation passes a threshold where the relative fitness cost is effectively neutral and the equilibrium viral load approaches the midpoint of the viral load range of $4.5\log_{10}$(viral copies per ml), and the genetic variation is maintained through mutation.

With more segregating sites, the mutation-selection balance maintains a more diverse quasispecies, where multiple viral types coexist at equilibrium rather than a single dominant type (supp. Figure 9.1). Higher diversity in the within-host population creates a broader pool of variants available for transmission, which influences the trajectory of between-host evolution.

By adjusting how selection strength scales with the number of segregating sites, we could alter the model outcomes. If strong selection (high *s*) was maintained across more sites, equilibrium viral loads would be higher because deleterious mutations would be more efficiently removed. Conversely, if selection became weak at a lower

number of sites, more deleterious mutations would accumulate, resulting in lower equilibrium viral load. While shifting the relationship between selection strength and number of sites would alter the specific equilibrium values, the qualitative behaviour of the system - and thus our main conclusions - would remain unchanged.

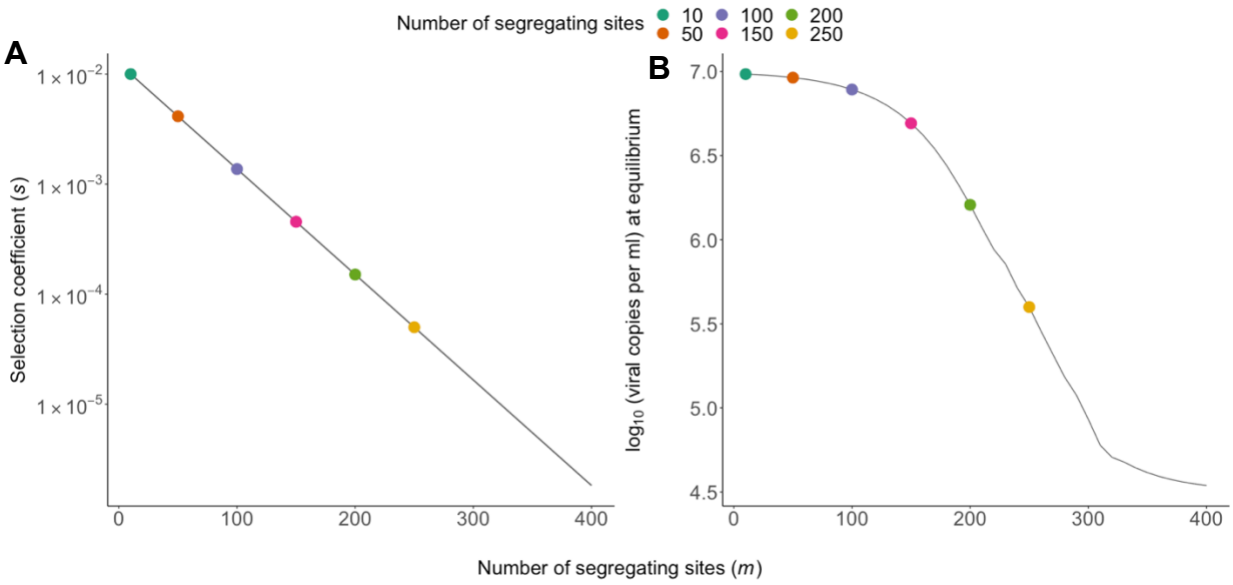

Figure 3.1 Within-host model fitness and equilibrium viral load A) The relationship between the number of segregating sites and the fitness cost (synonymous with selection coefficient) associated with an additional mutation. I explored a range of values of m, with the corresponding fitness cost of a mutation reducing on a log-linear scale as m increases. The six choices of m explored here are indicated by coloured points. B) The viral load at equilibrium of an infection with an increasing number of segregating sites. As the number of sites increase and the fitness costs decrease, the population at equilibrium is characterised by mutation-selection balance that maintains deleterious viral types in the population, thus lowering viral loads.

3.4.2 Having many mutations with low fitness cost slows the tempo of within-host evolution

To fully explore the model and its dynamics, we considered five scenarios differing by the number of segregating sites (10, 50, 100, 150, 200, 250) and the fitness cost *s* of each mutation. The viral load calculation ensures that the virus type with the maximum number of mutations *(m)* has an associated viral load at the lower tail of reported spVLs, approximately $2\ \log_{10}$(viral copies per ml), and the fittest virus has an spVL of $7\ \log_{10}$(viral copies per ml). For each of the scenarios, I determined the evolutionary dynamics of the infection for all possible starting infection types. As expected, if there are few segregating sites, each with mutations of large effect, a within-host equilibrium

was rapidly reached within months and was dominated by the fittest variants harbouring no mutations (Figure. 3.2). Increasing the number of sites to 50 and 100 slows the within-host dynamics, however all infections have reached the maximum possible viral load within 5 years and 15 years respectively, with initially low vial loads steadily increasing throughout chronic infection.

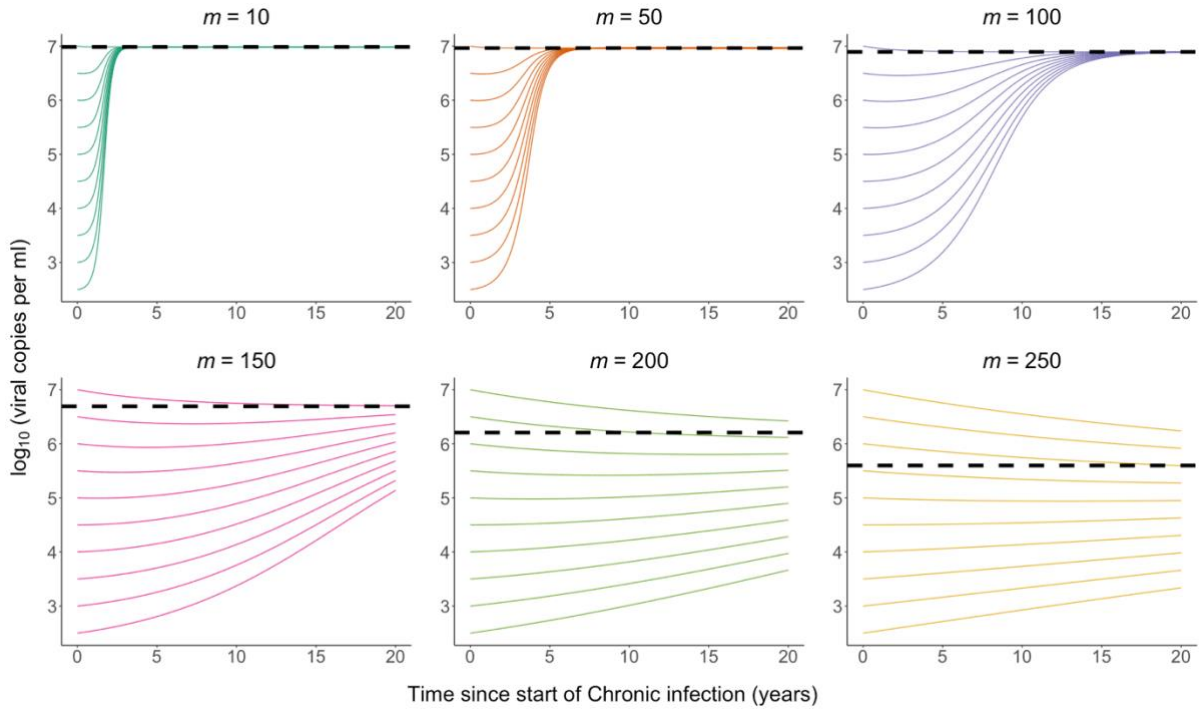

Figure 3.2 Within host viral load trajectories. The within-host viral load dynamics over time for 10, 50, 100, 150, 200 and 250 segregating sites, and varying initial numbers of mutations. The viral load of the equilibrium solution is shown by the black dashed horizontal line. The viral load changes as the virus population evolves over time, with new deleterious mutations appearing that are then lost due to purifying selection. When the cost of a mutation is high, viral loads climb to the maximum possible value. As we increase the number of segregating sites and reduce the fitness cost of a mutation, the selection against weaker virus types is balanced by the influx of mutations, and consequently the within-host dynamics are extremely slow relative to the typical duration of chronic infection.

When a large number (>100) sites are segregating and fitness costs are lower, a mutation-selection balance is reached at the within-host equilibrium state, lowering the viral load. Moreover, it took significantly longer than the assumed maximum infection duration to approach this equilibrium, depending on the number of mutations of the infecting strain. As we continue to increase the number of segregating sites from 200 to 250, the viral load dynamics do not qualitatively change within the time frame over which an infection typically occurs, and we continue to see stable viral loads.

3.4.3 Many mutations with low fitness cost leads to between-host diversity in viral loads

To determine the expected distribution of spVLs among individuals predicted by our model, I first used numerical integration to determine the within-host dynamics of all possible *j*-type infections. For each infection-type, *j*, we could then calculate the mean viral load during chronic infection (our proxy for spVL), and in addition the infectivity of *i*-type virus at time *t* since infection, $\beta_{ij}(t)$ for all *i*. Using this information we determined the distribution of *j*-type infections, and therefore spVLs, at equilibrium by using a next-generation framework.

The distribution of spVLs varies substantially as we increase the number of segregating sites (Figure 3.3A-F). As the number of segregating sites increases, the average viral load at endemic equilibrium across individuals falls to the range of previously reported average spVLs, with $5.14, 4.78,$ and 4.562 (viral copies per ml) for 150, 200 and 250 segregating sites respectively (Figure 3.3D-F). The increased number of segregating sites leads to slower within-host dynamics and more diverse quasispecies populations, enabling the virus to evolve between hosts towards an intermediate spVL that maximises transmission potential. As a result, the epidemic size increases (supp. Figure 9.2).

Whilst having many weakly deleterious mutations resulted in little variation in viral load over the course of an individual infection, the spVL during chronic infection differed by an order of magnitude between individuals. However, the variation in spVLs was still less than observed in infected populations (Fraser *et al.*, 2014). Host genetics are known to affect progression, and so we included host heterogeneity, such that the same infection type will induce different spVLs in individuals of different host type. For example, a host effect of 0.5 means that two individuals infected by the same virus type may differ in their spVL by a maximum of 1 $\log_{10}$ (viral copies per ml). I find that in the case *m*=250, spVLs at the endemic equilibrium are more realistically diverse when a host-effect is assumed, and the number of circulating virus types increases, and the between-host diversity in spVL increases with host effect as expected (Figure. 4). We find the same effect on all choices of *m* (supp. Figure 9.3).

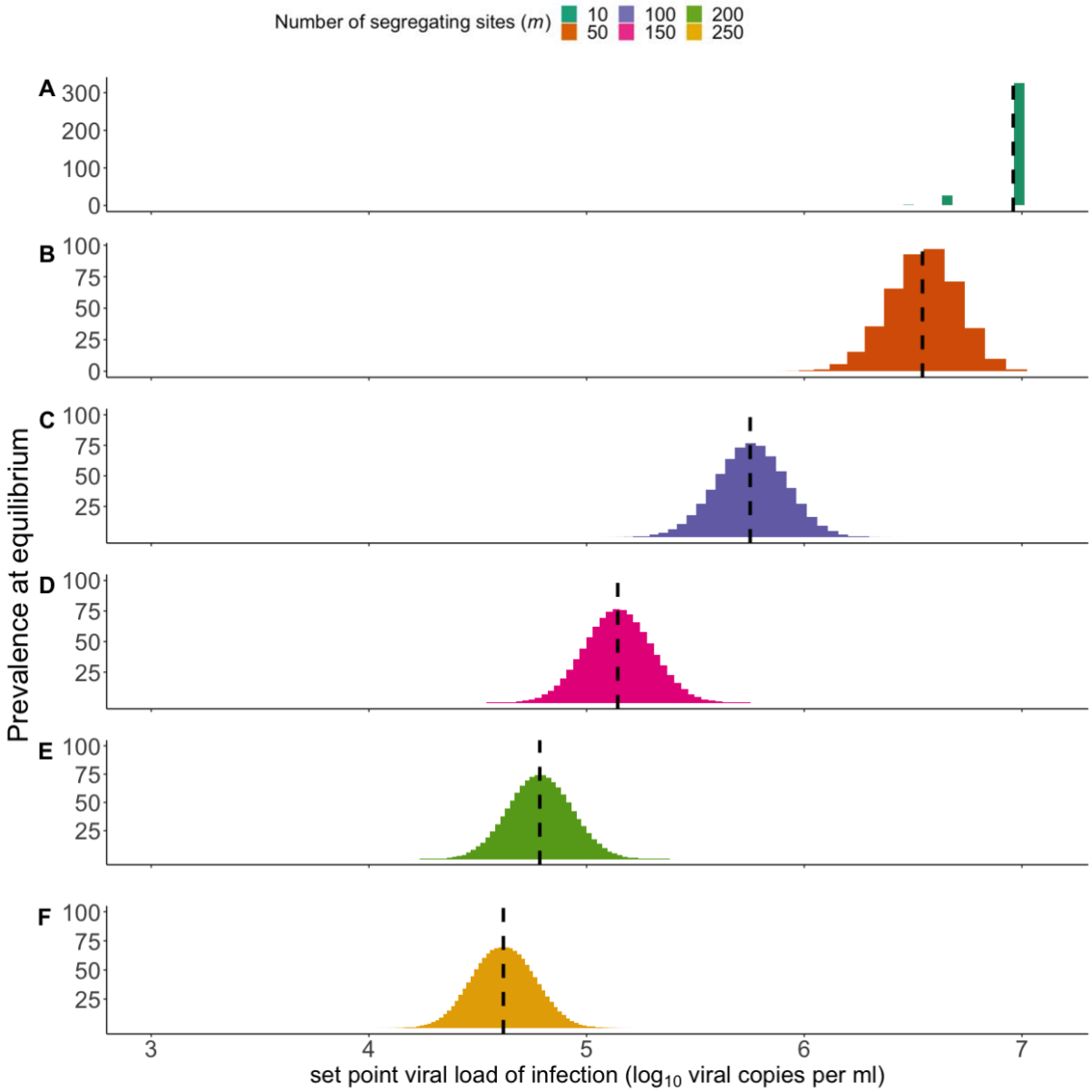

Figure 3.3 Between-host outcomes at endemic equilibrium A) The distribution of set-point viral loads (spVL) at the between-host equilibrium. The fittest virus type rapidly dominates within-host and short-sighted evolution dominates. Black dashed lines denote the average spVL. B-F) Greater within-host diversity provides a larger pool of variants that can be transmitted and on which between-host selection can act. As the number of segregating sites increases, the within-host dynamics slow and the virus is better able to evolve between-host towards an intermediate spVL that maximises transmission potential.

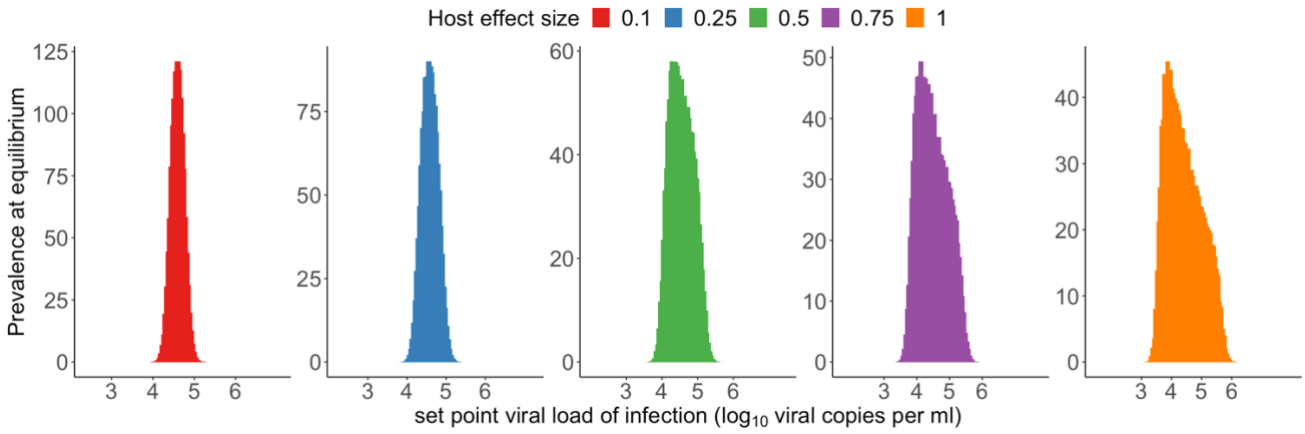

Figure 3.4 Between-host outcomes at equilibrium for an increasingly heterogenous host population. Histograms of spVLs in heterogenous infected population for different maximum host effect size, *e,* for 250 segregating sites. To account for the effect that host genetics has on viral load, we introduced a host specific additive effect to viral load. The size of the host effect is discretely uniformly distributed between -*e* and *e* and there are 50 host types. A maximum effect size of e=0.1 (A) results in a small increase in the range of viral loads observed, and as we increase *e* we observe a more realistic distribution of viral loads. Increasing the effect size towards *e*=1 further flattens and skews the distribution. Corresponding results for other choices of m are present in supp. Figure 9.3.

3.4.4 Many mutations with low fitness cost results in viral load evolving to intermediate levels between-hosts

To capture the epidemiological dynamics and how the mean spVL varies as the epidemic progresses, we numerically integrated the full nested model over the first 100 years of an epidemic. For each scenario (choice of *m*), three simulations were run where the average spVLs of infection types at the start of the epidemic differed and initially there is a single infected individual. Population average viral loads decrease or increase depending upon the initial viral load, with cumulative population changes occurring over approximately a century for m=100 (fig. 3.5C), and over two centuries for larger m values (fig. 3.5D-E).

When we considered few mutations of larger effect, the fittest and most virulent variant rapidly outcompeted other virus types on the within-host scale. As a result, short-sighted evolution blocks between-host adaptation and the fittest variant dominates across the population at the expense of a reduction in transmission potential (fig. 3.5A-B). As the number of sites increase, the within-host dynamics slow and selection for transmission potential on the between-host scale drives the population-level dynamics, leading to gradual evolution towards an intermediate viral load. Importantly, we observe this dynamic for 100 segregating sites despite a high viral load at the point

of mutation-selection balance. i.e. at within-host equilibrium. In this case, the fitness cost of each mutation is sufficiently small to slow within-host adaptation, resulting in stable viral loads during the first years of transmission opportunity.

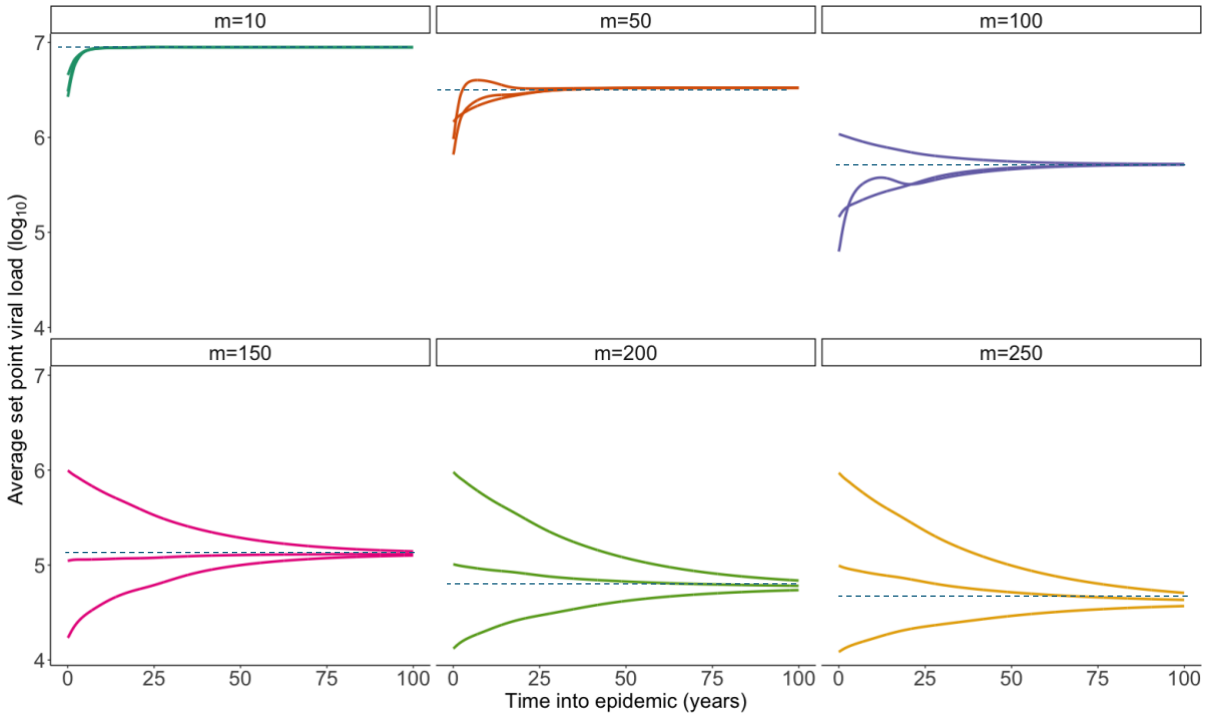

Figure 3.5 Average spVL over time in simulated epidemics. For *m*=10, short-sighted evolution leads to the rapid dominance of the fittest virus type as the single circulating variant within a few years of the epidemic. For *m*=50 and *m*=100, the within-host dynamics create a distinctive pattern at the population level. When epidemics start with lower viral load viruses that have a relatively long infection duration, the within-host evolution infections lead to the emergence and transmission of fitter variants with higher viral loads. This process accelerates the population-level increase in average spVL. When we assume a large number of weakly deleterious mutations (*m*=150,200,250) within-host dynamics are sufficiently slow for selection for transmission potential to influence the epidemiological dynamics and lower the average spVLs, despite the comparatively higher within-host equilibrium viral load. We observe slow cumulative changes in the average spVL, with over 100 years taken for convergence at higher values of *m*.

3.4.5 Viral loads are similar within transmission pairs

Within-host evolution can cause significant genetic change in the quasispecies between early infection and the time of onward transmission. In the absence of host heterogeneity or other environmental effects, we expect estimates of heritability to reduce in response to increased within-host evolution (Mitov and Stadler, 2018). We estimated heritability, h^2, by simulating transmission pairs in an infected population of homogeneous individuals at endemic equilibrium (Figure 3.6). Source infections were sampled based upon the infection-type population structure at the endemic steady

state. Recipient infections were sampled based upon the transmission potential of each virus type during the infection of the source, and therefore the heritability estimates accounts for within-host evolution of viral factors.

I find heritability is lowest for few mutations of large effect (m=10) due to rapid within-host evolution. However even with rapid evolution the heritability level is reasonably high, and this is due to the fact I have sampled source infections from the between-host equilibria solution, where there is very little diversity. In a homogenous population, all other scenarios have approximatley equal heritability, likely due to the initial period of stablity, high probability of transmission during acute infection, and the rapid changes occuring later into infection. As the host effect size increases, heritability decreases consistently across all scenarios, with the rate of decrease being similar regardless of the number of segregating sites, and heritability reduces to a range that has been reported in study estimates of broad-sense heritability. Heritability for m=50 and m=100 segregating sites is consistently greater than models with a larger number of segregating sites. This is because as the number of segregating sites increase, there is greater within-host variation in virus types, leading to more variation in transmitted viral loads between source and recipient.

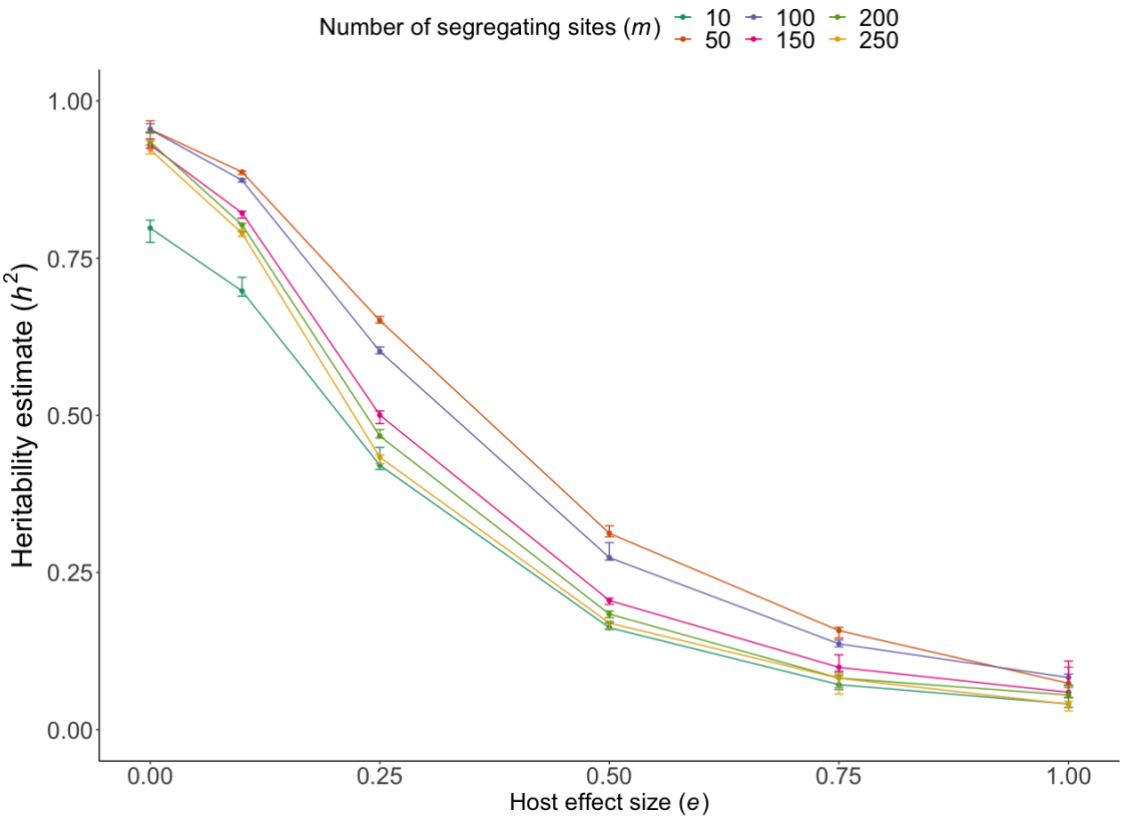

Figure 3.6 Heritability estimates. Heritability is estimated by a parent-offspring regression of spVL in simulated source and recipient pairs. The error bars indicate the standard deviation of the estimates taken from 1000 sampled sets of 500 transmission pairs. The infection type of the source (number of deleterious mutations of infecting virus type) is determined based upon the population-level prevalence of viral types at the endemic steady state. The virus type transmitted from source to recipient is determine based upon the transmission potential of each virus type over the course of the source infection. In a homogenous population, heritability is high, and naturally as a host-effect is introduced the amount of variability in spVL explained by the virus type, i.e. heritability, falls to within the range reported in multiple studies.

3.4.6 Support for model predictions in serodifferent studies dataset

An advantage of the study dataset upon which the analyses of subsequence chapters are built on, is provides longitudinal measures of viral load in phylogenetically confirmed transmission pairs. We measured heritability by regressing the spVL of the recipient against the spVL of the source and found $h^2 = 0.26$ (CI: 0.1 - 0.4) (Figure 3.6A). Our estimate of heritability is reassuringly within the range of previous study estimates, despite differences in study population and viral subtypes. In our model, we observe heritability of this magnitude when we allow viral loads to differ by up to one order of magnitude due to host effects.

To support our proposed mechanism of a mutation-selection balance of weakly deleterious mutations affecting viral load, we tested a prediction of the within-host model. Specifically, under a mutation selection balance, we expect initially high viral

loads to lower over time, and initially low viral loads to slowly increase. For each individual with greater than 1000 days of follow up, we determined the within-host trend by a linear regression of viral load measure against time, excluding any time points inferred to be taken during acute infection. For a between-host analysis, we regressed the slope of each within-host regression against the initial viral load measure. We found a significant negative relationship between initial viral load and the within-host viral load trend over time, in agreement with our model prediction (figure 3.6B).

Since noise in viral load measurements can lead to regression to the mean over time, we tested whether our observed results could be explained purely by noise. To do this, we simulated viral load trajectories that varied solely due to random noise. In all simulations, the relationship between within-host trends and initial viral load was consistently weaker than the observed values (supp. Figure 9.4). These findings suggest that the trend observed is unlikely to be explained by noise alone.

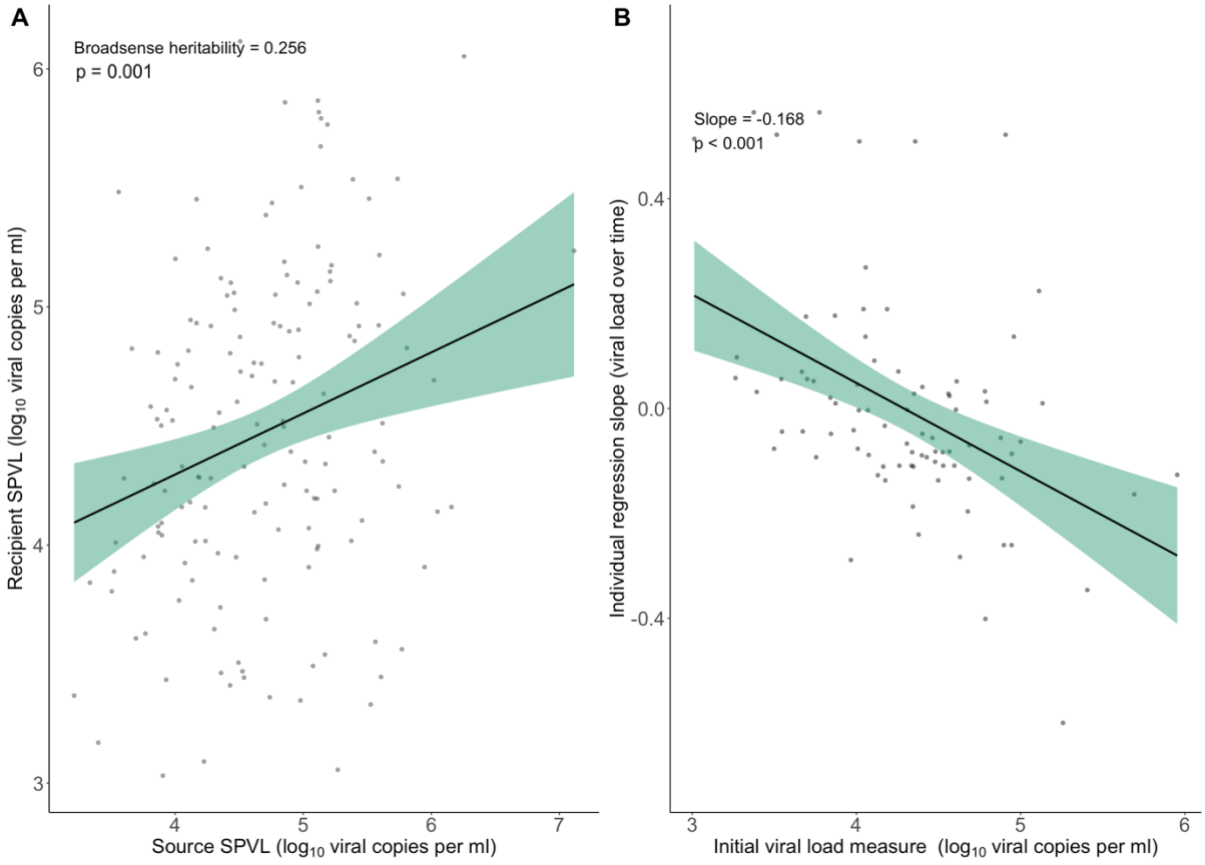

Figure 3.7 Viral load patterns in transmission pairs enrolled in serodifferent studies. A) The relationship between the spVL of the source infection and the spVL of the recipient infection. A linear regression estimated heritability as $\boldsymbol{h^2} = 0.26$ (CI: 0.1 = 0.4, p=0.001). B) Temporal trends in viral load over time. For each individual with over 1000 days of follow up, the regression slope of viral load over time was determined. A between-host level regression of the slope against the initial viral load measure found a significant negative relationship between initial viral load and the regression slope, supporting the model prediction that viral load evolves towards an intermediate value over time.

3.5 Discussion

There are several examples of mathematical models that predict pathogen dynamics by considering evolution across scales (Childs *et al.*, 2019). With a multi-strain model that nests within-host dynamics of HIV within a between-host model, Lythgoe *et al*. (Lythgoe, Pellis and Fraser, 2013) found that within-host evolution is rapid enough that HIV should evolve to high levels of virulence at the expense of fewer onward transmissions, and so within-host evolution was a greater prevailing force than evolution between hosts. By incorporating a reservoir into the model, evolutionary processes are delayed, and short-sighted evolution is prevented; however, the impact of the reservoir was highly sensitive to its assumed size (Doekes, Fraser and Lythgoe, 2017). Van Dorp *et al*. (Van Dorp, Van Boven and De Boer, 2014) proposed an alternative model that considers immune escape and found that high host-

heterogeneity in HLA types and consequently the escape and reversions of immune-escape mutations determine how spVL evolves. However, the model assumes that mutations are time-separated and occur according to a Markov process, effectively limiting the tempo of within-host evolution. Here, I applied the nested model framework from Lythgoe, Pellis and Fraser (2013) without a latent reservoir and consider whether many weakly deleterious mutations can provide an additional and more parsimonious mechanism for spVL evolution under two levels of selection. If the number of segregating sites is sufficiently large, I show this can reconcile multiple paradoxes in HIV biology: the stability of viral loads during chronic infection despite high mutation rates, the heritability of spVL despite considerable within-host evolution between transmission events, and the evolution of spVLs that maximise transmission. This mechanism could also explain why heritable viral factors determining viral load have been so difficult to identify, due to the high statistical power needed to detect them.

Alternative solutions to the problem of rapid within-host evolution have also been proposed, including the existence of rugged and complex fitness landscapes that are difficult for the within-host viral population to traverse (Lythgoe, Pellis and Fraser, 2013), the cycling of virus through the (unreplicating) HIV reservoir which then slows the rate of within-host evolution (Doekes, Fraser and Lythgoe, 2017), and viral factors that are heritable but are not under within-host selection, specifically polymorphisms that target the cell activation rate and are therefore beneficial to the entire virus population (Hool, Leventhal and Bonhoeffer, 2013). It is, however, difficult to reconcile the latter theory with the relationship between viral load and replicative capacity, and none of the described theories provide a fully satisfactory explanation of a heritable set-point viral load that varies by orders of magnitude between individuals, or why viral virulence factors have been so hard to identify (Bartha *et al.*, 2013).

I do not expect this mechanism alone to fully control viral load, as host factors—such as protective HLA alleles—are well-documented contributors to viraemic control (Kiepiela *et al.*, 2004; Zhang *et al.*, 2013; McLaren *et al.*, 2015; Leitman *et al.*, 2016). These host factors shape the immune system's ability to suppress viral replication and explain part of the variation in viral load across individuals. Specific viral mutations with large effects on fitness have been identified (Gabrielaite *et al.*, 2021), but these are relatively low in number and the total narrow-sense heritability from all genetic hits is significantly lower than our estimates of broad-sense heritability. It is therefore likely

that viral genetic factors each carry a small effect that studies do not have the statistical power to detect. Virus-host interactions further complicate this picture. The complex relationship between viral evolution and host immune responses creates a dynamic environment where the effects of individual genetic variants may be context-dependent, making it challenging to identify robust GWAS hits associated with spVL.

HLA footprints have been associated with spVL heritability, as viruses adapted to the source's HLA profile maintain higher viral loads when transmitted to recipients with similar HLA backgrounds (Carlson *et al.*, 2016). However, this mechanism cannot explain the dominant power of between-host evolution and the selection for maximal transmission fitness at a population level (Van Dorp, Van Boven and De Boer, 2014).

Studies of the HIV fitness landscape have identified the significant contribution from mutations that carry a very small fitness cost to replicative capacity, with these largely being synonymous (Zanini and Neher, 2013). Synonymous mutations, while not directly affecting protein structure, can still impede viral fitness by influencing RNA stability, translation efficiency, or protein folding. The deleterious component of the landscape has been shown to be universal across infections and fundamental to the high diversity within HIV group M diversity (Zanini *et al.*, 2017).

Further support for my model was also found by investigating testable predictions with the longitudinal measures of viral load. I expect to observe the viral load to increase or decrease towards an intermediate value during the course of infection, as the within-host model here has shown. I indeed observe this trend in individuals enrolled in one of the three serodifferent studies with follow up of 1,000 days. Confirmation of this result in a dataset with follow up over many years of infection would provide further support for this trend, as well as investigations into the effects of host factors on this relationship.

As discussed in the thesis introduction, the quasispecies framework has been the subject of controversy and criticism within the field of viral evolution (Holmes, 2010). Here, the framework was chosen primarily because its mathematical formulation naturally supports a nested modelling approach—an essential feature for capturing both within-host viral dynamics and between-host transmission. Importantly, this application does not rely on the population-level units of selection often viewed as problematic by critics. Instead, selection is modelled at the level of individual genomic

sites. This framework is particularly well-suited for modelling the accumulation of weakly deleterious mutations across many evolving sites. Rather than tracking specific alleles, the model focuses on the total mutational burden—specifically, the overall number of deleterious mutations without respect to the specific allele or locus — offering a level of abstraction that enhances computational tractability. By considering mutation–selection balance in terms of cumulative mutational load rather than the presence or frequency of specific variants, the approach captures the essential evolutionary forces hypothesised to be shaping the set point viral load. In this way, the model represents a pragmatic use of quasispecies theory, capitalising on its analytical strengths for testing the hypothesis in a way that would not be possible with classic population genetics and its focus on the fate of specific alleles.

As with all mathematical models, the model proposed here represents a significant simplification of complex biological processes for the sake of tractability and interpretation. Though this model reproduced known behaviour, in reality other processes will contribute to varying degrees, such as repeated immune escape and reversion across different host environments, and perhaps a small number of mutations of large effect. Developing an informed understanding of the virus factors that control virulence, and how they evolve in response to selection at the within- and between-host scales, will ultimately provide important insights into the severity of viral infections, including HIV, how this might change through time, and improve future treatments and public health policy.

4 HIV within-host recombination across the genome

4.1 Abstract

Recombination plays a pivotal role in generating within-host diversity and enabling HIV's evolutionary success, particularly in evading the host immune response. Despite this, the variability in recombination rates across different settings and the underlying factors that drive these differences remain poorly understood. In this study, I analysed a large dataset encompassing hundreds of untreated, longitudinally sampled infections using both whole-genome long-read and short-read sequencing datasets. By quantifying recombination rates, I uncover substantial variation across subtypes, viral loads, and stages of infection. I also map recombination hot and cold spots across the genome using a sliding window approach, finding that previously reported inter-subtype regions of high or low recombination are replicated at the within-host level. Importantly, these findings reveal the significant influence of selection on recombination, showing that the presence and success of recombinant genomes is strongly interconnected with the fitness landscape. These results offer valuable insights into the contribution of recombination to evolutionary dynamics and demonstrate the enhanced resolution that long-read sequencing offers for studying viral evolution.

4.2 Introduction

HIV diversifies rapidly within-host, with genetic diversity in a single virus population reaching upwards of five percent within a few years of infection in the envelope gene, Env (Shankarappa *et al.*, 1999). The short viral generation time, rapid turnover of virions and a high mutation rate generate mutations that enable the virus to continually survive against the immune response of the infected host (Perelson *et al.*, 1996). Nevertheless, the majority of mutations are deleterious (Zanini *et al.*, 2017), and the recombination of viral genomes is an important mechanism by which deleterious mutations are purged and genomic integrity is preserved (Rawson *et al.*, 2018). In addition, recombination combines beneficial mutations onto a single genome, with selection of these combinations linked to drug resistance, disease progression and immune-escape (Rambaut *et al.*, 2004), and it has been suggested that recombination accelerates the rate of adaptive evolution (Moradigaravand *et al.*, 2014; Sanborn *et*

al., 2015). The study of the dynamics of recombination is therefore an important piece in the puzzle for developing effective treatments and vaccines.

HIV is functionally diploid or “pseudodiploid”, meaning each virion contains two copies of its single-stranded, positive-sense RNA genome. When a host cell is co-infected by two genetically distinct virions, their respective RNA genomes can be co-packaged into the same viral particle—a prerequisite for recombination (Zhuang et al., 2002; Chen et al., 2009). Upon infection of a new target cell, reverse transcriptase can switch between these two co-packaged RNA genomes during reverse transcription, generating a recombinant DNA genome that is a mosaic of the two parental sequences. This template switching is the primary molecular mechanism of retroviral recombination and occurs with an estimated average of two to four switches per replication cycle (Zhuang et al., 2002; Onafuwa et al., 2003).

The overall rate and pattern of recombination observed *in vivo* are shaped by multiple interacting mechanisms beyond molecular template switching. One key factor is the probability of cellular co-infection, which is a prerequisite for recombination between distinct viral genomes. Studies in mice and cell cultures have shown that higher levels of co-infection correlate with increased recombination rates (Levy et al., 2004). Additionally, Romero and Feder (2024) found a positive association between viral load—a proxy for viral abundance—and recombination rates inferred from viral genomes sampled in human infections. However, it remains unclear whether this relationship extends to tissue compartments beyond blood plasma. Notably, cell-to-cell transmission may significantly enhance the likelihood of co-infection compared to cell-free infection, thereby increasing opportunities for co-packaging and recombination (Hübner et al., 2009; Komarova et al., 2013).

The efficiency of co-packaging genetically distinct genomes depends on the relative frequencies of different viral variants in the population, as well as potential biases in genome dimerisation and packaging signals that can favour or disfavour heterozygous virions (Nikolaitchik et al. 2013). Moreover, not all co-packaged genomes result in successful recombination—template switching depends on sequence similarity between the two genomes and may be suppressed by structural constraints or highly divergent regions (Balakrishnan et al 2001, An et al 2002, Balakrishnan et al 2003, Nikolaitchik et al. 2011).

Finally, the recombination patterns observed in clinical or experimental data reflect not only these mechanistic processes but also the action of natural selection. Recombinant genomes may rise in frequency if they confer selective advantages, such as combining mutations that enhance immune escape or drug resistance (Nagaraja et al 2016). Conversely, recombination may disrupt co-adapted sets of mutations, leading to reduced fitness and purifying selection against certain recombinant forms (Galli et al. 2010). As a result, the observed recombination rate is an *effective* rate that integrates both the underlying mechanistic frequency of recombination events and the selective forces acting on recombinant progeny

If an individual is infected by two different subtypes of HIV, a recombinant virus may be formed that is composed of a mixture of the genetic material of the two subtypes (Song *et al.*, 2018). This process has led to the creation of unique recombinant forms (URFs) of HIV, which can be successfully transmitted and spread within populations (Neogi *et al.*, 2017; Liu *et al.*, 2019). Over time, some URFs become prevalent and establish themselves as circulating recombinant forms (CRFs) (Bbosa, Kaleebu and Ssemwanga, 2019). To date, over 100 CRFs have been identified globally.(Williams *et al.*, 2023)

There exist numerous software applications for detecting recombination from sequence data (Lemey, Vandamme and Salemi, 2009), however these methods are typically developed for diverse populations or inter-subtype recombination, and are therefore not suitably robust for within-host virus populations where diversity is relatively limited. *In-vitro* methods include the use of retroviral reporter systems (Hu *et al.*, 1997; Chen, Rhodes and Hu, 2005; Chen, Powell and Hu, 2006), as well as approaches that leverage the genetic differences across subtypes to monitor recombination events (Jetzt *et al.*, 2000; Zhuang *et al.*, 2002; Levy *et al.*, 2004), however controlled laboratory conditions cannot mimic the complexities of the host environment (Schlub *et al.*, 2010) and any conclusions are further limited due to confounding by factors, including RNA homology (Balakrishnan, Fay and Bambara, 2001; Balakrishnan *et al.*, 2003).

In-vivo approaches better reflect the within-host environment, in particular the dynamics of evolution and recombination under immune pressure. Inevitably, observed recombination in natural infections will underestimate the true recombination rate, as recombination between identical genomes or the generation of a fatal

combination of mutations on the same genome is undetectable. Instead, *in-vivo* estimates reflect the *effective recombination rate*, which quantifies the amount of recombination that contributes to the diversity and evolution of the virus population.

Methods can be broadly characterised as breakpoint detection or rate estimation. Breakpoint detection identifies the likely genome location at which template switching has occurred and an array of methodologies and software have been developed specifically for application to viral sequencing data (Jaya, Brito and Darling, 2023). By categorising regions of above or below average numbers of breakpoints, it has been shown that recombination does not occur randomly across the genome both within and between subtypes (Fan, Negroni and Robertson, 2007; Archer *et al.*, 2008; Smyth *et al.*, 2014; Jia *et al.*, 2016; Tongo, De Oliveira and Martin, 2018; Grant *et al.*, 2020). Whether hot and cold spots for recombination are consistent across large numbers of individuals and subtypes is not known, and it is difficult to remove the possibility of sequencing artefacts, such as primer locations, that leave a false signal in the data.

In contrast, recombination rate estimation takes a larger scale approach, and the aim is to quantify the frequency at which recombination occurs over a given genome region – typically the entire genome – and provide a continuous measure that is often expressed per generation per genome site. Reported estimates range from 10^{-5} to 2×10^{-4} recombination events/site/generation, suggesting recombination may be occurring even more frequently than point mutations (Shriner *et al.*, 2004; Neher and Leitner, 2010; Batorsky *et al.*, 2011; Romero and Feder, 2024).

Advances in next-generation deep sequencing have revolutionised the study of within-host dynamics, enabling fine-scale mapping of genetic variation across entire genomes. Given longitudinally sampled sequences, it is now possible to accurately measure decay of linkage disequilibrium (LD) over time. Mutations that initially arise on the same genome exhibit strong LD, but this linkage diminishes as recombination events occur. The rate of LD decay is influenced by the genomic distance between two sites; the greater the distance, the more potential recombination breakpoints exist, leading to a faster decay in linkage. Recent methods leverage the relationship between LD decay and recombination to infer recombination rates, providing deeper insights into viral evolution (Neher and Leitner, 2010; Romero and Feder, 2024).

The success of a recombinant genome is likely influenced by the fitness landscape of HIV, and functional constraints have been reported to lower the contribution of recombination to viral diversity in ENV (Simon-Loriere *et al.*, 2009). Despite the clinical significance of identifying factors that promote or inhibit the effective recombination rate, the small number of individual infections included in most studies has made characterising the correlates of recombination during natural infection problematic.

Reported rates of recombination typically represent the average across the entire genome, yet the non-random distribution of breakpoints implies there may be extensive undetermined variation within and between genes. The success of a recombinant genome is likely influenced by the fitness landscape of HIV, and functional constraints have been reported to lower the contribution of recombination to viral diversity in ENV (Simon-Loriere *et al.*, 2009). Other proposed mechanistic drivers of recombination include sequence identity and RNA structure (Galetto *et al.*, 2004; Baird *et al.*, 2006; Archer *et al.*, 2008).

In this study, I analysed long-read, whole genome, next-generation sequences (NGS) from hundreds of individual infections sampled from cohorts of HIV serodifferent African heterosexual couples. As a result, the dynamics of *in-vivo* recombination can be studied with an unprecedented level of detail. I applied the method of Romero and Feder (2024) for recombination analysis via time series linkage decay (RATS-LD), which uncovered significant differences in recombination rate by subtype, viral load and infection stage in our data. I also observed recombination rates to vary by orders of magnitude across the genome, generate supporting evidence for previously identified regions of high and low amounts of recombination, and revealed previously unreported hot and cold spots of recombination. Importantly, patterns across the genome did not depend upon the sequencing platform (PacBio vs Illumina). The findings of the study contribute to our understanding of how recombination shapes the viral population at both within and between host scales and may provide vital insights for preventing or slowing the emergence of drug resistance and developing successful vaccines.

4.3 Methods

4.3.1 Study cohort

All individuals analysed in this study were enrolled in one of the three African HIV serodifferent heterosexual couples studies detailed in chapter 2. “Source” individuals were living with HIV-1 when recruited, and “recipient” individuals were initially HIV-1 negative and seroconverted at some point during the study. For the main analysis dataset, individuals were selected if they had at least two samples sequenced on the PacBio long-read platform and subsequently processed with Shiver. Details of the sample preparation, sequencing, and processing are provided in Chapter 2. Both recipients and source individuals were included, however as I do not make use of transmission information, there was no requirement that for a recipient to be included their associated source must also be and vice versa. For a sensitivity analysis, I constructed a secondary dataset, consisting of illumina-short reads, with sequencing methods again provided in Chapter 2.

The software Phyloscanner (Wymant, Hall, *et al.*, 2018) was used to generate sliding windows of alignments for each individual. Windows covered the whole genome from position 520 (with respect to HXB2 sequence). The reads in every window were aligned to HXB2 positions, and so insertions and deletions relative to HXB2 were excluded from the analysis. For the primary dataset of long-read sequences, a window length of 750bps was chosen to balance the need for both long reads to measure the effect of distance on linkage measures and the need for sufficient depth to accurately quantify allele frequencies. For the secondary dataset, a window length of 250bps was chosen as the Illumina platform produces short reads. A window needed a depth greater than 10 reads to be included in either dataset.

4.3.2 Linkage Disequilibrium Calculation

For each window of sequences outputted by Phyloscanner, allele frequencies at every position were calculated. Any position at which the proportion of gaps was greater than 10% was excluded, and to be included in the recombination rate analysis the minor allele frequency (MAF) must exceed 5%. For all such sites, linkage disequilibrium (LD) was calculated for each pair of sites *(X, Y)* in the window. The standardised LD metric is defined as the D' statistic:

$$D' = \frac{|p_{AB} - p_A p_B|}{D_{max}}$$

Where allele A and allele B are the major alleles at position *X* and *Y* respectively, p_A is the frequency of allele A, p_B is the frequency if allele B, p_{AB} is the frequency of the haplotype AB, and D_{max} is the maximal possible value of linkage given p_A and p_B. For a pair of sites (X,Y) sampled at time t to be included in the recombination analysis, all four possible haplotypes must be observed. Additionally, the standardised LD metric for a pair of sites (X,Y) at the earliest included time point must exceed 0.2. LD calculations were performed in Python 3.

4.3.3 Recombination Analysis

Recombination Analysis by Time Series Linkage Decay (RATS-LD) is a method developed by Romero and Feder (2024) that exploits the relationship between linkage, recombination, and distance and time. In brief, mutations that appear in the population on the same background are initially in high linkage disequilibrium, and the rate at which linkage decays as a result of recombination is related to the physical distance between the sites at which the mutations appeared. The RATS-LD method assumes that between time *t* and time $t + \Delta t$, the linkage between allele A and allele B decays by the equation:

$$D'(t + \Delta t) = D'(t) e^{-\rho d \Delta t}$$

Where *d* is the distance between sites, Δt is the number of generations between time points and ρ is the recombination rate per site per generation (/site/generation). The RATS-LD method calculates the metric, D'_{ratio}, for each pair of sites across two time points, defined as:

$$D'_{ratio} = -\log\left(\frac{D'(t+\Delta t)}{D'(t)}\right) = \rho d \Delta t$$

This equation implies a linear relationship between D'_{ratio} and $d\Delta t$ with ρ. However, to account for the effect of small differences arising from sampling errors, as well as the effect of other evolutionary forces for large value of $d\Delta t$, the method expresses the relationship between linkage decay and time-scaled distance in the functional form:

$$f(x) = c_0 + c_1(1 - e^{-c_2 x})$$

Near zero, the slope approaches ρ, and so the estimate of the recombination rate /site/generation, $\hat{\rho}$, is given by

$$\hat{\rho} = f'(x = 0) = c_1 c_2$$

The calculation of D'_{ratios} and fitting of the curve $f(x)$ were performed in R v4.2.1. For a full explanation of RATS-LD and validation of the method with simulations considering recombination rates, selection effects and population size, refer to Romero and Feder (2024).

As the pairs of sites are identified from overlapping windows of 750bp reads over the genome, the same pair of sites can appear in the dataset multiple times. To account for this, the mean D'_{ratio} for a given pair of sites and timepoints is included in the dataset for curve fitting. Sites within the variable loops V1-V5 were excluded from the analysis as the method performed poorly in these regions, most likely due to alignment issues.

Confidence intervals were generated by bootstrapping. For each rate estimate, 1000 bootstrap replicates were generated and analysed. Bootstrapping was performed at an individual level, such that for an individual with *n* pairs of sites, *n* pairs would be sampled with replacement for curve fitting.

4.3.4 Viral Load Matching

To account for the effect of differences in viral load (VL) when comparing recombination rates by host factors, I adjusted datasets (e.g. subtype specific, recipient only) such that the viral load distributions approximately matched. Datasets filtered by subtype were subsampled to match the VL distribution of subtype A infections. To compare the recombination rate of source individuals against recipient individuals, the source dataset was subsampled to match the viral load distribution of recipients. Specifically, the viral loads of all diverse pairs identified from sequences sampled from recipient individuals were split into 5 equally sized quantiles, and the diverse pairs from source individuals were then subsampled such that the proportion of data points in each quantile was consistent between the two groups. The subsampling approach was repeated on the subtype data, with subtype C and D subsampled to match subtype A viral loads.

The VL associated with a D'_{ratio} is the mean of the individual's VL at time t and time $t + \Delta t$. If a viral load measure was not available for the same date that the virus sampling took place, the most recent viral load measure is imputed. The viral load measures of the dataset that is to be matched against were split into three groups containing an equal number of D'_{ratios}. The second dataset is then sampled such that proportion of D'_{ratios} in each of the three groups is equal to that of the matching dataset.

4.3.5 Recombination estimates by sliding windows

To identify local and global patterns in recombination rates, a sliding window approach was taken. For analyses of the main dataset, the recombination rate of windows of 500bps length with overlaps of 450bps was estimated. A window was excluded if the bootstraps from curve fitting did not converge sufficiently, with insufficient convergence defined as when the log-transformed ratio of the upper and lower confidence interval bounds was lower than 0.8. Poor convergence across bootstraps was most likely caused by proximity to both strong hot and cold spots within a single window, too few pairs of sites or alignment issues masking recombination signal.

Sliding window rates were also measured for subtype-level comparison and for the Illumina dataset, where a window length of 750bps length was chosen to account for the reduction in the total number of pairs in the analysis.

4.3.6 Estimation of correlation for window-specific rates

We calculated the Pearson correlation coefficient for the median recombination rate at each 750bp genome window, with three comparisons in total: subtype A vs subtype C, subtype A vs subtype D, and Illumina vs PacBio. By considering the correlations between windows, we are measuring the extent of similarity in patterns of hot and cold spots for recombination along the genome across subtypes. To test the statistical significance of the correlation, we performed permutation tests. The window IDs for one subtype were shuffled and a correlation test was performed. We repeated the permutated correlation calculation 10,000 times to provide a distribution of coefficients to compare against the correlation coefficient of the non-permutated dataset in order to test for statistical significance of the observed correlation. To determine the windows that differed significantly across subtypes, we performed a linear regression of subtype

A window rates at window *i* against subtype C/D window rates at position *i.* Outliers were defined as windows at which the standardised residual was greater than 2.

4.3.7 Simulations

As was presented in Romero and Feder (2024), simulations of neutrally evolving populations were generated in SLiM (Haller and Messer, 2023), with a mutation rate of 10^{-5} and an effective population size of 10^4. Generated sequences were 1000 bps long and simulations ran for 5×10^4 generations, after which 100 sequences were sampled every 50 generations for 60 generations. The recombination rate for each simulation was fixed, ranging from 7.5×10^{-6} to 5×10^{-5} recombination events /site/generation. For each choice of recombination rate, 100 simulations were performed, and D'_{ratios} for the outputted sequences were pooled.

4.4 Results

4.4.1 Dataset characteristics

A total of 306 individuals had at least two viral samples sequenced with the long-read deep-sequencing PacBio pipeline and were therefore included in the main analysis dataset, representing 163 sources and 143 recipients. A secondary dataset was generated to allow for verification of the results of the analysis. This consisted of individuals with at least two viral samples of sufficient depth sequenced with the Illumina pipeline, a minority of whom also featured in the primary dataset. In total, 356 individuals were included in the Illumina dataset, of which 171 individuals also appear in the PacBio dataset.

Log-transformed viral load measurements were approximately normally distributed with a mean of $4.79\ \log_{10}$ copies per ml. Source individuals exhibited a significantly lower average viral load compared to recipients (t-test of log-transformed viral loads, p-value 1.3×10^{-5}, mean of $4.90\ \log_{10}$ copies per ml in recipients versus $4.66\ \log_{10}$ copies per ml in sources). Higher viral loads in recipients are likely due to recipients being captured earlier into reflection, with some recipients captured towards the end of acute infection. Infections were predominantly classified as subtype A1 (53%), with 18% identified as subtype D, 18% as subtype C, and the remaining 11% largely composed of circulating recombinant forms (CRFs) and unique recombinant forms (URFs).

The recombination rate, defined as the number of recombination events per site per viral generation, was inferred from linkage decay over time between alleles at pairs of diverse sites using the PacBio sequencing data. We applied the RATS-LD method which exploits the relationship between LD and recombination, by measuring the rate at which linkage decays between two sampling points as a function of the product of time and genomic distance. Source individuals contributed the majority (~95%) of site pairs due to the higher genetic diversity observed later in infection, supporting the interpretation that source individuals were typically at a more advanced stage of infection than recipients. Overall, the dataset encompasses 775,467 unique site pairs, with 33% located in *env*, 27% in *pol*, and 13% in *gag*, and 28% in other regions of the genome.

4.4.2 Measured recombination saturates across long genomic distances and timescales

The length of sequencing reads significantly impacts the resolution and accuracy of detecting linkage decay between alleles, with longer reads capturing genetic variants that are positioned further apart on the same genome, providing a more comprehensive view of how linkage decays over distance. The time between sampling points is also important, with the higher the number of generations the lower the resolution at which recombination can be measured. This is due to saturation caused by multiple recombination events between pairs of sites between time points. Here, I used the long reads produced from PacBio sequencing, and I also benefited from the relatively frequent sampling of individuals.

Although the aim is to quantify the *effective recombination rate*, which captures the recombination that contributes to viral population diversity, the constraints imposed by sequence length and time between sampling make this challenging. Instead, I report the *measured recombination rate*, which is an outcome of recombination (composed of the frequency of cellular co-infection, genome co-packing and template switching), evolutionary processes (e.g. selection) and linkage decay. I first tested the expectation that both read length and time between sampling affect measured rates of recombination by varying the time-scaled distance threshold. Time-scaled distance, $d\Delta t$, is the product of the number of generations over which we measure linkage and the genomic distance between the sites, and the threshold is the maximum possible value of this product. A lower threshold therefore excludes pairs of sites that are either

genomically close but have sampling points spread far apart in calendar time or sites that are distant on the genome but have sampling points closely spaced in time.

First, I analysed simulated populations with known recombination rates using the package SLiM (Haller and Messer, 2023). As the maximum value of $d\Delta t$ increased, the measured recombination rate decreased (Figure 4.1A). When increasing the choice of the true rate of recombination, shorter distances between sites and/or shorter time intervals between sampling are required to accurately recover the true recombination rate. Repeating the analysis on the PacBio sequencing dataset, we observed a similar relationship between the estimated genome-wide recombination rate and the maximum value of $d\Delta t$. (Figure 4.1B).

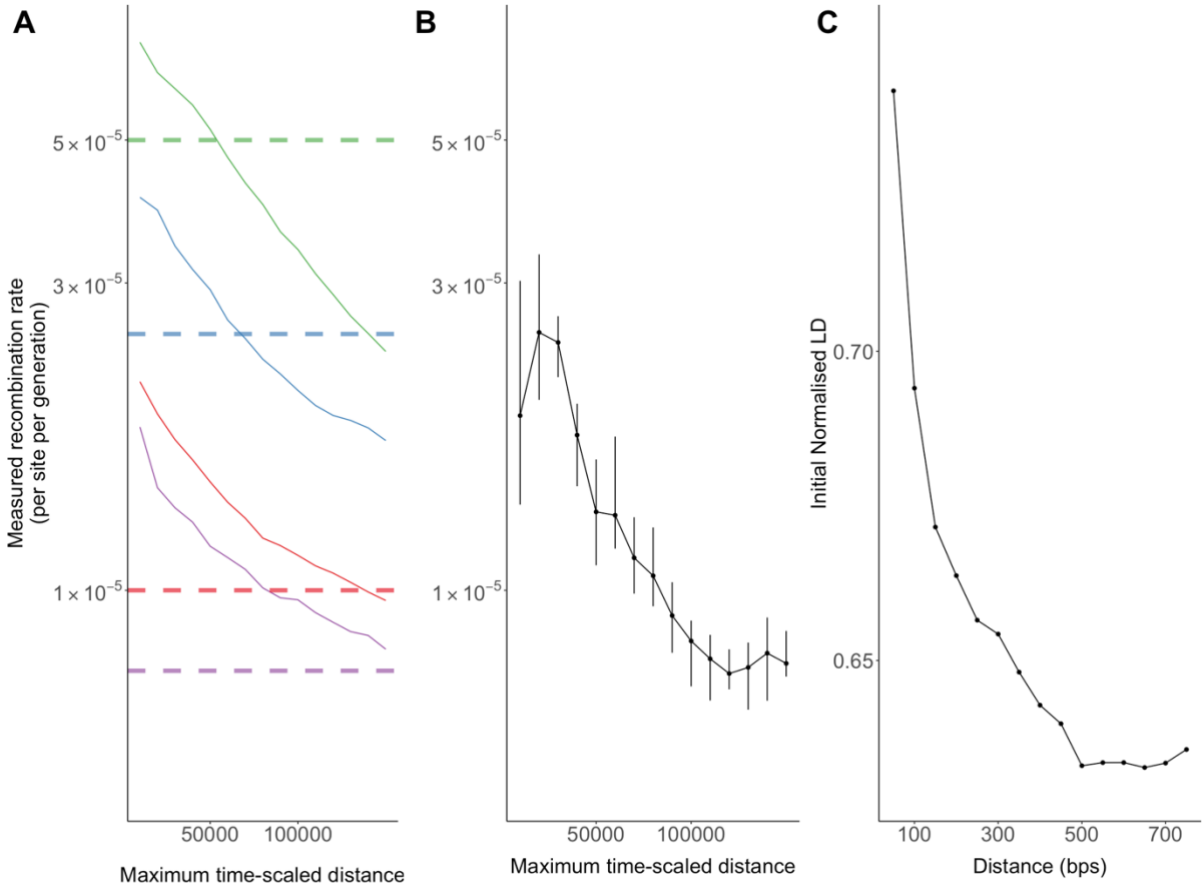

Figure 4.1 Recombination rate decreases with time-scaled distance. A) The recombination rate/site/generation calculated by the RATS-LD method (Romero and Feder, 2024) with sequences representing simulated virus populations evolving over time (Haller and Messer, 2023). Four recombination rates (/site/generation) values were considered ($\mathbf{5 \times 10^{-5}}$ in green, $\mathbf{2.5 \times 10^{-5}}$ in blue, $\mathbf{10^{-5}}$ in red, $\mathbf{7.5 \times 10^{-6}}$ in purple). The dashed lines indicate the true recombination rate fixed as a parameter in the simulation. As the maximum time-scaled distance increases, the inferred recombination rate reduces. The best choice of cut-off for time-scaled distance for accurate estimation depends upon the magnitude of the true rate of recombination. B) The whole-genome recombination rate for increasing maximum time-scaled distance threshold decreases for the PacBio sequences dataset. Time-scaled distance is the product of the genomic distance between two sites and number of generations between sampling points. Error bars indicate 95% confidence intervals generated from 1000 bootstrapped replicates. C) The initial normalised LD (LD standardised by the maximum possible value) of a pair of sites given the genomic distance between sites for the PacBio sequence dataset. Data is binned into groups spanning 50bps, ranging from 0 - 50bps to 700 - 750bps, and the mean linkage of the group is denoted by a point. The decrease in LD plateaus at 500bps, and linkage at sites situated greater than 500bps likely reflects residual background linkage and are not informative for recombination estimation.

This relationship likely occurs because as the time-scaled distance increases, recombination reaches a point of saturation where the linkage between alleles at two sites is completely broken down. Beyond this point, further increases in distance or number of generations do not reveal additional recombination events because the alleles are already effectively unlinked. This saturation can artificially lower the

measured recombination rate as the decay of LD plateaus. The distance at which sites can be in significant linkage along the HIV genome has previously been determined to be lower than 200bps (Neher and Leitner, 2010; Zanini *et al.*, 2015), however we found linkage continued to decay as distance increased until plateauing at around 500bps, at which point the linkage estimates reflect residual background LD that is not dependent upon distance and is therefore not informative for recombination estimation (Figure 4.1C). To account for the maximum number of base pairs over which we observe linkage decaying with distance, and a mean number of approximately 150 generations between time points, we apply a maximum value of $d\Delta t = 75{,}000$ for our remaining analyses, with the exception of analysis on short-read Illumina data which has a maximum of $d\Delta t = 50{,}000$. Although measured recombination rates are sensitive to the data factors, and the optimal maximum time-scaled distance is uncertain when the true recombination rate is unknown, consistent distance thresholds or appropriate adjustments can still provide valuable insights into the factors influencing recombination.

4.4.3 Recombination rate varies by viral load stage of infection

It has been previously observed in a small cohort of 10 individuals that the recombination rate varied by viral load both within and between individuals, with high viral load (>$4.9\ \log_{10}$ copies per ml) associated with elevated rates (Romero and Feder, 2024). We performed a similar analysis on our dataset to verify the finding in a large independent dataset. Viral load measures were grouped into tertiles of increasing value, with each group containing equal numbers of pairs of sites. Tertiles were chosen due to maximise the number of data points contributing to each group. I found substantial variation across the groups, with high viral load (>$4.8\ \log_{10}$ copies per ml) associated with a larger genome-wide recombination rate estimate (Figure 4.2A), corresponding closely to the existing estimate of the viral load values at which recombination rates were elevated (Romero and Feder, 2024).

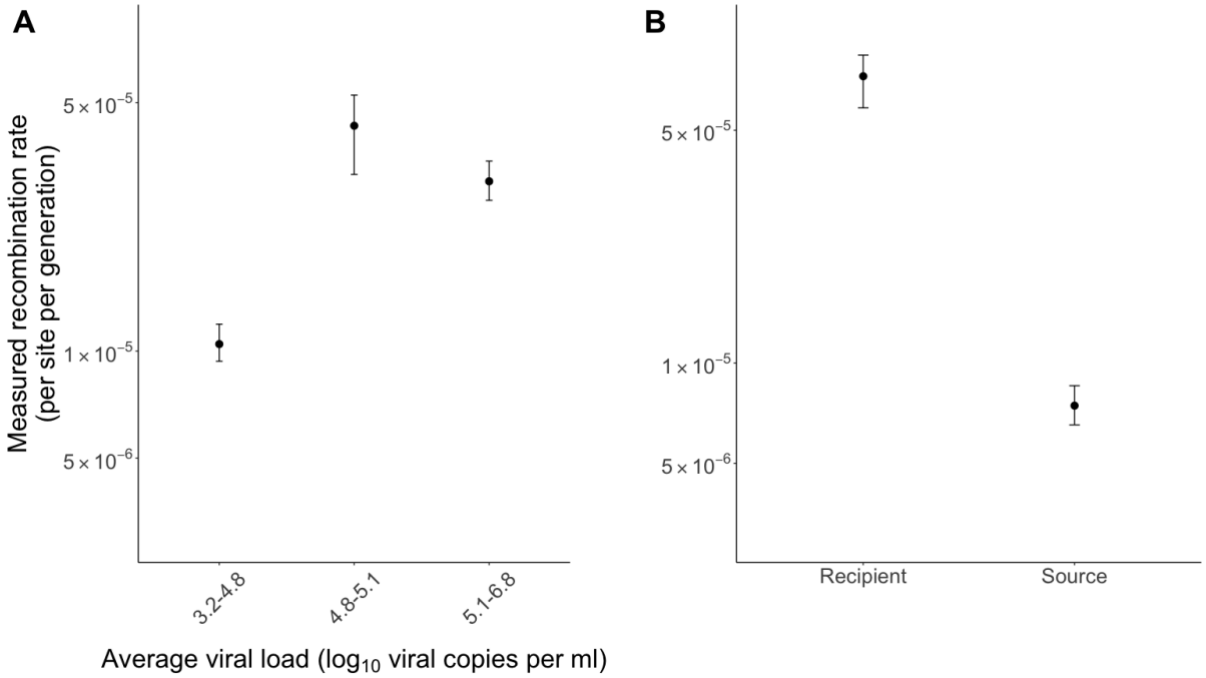

Figure 4.2 Average recombination rate varies by viral load and stage of infection. The recombination rate /site/generation by viral load group for subtype A infections. The median of bootstrapped replicates is represented by a point and 95% confidence intervals are represented by bars. Groups represent tertiles of viral load values that capture average viral load across the two time points that contribute to the *D'* ratios. Each tertile contains the same number of *D'* ratios. In agreement with Romero and Feder (2024), recombination estimates increase with viral load. B) The recombination rate for the recipient versus the source group. The median and confidence intervals of the bootstrap replicates are represented as they are in figure A. For individuals sampled later into infection (i.e. sources) the rate is significantly lower than in individuals sampled within the first 12-18 months following infection (i.e. recipients). The source group is subsampled such that the distribution of viral load approximately matches that of the viral load distribution of the recipient group, due to the recipients exhibiting higher viral loads which may drive higher recombination rates. To match the groups, the viral loads of all diverse pairs identified from sequences sampled from recipient individuals were split into 5 equally sized quantiles, and the diverse pairs from source individuals were then subsampled such that the proportion of data points in each quantile was consistent between the two groups.

To investigate whether the effective recombination rate is a dynamic quantity that varies over the course of infection, we compared rate estimates across source and recipient groups. The evolutionary dynamics between the groups are likely to be distinctive due to the level of adaptation towards the host immune response - recipients were sampled within the first 12 to 18 months of infection, while source individuals had been infected for several years. The measured recombination rate in recipients was $7.3 \times 10^{-5}(\text{CI: } 6.2 \times 10^{-5} - \ 8.9 \times 10^{-5})$ /site/generation, while the measured recombination rate in the source group was substantially lower at $8.7 \times 10^{-6}(\text{CI: } 7.8 \times 10^{-6} - \ 1.0 \times 10^{-5})$ /site/generation (figure 4.2B). As the source dataset was subsampled to match the viral load distribution of the recipient dataset, it

is unlikely that the differences is driven by variation in viral load. Rather, I hypothesise that the reduction in the effective recombination rate is likely due to differences in selection for and against recombinants, rather than lower rates of cell co-infection or template switching.

4.4.4 Variation in rates of recombination between and within genes

When the recombination rate is inferred from pairs of sites across the entire genome (i.e. genome wide), the median number of events /site/generation is approximately 1.04×10^{-5} (CI: $0.91 \times 10^{-6} - 1.21 \times 10^{-5}$), which falls within the range of several other *in-vivo* estimates (Shriner *et al.*, 2004; Neher and Leitner, 2010; Batorsky *et al.*, 2011; Romero and Feder, 2024). However, when the genome is divided into specific genes and protein domains, significant fluctuations in recombination rates are observed (Table 4.1). The recombination rates for the entire *gag* and *pol* genes are relatively similar, at 1.28×10^{-5} and 1.15×10^{-5} recombination events/site/generation respectively, however POL contains protein domains with exceptionally high recombination rates, with the rate in the p31 integrase domain exceeding the average by more than fivefold. Conversely, in the ENV gene, the recombination rate is notably lower than average, a consequence of a relatively low rate of recombination in GP41, however the bootstrap intervals overlap across ENV, POL and GAG estimates.

Table 4.1 The recombination rate /site/generation by gene or protein domain. The difference to the genome average is calculated as a ratio. Above-average rates are coloured in red and below-average in blue.

* Variable loops are excluded from the analysis and therefore only 70% of the protein nucleotides are included.

Gene	HXB2 coordinates	Recombination rate ($\times 10^{-5}$)	Ratio difference to genome median
gag	790 – 2292	1.33 (CI: 1.02 - 1.66)	1.26
P17	790 – 1186	1.93 (CI: 1.18 – 3.04)	1.84
P24	1186 – 1879	1.38 (CI: 0.85 - 2.01)	1.31
pol	2085 – 5096	1.00 (CI: 0.79 – 1.24)	0.96
prot	2253 - 2550	1.42 (CI: 1.02 – 1.88)	1.35
p51 RT	2550 – 3870	1.00 (CI: 0.78 – 1.43)	0.95
p15 RNase	3870 - 4230	2.82 (CI: 2.06 – 5.02)	2.69
p31 int	4230 - 5096	5.31 (CI: 1.52 – 8.82)	5.05
vif	5041 – 5619	0.38 (CI: 0.30 – 0.59)	0.36
vpr	5559 – 5850	2.25 (CI: 1.48 – 2.83)	2.14
vpu	6062 – 6310	8.38 (CI: 4.23 – 14.5)	7.99
env	6225 - 8795	0.66 (CI: 0.41 – 1.04)	0.62
gp120*	6225 – 7758	0.82 (CI: 0.45 – 2.5)	0.79
gp41	7758 - 8795	0.58 (CI: 0.31 – 1.11)	0.55

To better understand the drivers of inter and intra- gene variation, I applied a sliding window approach to quantify the relative recombination rate across the genome (Figure 4.3). The range of rate estimates was very wide, spanning as low as 1.35×10^{-6} (HXB2: 8080-8580) up to 1.27×10^{-4} (HXB2: 4510-5010) recombination events/site/generation a difference of 100-fold. Consistent with previous observations (Fan, Negroni and Robertson, 2007; Jia *et al.*, 2016; Grant *et al.*, 2020), I found that *env* is flanked by two short regions of frequent recombination -*TAT/VPU* and the C1 region of the 5' end of the gene and *REV/TAT* exons at the 3' end of the gene – while recombination occurs at a consistently low rate in the interior region.

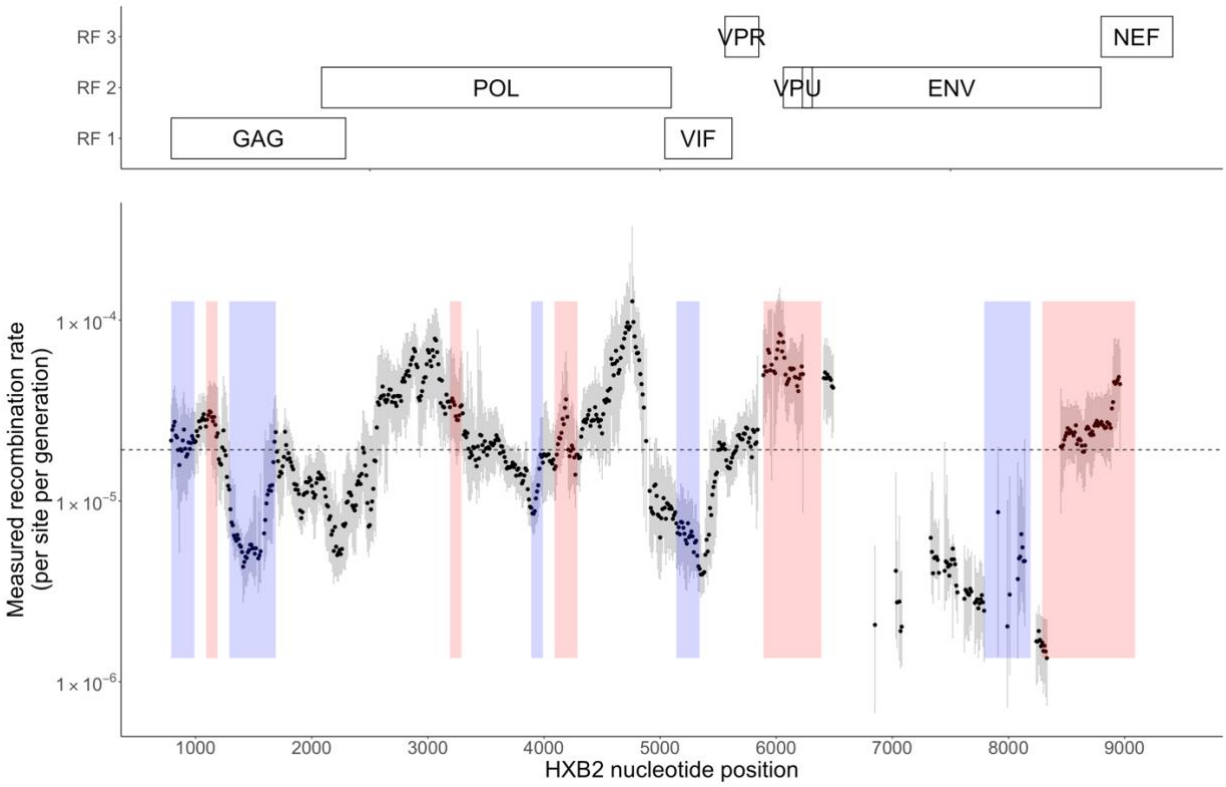

Figure 4.3 Recombination rate /site/generation for overlapping sliding windows of length 500bps, moving in increments of 10bps. Genes and reading frames are described above. Shaded grey bars represent a 95% confidence interval calculated by 100 bootstrap replicates. Windows in which variation across bootstrap replicates was extremely large were discarded, defined as when the log-transformed ratio of the upper and lower confidence interval bounds was lower than 0.8. Poor convergence across bootstraps was most likely caused by proximity to both strong hot and cold spots within a single window, too few pairs of sites or alignment issues masking recombination signal. The vast majority of discarded window were located in ENV. Regions shaded in red or blue represent previously identified hot or cold spots for inter-subtype recombination in unique recombinant forms and circulating recombinant forms identified in Africa (Jia *et al.*, 2016), for which we observe a correspondence between hot and cold spots with peaks and troughs respectively.

Analyses of breakpoints in URFs and CRFs have produced descriptions of the distribution of hot and cold spots across the genome at the inter-subtype level, however it remains unclear whether inter-subtype breakpoints align with the recombination hot and cold spots observed within a single-virus infection. Using published data, I compared windows with high and low numbers of breakpoints identified in CRFs and URFs circulating in Africa to the patterns observed in our analysis (Jia *et al.*, 2016), and found strong concordance (figure 4.3). All five inter-subtype hot spots identified by Jia et al exhibited above-average recombination rates in our analysis, and four of the five cold spots had below average rates (Table 4.2).

Table 4.2 The recombination rate per site per generation for hot and cold spots identified in unique recombinant forms and circulating recombinant forms seen in Africa (Jia *et al.*, 2016). Rates in red indicate a value greater than the average (adjusted for the length of the region) and cold indicates a value lower than average. For window 790-990, the model did not return a positive rate, likely due to poor model fitting as a result of minimal evidence of recombination.

HXB2 coordinates	Length (bps)	Recombination rate
Hot spots:		
1090 - 1190	100	**11.3 (CI: 5.9 – 19.0)**
3190 – 3290	100	**8.0 (CI: 5.2 – 15.1)**
4090 – 4290	200	**4.5 (CI: 2.4 – 9.8)**
5890 – 6390	500	**4.8 (CI: 3.0 – 8.3)**
8290 - 9090	800	**4.7 (CI: 2.8 – 6.3)**
Cold spots:		
790 – 990	200	**NA**
1290 – 1690	400	**0.4 (CI: 0.25 – 0.55)**
3890 – 3990	100	**6.7 (CI: 3.2 – 10.03)**
5140 – 5340	200	**0.9 (CI: 0.57 – 1.66)**
7790 - 8190	300	**0.19 (CI: -0.05 – 1.48)**

4.4.5 Subtype-specific patterns of recombination

The genome fragments contributing to CRFs and URFs have been shown to vary across subtypes, suggesting a degree of variation in the breakpoint distribution between subtypes. Biologically, it is unlikely that co-infection rates or frequency of template switching alone varies by subtype, and differences in the observed *in-vivo* recombination rates may arise due to varying degrees of diversity or disparities in selection pressures. To account for the possibility of viral load variation influencing inter-subtype variation, I matched datasets filtered by subtypes (C and D) to the viral load distribution of subtype A *D'* ratios. The highest genome-wide recombination rate was observed in the subtype C dataset (3.64×10^{-5} ($\mathrm{CI}: 2.08 \times 10^{-5} - 5.34 \times 10^{-5}$) /site/generation), followed by subtype A (1.56×10^{-5} ($\mathrm{CI}: 1.28 \times 10^{-5} - 1.93 \times 10^{-5}$) /site/generation), while subtype D has the lowest rate (0.81×10^{-5} ($\mathrm{CI}: 0.56 \times 10^{-5} - 0.13 \times 10^{-5}$) /site/generation) (figure 4.4A).

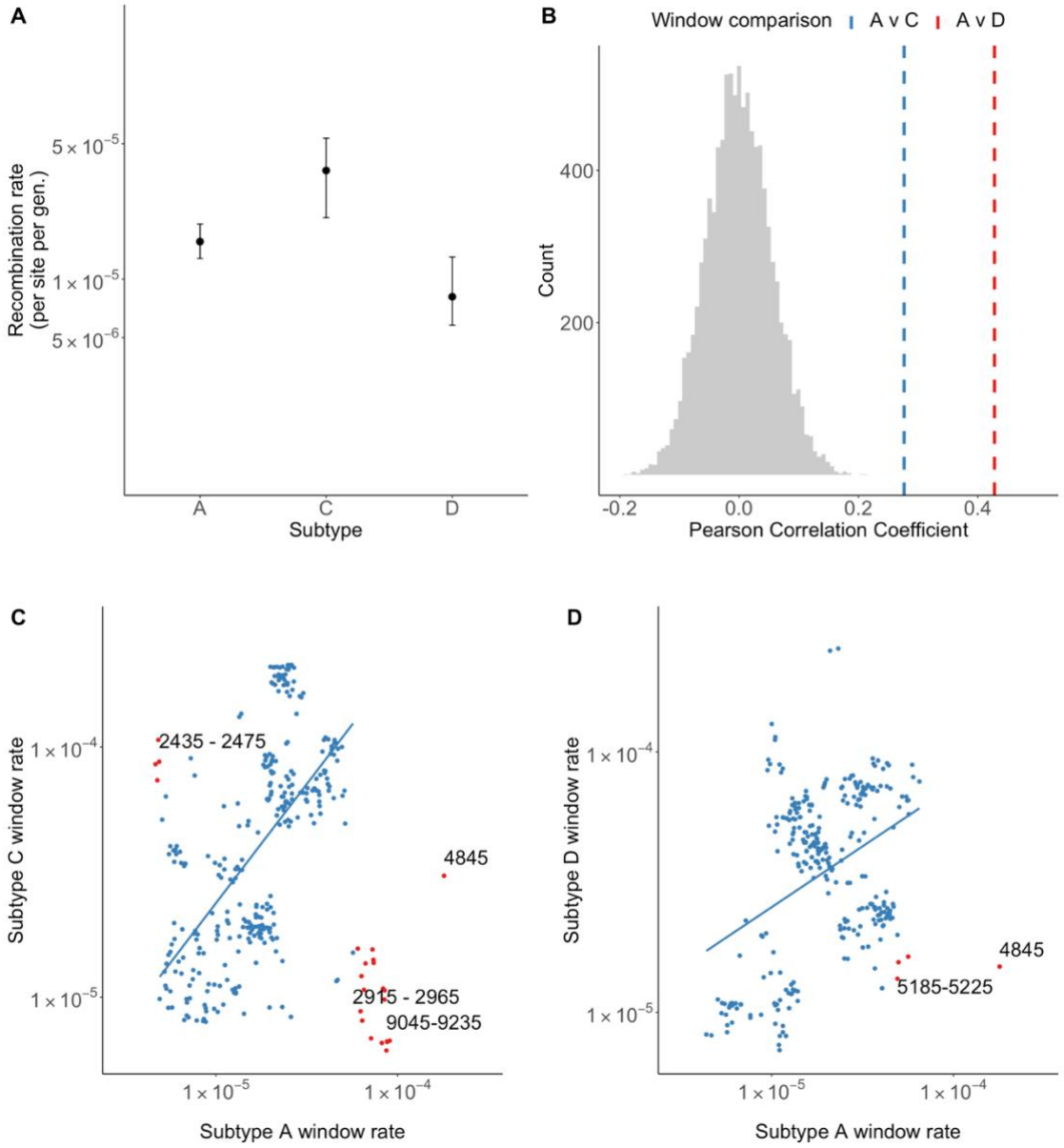

Figure 4.4 Variation in recombination rates by subtype. A) The average recombination rate for subtypes A, C and D. Error bars represent a 95% confidence interval calculated by bootstrapping. Subtype C and D datasets are subsampled by viral load tertiles in order to balance differences. Subtype C has the highest observed rate, and subtype D the lowest. B) Correlation test results for comparison between subtype A and C and subtype A and D. The histogram represents correlations of 10,000 replicates of a permutated dataset, and dotted lines represent the true correlations. C-D) Red points indicate outliers that are determined by standardised residual size (outlier has a value greater than 2), with labels denoting the midpoint of the window. Windows in close proximity are labelled together as a sequence of coordinates. The slope signifies a linear regression line with outliers removed. For subtype comparisons, window lengths of 750bps were applied to the sliding windows to ensure windows included a sufficient number of data points to fit the model.

Next, I repeated the sliding window analysis on subtype filtered datasets, discarding any windows for which the bootstraps did not converge. A lack of convergence is most

likely a result of too few data points or windows containing both hot and cold spots. To assess the similarity in the location of hot and cold spots we measure the correlation between the median rate of each window for subtype A versus C and subtype A verses subtype D. To assess the significance of the result, I performed a permutation test where window coordinates of one subtype are shuffled, the correlation coefficient is then calculated, and the process is repeated 10,000 times. For both comparisons I find the true correlation to be higher than the extremes of the distribution of correlations from the permutated datasets (figure 4.4B), signifying that local changes in recombination rates are broadly similar across subtypes.

By performing a linear regression by window-matched rates and calculating standardised residuals, I identified outliers that indicate any hot or cold spots that are subtype-specific (figure 4.4C-D). For example, a hot spot located between 2435-2475 (HXB2 coordinates) is identified in subtype C and not subtype A. Interestingly, the majority of the variation across the subtypes is in located in POL.

4.4.6 Recombination patterns supported in Illumina pipeline sequences

I repeated the sliding window analysis on the illumina dataset. The *D'* ratios in the illumina dataset are derived from sequences processed via the illumina short-read pipeline. Approximately a third of the infections in the illumina dataset also feature in the PacBio dataset, and the two datasets have an overlap of 58,305 *D'* ratios (25% of the illumina set, 10% of the PacBio dataset). To determine whether the patterns across the genome, I repeated the permutation test performed previously for a subtype comparison. The correlation between windows was 0.43, which was consistently greater than any correlation coefficient derived from correlations inferred from 10,000 permutated datasets. The strong positive relationship between corresponding windows of recombination rates in the two datasets provides evidence of a 'universal landscape' for hot and cold spots for recombination that is independent of the sequencing pipeline chosen (Figure 4.5A). In addition, I identified the strongest outliers from the relationship. Outliers were classified as points with standardised residuals in a linear regression with a value greater than 3. Outliers were located in regions where we found rapid changes from low to high frequency, specifically windows with start coordinates in the 8310-8710 region incorporating both regions of low recombination in GP41 and regions featuring high recombination located at the end of the ENV gene and start of NEF.

Differences by source and recipients are also evident, and the ordering of subtype average rates remains the same, albeit the confidence is lower due to a reduced number of *D'* ratios per group (figure 4.5B-C). Finally, we again see an association between increasing viral load and recombination rate when we compare low and high viral loads (defined as an average viral load lower or greater than 5.4 ($\log_{10}$ copies per ml) to make the two groups include the same number of *D'* ratios) (figure 4.5D).

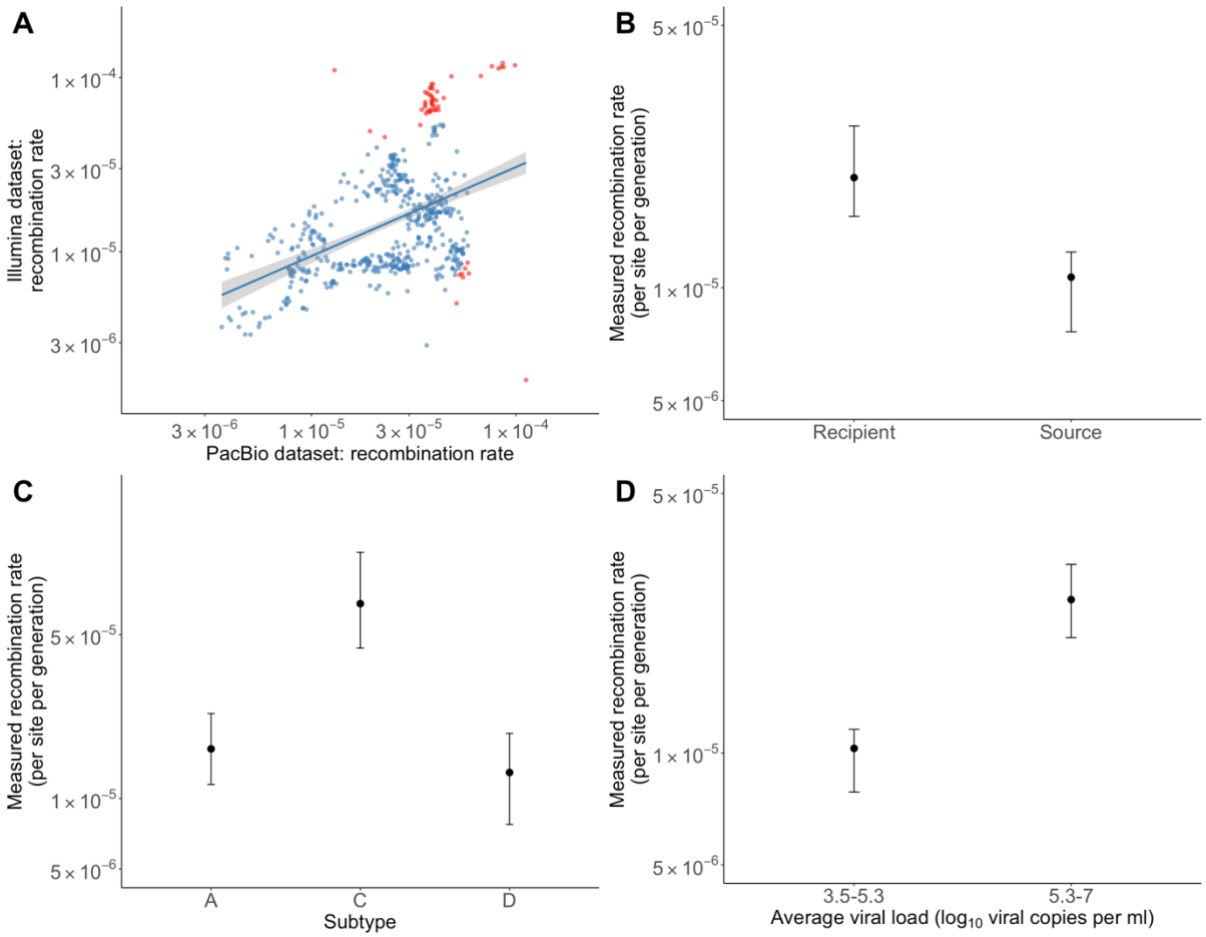

Figure 4.5 Findings are supported by Illumina dataset. A) Histogram of correlation coefficients between sliding windows of PacBio and Illumina data with the latter's window coordinates shuffled. The black dotted line indicates the true correlation coefficient. For sequence platform comparison, window lengths of 750bps were applied to the sliding windows for both datasets to ensure windows in the Illumina set included a sufficient number of data points to fit the model, as fewer pairs of diverse sites were identified as a consequence of short Illumina reads. B) Boxplot of bootstraps of recombination rates by recipient versus source C) Boxplot of bootstraps of recombination rates for each viral load tertile. D) The average recombination rate by subtype. Bars indicate 95% confidence interval determined from bootstrapped replicates.

4.4.7 Recombinants selected for in *pol* but against in *env*

Selection and recombination are intrinsically linked processes: a recombination event that produces combinations of mutations that lower viral fitness will be selected against, and the effective recombination rate will be reduced. To investigate the interplay between recombination and selection, I determined the gene-level (GAG, POL, ENV (GP120)) recombination rate from codon partitioned data, excluding any sites within overlapping reading frames (figure 4.6). For ENV, due to overlapping reading frames in *GP41*, I only considered GP120. In GAG there was minimal difference between codon-partitioned estimates. In POL, the rate was 3.5 times higher in pairs of sites both situated in codon positions 1 and/or 2 compared to pairs of sites both at codon position 3, signifying possible positive selection of recombinant forms. In ENV, the difference was extremely large and in the opposite direction, with pairs at codon position 3 having a rate approximately 30 times greater than pairs at positions 1 and 2. This indicates strong epistasis between non-synonymous mutations that conserve linkage and prevent the emergence of recombinants in ENV.

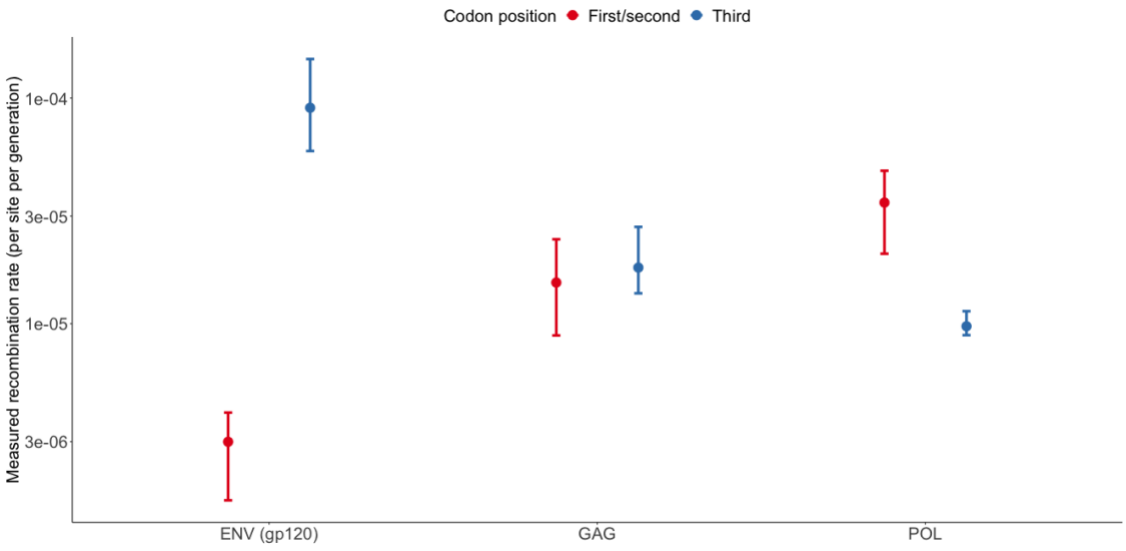

Figure 4.6 Substantial differences in recombination rate in ENV by codon position. The recombination rate is estimated separately by codon position (1 and 2 vs 3) for non-overlapping regions of ENV, GAG and POL. For example, pairs of sites included in the 'first/second' group are restricted to pairs where both sites at codon position 1 or 2. Error bars represent 95% confidence intervals generated by bootstrap replicates. Mutations at third codon positions are used as a proxy for synonymous changes and first/second for non-synonymous. A lower rate of measured recombination in pairs of sites at first/second positions signifies selection against recombinants.

I investigated whether the differences in the rates can be explained by rapid changes in allele frequency affecting the measurement of LD. I found that the differences cannot be explained by strong selection acting on specific sites, as the directions of the differences in rates are preserved when we only consider sites at which the change in allele frequency is low, defined as lower than 10% change in absolute frequency between time points (supp figure 9.5).

4.5 Discussion

In this study, I employed a time-series linkage decay method on a large cohort of individual infections to investigate the patterns and mechanisms driving recombination across the viral genome. The analysis focused on how viral and environmental factors influence the extent to which recombination shapes the viral population. By analysing overlapping windows across the genome, I identified both local and global recombination hot and cold spots, uncovering evidence of a 'universal landscape' of recombination breakpoints that operates at both intra- and inter-subtype levels. Furthermore, I validated our findings by analysing sequences derived from a different sequencing platform and a largely independent cohort of infections, confirming that our results were not artifacts of sequencing bias or noise.

It has been previously demonstrated that hot and cold spots for recombination exist: consistent hot and cold spots within the POL and GAG regions have been reported across multiple infections (Smyth *et al.*, 2014), and recombination hot spots have been repeatedly described in studies of inter-subtype recombination (Magiorkinis *et al.*, 2003; Fan, Negroni and Robertson, 2007; Jia *et al.*, 2016; Grant *et al.*, 2020) and in experimental settings (Zhuang *et al.*, 2002; Simon-Loriere *et al.*, 2010). Recombination patterns are therefore likely to be intrinsic to the genome structure and governed by underlying mechanisms that operate broadly across individuals. I provided strong evidence that recombination breakpoint locations are indeed systematic, with significant correlations in window-specific rates observed across subtypes and sequencing platforms, implying consistency in regions of high and low recombination. I also confirmed that regions identified as recombination hot or cold spots at the inter-subtype level (Jia *et al.*, 2016) also exhibit similar patterns within individual hosts infected with a single subtype. This finding suggests that the formation of URFs and CRFs is at least partially driven by within-host recombination dynamics. Consequently,

it may be possible to predict future emerging recombinant forms based on within-host recombination patterns.

By considering variation across the entire genome with a sliding window approach, I have identified novel hot and cold spots, including a strong hot spot within the p31 region, which encodes the integrase protein in the POL gene. The protein is of significant clinical importance, as it is the target of integrase strand transfer inhibitors (Smith *et al.*, 2021). Our study cohort are treatment naïve, however this finding suggests a natural susceptibility to recombination in this region, which could increase the likelihood of drug resistance mutations emerging over time.

The measurable frequency of recombination events in natural infection partly reflects evolutionary processes - primarily selection pressure - and rate variation within and between genes is very likely influenced by the fitness of recombinant genomes. In agreement with previous studies, I observed a low rate of measurable recombination in GP120, a protein that is required for cell entry and recombination is likely to be suppressed due to functional constraints (Simon-Loriere *et al.*, 2009). Additionally, interdependencies between GP120 and GP41 proteins have been said to lower the fitness of recombinant genomes (Golden *et al.*, 2014; Woo, Robertson and Lovell, 2014; Bagaya *et al.*, 2015). By partitioning sequence data by codon position, I showed linkage was much more rapidly broken down between pairs of third codon position sites (synonymous) in comparison to first and second position pairs (non-synonymous). Infrequent recombination in ENV is therefore likely the result of strong epistasis and selection against recombinants, emphasising the disruptive potential of recombination in virus adaptation and a probable large discrepancy in the true frequency of recombination with measurable recombination *in-vivo,* as well as the role of interacting and interdependent mutations in the evolutionary success of mutations in ENV. More broadly, I hypothesise that the reduction in the measurable recombination rate in individuals sampled later in infection reflects the interaction between recombination and selection. I propose that the virus, having become well-adapted to the host environment after lengthy infection, rapidly purges deleterious recombinants, hence lowering the measured rate of recombination.

The study is limited by potential noise in the data and alignment issues, however verification with a dataset comprised of sequences generated by a distinct sequencing pipeline provides confidence that the findings are not substantially affected by noise

for the majority of the genome. The method did not perform well in the variable loops, which may be a consequence of alignment errors, thus limiting the conclusions we can draw on variation within the ENV gene, specifically GP120. Additionally, regions of high or low rates of recombination rates are limited to a minimum window span of 500bps in this analysis, narrowing our ability to pinpoint specific breakpoints.

This study is unique due to the large sample size, the use of long-read deep sequencing from the PacBio platform and the replication of results with a dataset built on sequences generated on the Illumina platform. With LD measures from sequences outputted from long-read PacBio and short-read Illumina platforms, I have shown that key findings are not a consequence of the sequencing process or noise, and that they are replicated in a dataset of pairs of sites with minimal overlap and a largely separate set of infections. By leveraging the read length of PacBio sequences, I considered the effect that distance and time has on recombination estimation, as previous proposals for the range at which linkage is observed (100-200bps) have been limited by read length and relatively sparse sampling (Zanini *et al.*, 2015). By applying the method to simulated data, we found that the required read length for accurate rate estimation depends on sampling frequency and the true recombination rate, with more frequent sampling necessitating longer reads to accurately capture linked mutations before recombination occurs. In the empirical dataset, LD continued to decay with increasing genetic distance up to 500 bps, indicating that LD can be seen to persist over longer distances if sampling is sufficiently frequent. Consequently, careful consideration of the necessary read length based on the sampling regime is crucial for analysing linkage or conducting phylogenetic analyses, and comparisons of LD patterns across different datasets should be made cautiously. Nonetheless, when controlling for time-scaled distance, LD decay methods can still provide valuable insights into the viral and host factors that influence the effective recombination rate.

In summary, the findings of this study reveal substantial variation in recombination rates across the genome that are consistently replicated across subtypes, sequencing platforms, and at the inter-subtype level, suggesting an intrinsic viral property driving recombination dynamics. The contribution of recombination to within host populations is shown to be strongly tied to selection processes, emphasising the need for *in-vivo* studies and highlighting the role of strong epistasis in ENV. Ultimately, understanding

the recombinogenic properties of HIV and recombination dynamics can guide the development of treatments and vaccines that avoid viral escape strategies.

5 The evolutionary rate of HIV at the within and between host scale

5.1 Abstract:

HIV evolutionary rates are consistently 2-5 times higher within hosts compared to between hosts. While several explanations exist, including the transmission of ancestral virus variants and ongoing cycles of immune escape and reversion, previous studies have been limited in size or genome region. Using longitudinal sampling of a large cohort during early infection and population-level consensus sequences from sub-Saharan Africa, I applied multiple methodological approaches (BEAST and divergence) to quantify evolutionary dynamics across genomic regions and ecological scales. While evolutionary rates differed between scales, I found strong correlation in patterns of sequence divergence across the genome at both within-host and between-host levels, suggesting consistent evolutionary constraints operate from early infection through long-term epidemic spread. Different methodological approaches to measuring evolution yield substantially different rate estimates, which I propose is driven by "toggling" mutations - a dynamic that proves pervasive across both synonymous and non-synonymous sites and may explain the disparity in rates across ecological and temporal scales.

5.2 Introduction

HIV evolves rapidly and within-host dynamics are characterised by remarkable levels of diversity and divergence during the course of a single infection (Shankarappa *et al.*, 1999). The rapid and continuous creation of viral mutants provides a major advantage for the long-term survival of the virus, as extensive diversity mediates escape from humoral and cellular host immune responses that have been observe to emerge throughout infection (Phillips *et al.*, 1991; Wei *et al.*, 2003; Jones *et al.*, 2004; Bernardin *et al.*, 2005; Liu *et al.*, 2006; Goonetilleke *et al.*, 2009) and creates the potential for the rapid emergence of drug-resistance (Little *et al.*, 2008; Hedskog *et al.*, 2010). Despite this, between-host evolution and the diversification of HIV at an epidemiological scale shows evidence of neutral evolution (Lemey, Rambaut and Pybus, 2006; Theys *et al.*, 2018), with the co-circulation of many subtypes and relatively low rates of drug resistance emerging over time (Theys *et al.*, 2018). The tempo of evolution has been observed in multiple datasets to be substantially slower

when measured at the epidemiological scale compared to over the course of a single infection (Lemey, Rambaut and Pybus, 2006; Alizon and Fraser, 2013; Novitsky *et al.*, 2013; Raghwani *et al.*, 2018). Unravelling the mechanisms behind the mismatch in evolutionary rates across scales and understanding how within-host evolutionary processes manifest at a population level are critical for the development of treatments and vaccines, predicting the spread of immune escape and drug-resistant mutations, and may offer insights into more general features of pathogen dynamics for other understudied chronic RNA viruses.

The importance of selection at the within-host scale on how evolution proceeds between hosts depends upon viral and host factors. First, HIV has a severe bottleneck of typically one viral genome that limits inherited genetic diversity (Carlson *et al.*, 2014). If all circulating variants at the point of transmission had equal chance of establishing onward infection, we would expect the evolutionary rate of neutral or nearly-neutral mutations across scales to be of similar magnitude. It has therefore been suggested that the lower rate between hosts is because not all viral genomes have an equivalent chance of successfully transmitting. Indeed, some viral phenotypes have been linked with increased probability of transmission, such as use of the CCR5 co-receptor for cell entry and envelopes with lower levels of glycosylation (Dragic *et al.*, 1996; Chohan *et al.*, 2005; Checkley, Luttge and Freed, 2011). Transmission fitness has also been associated with genetic similarity to the transmitted/founder (T/F) virus in the source infection, which is made possible by ancestral virus circulating within the latent reservoir (Redd *et al.*, 2012).

The tempo of evolution may be quantified by measuring the number of nucleotide or amino acid substitutions that accumulate per site per unit time. This is known as the molecular clock. Methods to quantify the rate at which the clock ticks vary in their complexity and statistical sophistication. At the most basic level, genetic divergence can be measured over time from a putative founder sequence. More advanced phylogenetic approaches, such as root-to-tip regression analysis in maximum likelihood trees, can account for the underlying substitution process. However, these methods are sensitive to root placement, and uncertainty in tree topology and branch lengths cannot be easily quantified. More sophisticated approaches have been developed within a Bayesian framework, including phylogenetic methods that explicitly model the molecular clock during tree construction. Bayesian Evolutionary Analysis

Sampling Trees (BEAST) represents a powerful, though computationally intensive, approach for estimating evolutionary rates (Suchard *et al.*, 2018). This method allows rates to vary across branches of the phylogenetic tree and provides a framework for quantifying uncertainty in both the tree topology and rate estimates through the posterior distribution. Rate estimates can differ substantially between these methods due to their varying abilities to capture different aspects of the evolutionary process, for example multiple substitutions at the same position or parallel evolution across lineages.

The disparity in clock rates across scales was first documented by Lemey, Pybus, and Rambaut in a single individual, focusing on the C2V5 region of the genome (Lemey, Rambaut and Pybus, 2006). Their initial observation, based on a Bayesian time tree analysis of sequences sampled over a decade, revealed that the molecular clock rate was approximately twice as high within-host compared to between-host, even when comparing sequencing data collected over similar timescales. Alizon and Fraser (2013) later expanded on the result by studying within-host sequences from five infections, though comprehensive genome coverage was only available for one chronically infected individual. Their analysis confirmed consistently higher within-host evolutionary rates across the entire genome, with the most pronounced difference observed in the envelope gene. A subsequent 2013 study of within-host evolution during primary subtype-C infection further described within-host clock rates and genome rate variation in a larger cohort, focusing on ENV and POL regions (Novitsky *et al.*, 2013). The within-host rates were higher than previously reported, likely to more frequent sampling that captured transient mutations. Notably, the elevated rate of within-host evolution has been demonstrated for both non-synonymous and synonymous mutations, with the latter typically assumed to evolve under neutral conditions (Lemey *et al.*, 2007; Abecasis, Vandamme and Lemey, 2009; Lythgoe and Fraser, 2012; Raghwani *et al.*, 2018). Differences in evolutionary rates have also been observed in other chronic viruses, specifically HCV, suggesting underlying drivers that have more general applicability to chronic viruses (Raghwani *et al.*, 2019).

Collectively, these studies provide strong evidence that HIV evolves 2-5 times faster within hosts compared to the population scale. However, the majority of studies have been limited to reasonably small cohorts, making comparisons across individuals and subtypes challenging. Earlier studies typically included a small number of sequences

per time point, missing the full diversity of viral populations. With the increasing use of next-generation sequencing technologies, it has been possible to describe within-host populations in much greater detail, with deep-reads and the detection of minor variants (Zanini *et al.*, 2015). However, more recent studies have been limited by short-read data, affecting insights into genetic linkage. There are also significant methodological differences across studies, including:

1. Genome regions covered (e.g., full genome, specific genes like env or pol)
2. Sampling frequency and timescales (ranging from weeks to years)
3. Sequencing methods (Sanger, NGS platforms, single-genome amplification)
4. Phylogenetic approaches (Bayesian, maximum likelihood, genetic divergence)
5. Evolutionary models (strict vs. relaxed molecular clocks, demographic models)

These variations in methodology, combined with differences in cohort sizes and characteristics, make it difficult to directly compare and collate estimates across studies.

Several hypotheses have been proposed for the mechanisms that explain the evolutionary rate mismatch, however unanimous agreement over the primary drivers has not been reached. The hypothesised mechanisms are not mutually exclusive, and the relative roles may differ across genomic regions. First, is the “stage-specific selection” hypothesis that argues that selection is weaker during early infection and intensifies during chronic infection (Pybus and Rambaut, 2009). If transmission typically occurs early in infection, then evolution measured across multiple infections will be lower. While supported by evidence of differences lower in populations where transmission typically occurs early in infection, specifically injecting-drug users (Maljkovic Berry *et al.*, 2009), it is unlikely to contribute significantly to the rate mismatch given the repeated observations of early diversification and rapid evolution in response to cellular and humoral immune response during early infection.

The second hypothesis, termed "escape and revert," posits that adaptive mutations selected in response to the host's cellular immune response are highly individualised based on the host's HLA background and often carry intrinsic fitness costs (Leslie *et al.*, 2005; Brumme *et al.*, 2009; Carlson *et al.*, 2015; Kløverpris, Leslie and Goulder, 2016). After transmission to a new host, these adaptive mutations may revert to their

ancestral states (Leslie *et al.*, 2004; Herbeck *et al.*, 2006; Boutwell *et al.*, 2010; Zanini *et al.*, 2015). Support for this hypothesis is bolstered by observations that evolution tends to favour subtype-specific population-level consensus sequences, which may represent a 'universally fit' genome (Zanini *et al.*, 2015). More recently, a study on within-host selection and dynamics of haplotypes showed evidence of the escape and revert dynamic also affecting antibody sites . Nonetheless, "escape and revert" does not adequately explain the elevated rates of synonymous mutations that are not under selection for CTL escape or account for the overall elevation in evolutionary rates observed across the entire genome, including non-immunogenic regions. Please see chapter 6 for an in-depth review of the escape and revert mechanism.

The final hypothesis suggests that ancestral viruses present during early infections are more likely to be transmitted, termed “store and retrieve” (Lythgoe and Fraser, 2012). This could be due to inherent properties of T/F viruses that enhance their fitness for establishing subsequent infections. Studies have indicated lower evolutionary rates in branches associated with transmission events and provide evidence supporting the transmission of viruses present early in source infections (Vrancken *et al.*, 2014, 2015). Store and retrieve is considered the most parsimonious explanation, as both non-synonymous and synonymous mutations across the entire genome are affected. However, an analysis of within-host evolution in a cohort of untreated individuals in Uganda found that the size of the transmission advantage must be large to explain the observed evolutionary mismatch (Raghwani *et al.*, 2018), yet recent analysis of deep-sequenced long-reads of a large cohort of transmission pairs did not find evidence for the transmission of ancestral viral variants (Zhao *et al,* unpublished).

Beyond these three main hypotheses, the time-dependent rate phenomenon provides additional insight into the evolutionary rate mismatch. This phenomenon describes how the molecular clock rate decreases as the measurement time-scale increases due to saturation (Ho *et al.*, 2011; Duchêne, Holmes and Ho, 2014; Aiewsakun and Katzourakis, 2016). Recent work by Druelle and Neher (2023) demonstrated saturation of divergence over the course of the pandemic, linking this pattern to the preference for reversions of host-specific adaptations to population-level consensus sequences. Additionally, rates measured over shorter time scales are naturally elevated due to the detection of weakly deleterious mutations that are later removed by purifying selection. Both time-dependent selection and transiently beneficial

mutations – including both synonymous and some non-synonymous changes – have been identified from longitudinal within-host data of the C2-V5 region of the ENV gene (Zanini and Neher, 2013). Consequently, measurements taken over longer time scales may fail to capture the full complexity of evolutionary dynamics and fitness landscapes.

Underpinning our understanding of the mechanisms driving short- and long-term rates of viral evolution is the need for a comprehensive description of within-host divergence and evolution across the genome, individuals, and subtypes. In this study, I investigate the evolutionary dynamics during the first 12–24 months of untreated infection in 89 individuals longitudinally sampled in Sub-Saharan Africa. Leveraging deep-sequencing with long-read next-generation sequencing technology, our dataset is characterized by more frequent sampling than most comparable studies, providing a unique opportunity to track viral evolution with high temporal resolution. I examine variation across the genome, identifying regions with evidence of a molecular clock signal, and generate BEAST phylogenetic trees to quantify evolution.

This analysis compares two different methodological approaches: a population genetic method and a phylogenetic approach. I explore the drivers of variation in evolutionary rates across infections, including associations with viral load, correlations across the genome, and the influence of the sampling timeframe. In addition, I investigate the differences between these methods, particularly highlighting the role of transient mutations in elevating rates estimated by BEAST. Finally, I show that within-host evolutionary dynamics are characterised by a high turnover of mutations—what I term "within-host toggling"—reflecting the transient rise and fall of many mutations, including a significant proportion of non-synonymous changes. The toggling dynamic can help explain the overall evolutionary rate mismatch, why the mismatch is larger and when sampling is denser, and affects both synonymous and non-synonymous sites.

5.3 Methods

5.3.1 Study Population

All individuals whose samples were analysed were enrolled in one of three African HIV serodifferent heterosexual couples studies outlined in Chapter 2. As I am interested in evolution during the first 1-2 years of infection, I only consider individuals who seroconvert during the course of the study (N=278). The time since infection for

recipient individuals was taken as the mid-point between the last HIV- negative date and first HIV-positive date, which were on average (median) 84 days apart. Individuals where the first sampling points was over three months post the estimated time of sero-conversion were excluded from the analysis (N=10 excluded).

5.3.2 Sequence processing and filtering

Details on sample preparation, sequencing and initial processing steps are provided in Chapter 2. Here, I utilise PacBio sequences as they have the advantage of longer reads. All individuals included in the study had at least three longitudinal samples that have been sequenced on the PacBio pipeline and subsequently processed with Shiver (N= 107). The Phyloscanner software was then used to generate sliding windows of alignments for each individual, with windows of length 500bps and an overlap of 490bps to neighbouring windows. The length was chosen due to high recombination significantly minimising linkage at distances greater than this length as described in chapter 4, and so the 500bp length was chosen as a compromise between capturing the evolution of haplotypes with the increased impact of recombination across longer genomic distances. The reads of all windows were aligned to HXB2 positions to allow for comparison across datasets, and therefore insertions relative to HXB2 were excluded from the analysis. For each window, the minimum number of reads at each time point is set at 10, and any individuals for which phyloscanner windows following filtering did not cover 50% of the genome were excluded from the analysis (N=13 excluded).

For the subtype-specific analyses, only subtypes A1, C and D were included as other subtypes were insufficiently represented. Samples where the best reference subtype was not consistent across samples were not included in the subtype-specific datasets.

As well as quantifying evolutionary rates, I also studied the dynamics of minor variants contributing to genetic divergence over time. I tracked allele frequencies using base frequency files generated by the shiver pipeline. To improve accuracy in alignments across timepoints, I first generated a consensus sequence for each timepoint by taking the majority allele at each genomic position. These per-timepoint consensus sequences were then aligned together with the HXB2 reference sequence using MAFFT (Katoh, 2002), creating a master alignment that allowed us to map equivalent positions between timepoints. I was primarily interested in mutations that emerge

within the recipient, termed de novo mutations, which we define a variant that initially has a frequency lower than 5%. For these minor variants, I classified them as non-synonymous if they would cause an amino acid change when combined with the major alleles at the other two codon positions. If no amino acid change would result, the variant was classified as synonymous.

To control for potential dual infections, which could confound evolutionary rate estimates, I implemented a diversity filter, where initial diversity was defined as the average per-read hamming distance to the consensus sequence at the first time point. Individuals showing unusually high initial diversity (>0.1, representing two standard deviations above the mean) were excluded from further analysis, resulting in the removal of five individuals whose viral populations suggested possible dual infection or transmission of multiple viral variants.

Following all filtering steps, 89 individuals were included in the main dataset.

5.3.3 Evolutionary Analyses

Within-host divergence estimation

The genetic divergence of the within-host viral population was tracked over time using sliding windows across the genome. At each timepoint, divergence was calculated by comparing individual sequence reads to the consensus sequence from the first timepoint, which serves as a proxy for the founder virus. The divergence at time t is defined as,

$$\text{divergence}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{d(r_i, F)}{L}$$

Where N is the number of reads, L is the read length and $d(r_i, F)$ is the hamming distance between read r_i and the founder sequence F, for which the consensus sequence at time $t = 0$ is a proxy. At time $t = 0$ the divergence estimate is equivalent to diversity. To determine an estimate of noise in the data, I determined the average diversity across all windows in individuals where the first sample was dated within 2 weeks of the estimated date of seroconversion, which included individuals mainly included in the PrEP study that tested HIV negative individuals monthly. Using data from individuals sampled within 2 weeks of seroconversion, representing the PrEP study participants with monthly HIV testing, I established that initial diversity was

relatively low (mean 0.37%, median 0.28% (IQR: 0.21%-0.46%), equivalent to fewer than two differences per read). This initial diversity measure was used as a “noise” threshold in subsequent analyses. Bootstrapping across windows showed minimal effect on the initial diversity average.

To determine a divergence rate for a window, a standard linear regression was fitted to the data, where the slope coefficient was taken as the divergence rate. The initial diversity, i.e. the first time point, was included in the fit, to account for initial noise and background diversity, and time, t, was measured relative to the first sampling point.

To distinguish between selective and neutral evolution, separate calculations were performed for different codon positions: positions 1 and 2, where mutations typically cause amino acid changes (non-synonymous), and position 3, where mutations are often silent (synonymous). The codon position analysis excluded genome regions with overlapping reading frames.

Molecular clock analysis

To estimate evolutionary rates at both within-host and between-host scales, I followed the methodological framework outlined by Alizon and Fraser (2013). Before conducting molecular clock analyses, I first assessed whether our sequence data contained sufficient temporal signal. For within-host analysis, I evaluated the relationship between genetic divergence and time since infection using both divergence rate measures and a systematic approach to handle overlapping genomic windows.

For each genomic window in each individual, I applied three filters: (1) the maximum divergence over the sampling period must exceed the estimate of background noise (defined as the observed diversity at the earliest time point), (2) the rate of divergence must be positive, consistent with the expectation of accumulating mutations over time; and (3) the relationship between divergence and time must show strong temporal structure, as measured by an adjusted R-squared value exceeding 0.8. A higher threshold of 0.9 was also considered, however differences in the final filtered dataset were minimal.

After applying the initial filters, I addressed the overlap between adjacent windows to prevent redundancy in our final dataset. I applied a greedy algorithm to select the most informative non-redundant set of windows: for each individual, windows were first

ranked by their adjusted R-squared values, and proceeding from highest to lowest, a window was only retained if less than 10% of its sequence overlapped with any previously selected window. This approach ensures that each region of the genome is represented by its most clock-like window while avoiding oversampling of any particular region.

This systematic filtering process yielded 848 windows suitable for subsequent Bayesian Evolutionary Analysis by Sampling Trees (BEAST) analysis (Suchard *et al.*, 2018). These windows represent genomic regions with robust temporal signal and minimal redundancy. By only analysing regions with evidence of the molecular clock signal, I bias the analyses towards finding faster clock rates. However, this step is necessary to avoid invalidating model assumptions, and across all hosts we cover the entire genome.

5.3.4 BEAST Analysis

Data subsampling

Due to the computational demands of BEAST analysis, it is not possible to analyse all reads. I implemented a random sampling strategy for each genomic window, where at each timepoint, 30 sequences were randomly selected to represent the viral population diversity. All windows were trimmed to ensure alignment with codon positions, with the first nucleotide corresponding to the first codon position.

Substitution Model Selection

Model selection was performed using IQ-TREE's ModelFinder algorithm (Kalyaanamoorthy *et al.*, 2017) on a representative subset of genomic windows across multiple individuals. Specifically, one window was randomly selected for each individual. The HKY+F model was consistently selected as the best-fit model according to the Bayesian Information Criterion (BIC), which penalises model complexity more strongly than AIC. To assess model robustness, analyses were also performed on a subset of the data (10 randomly selected windows) using the more parameter-rich GTR+F+Γ4 model. Key parameter estimates, including evolutionary rates and tree topology, remained qualitatively unchanged between models. As the clock rate of first/second versus third codon position were modelled separately, model selection was performed on codon portioned data, however the result did not differ to that to unpartitioned data.

Clock Model Implementation

Based on extensive evidence supporting its suitability for HIV sequence evolution at both within- and between-host scales, I implemented an uncorrelated lognormal relaxed molecular clock. The appropriateness of this model choice was validated by examining the coefficient of variation (COV) posterior distributions. For nearly all windows, the COV posterior distribution excluded zero, providing strong support for rate heterogeneity across within-host and between-host trees.

Coalescent model selection

For the choice of coalescent model, I considered a constant population size, exponential growth and Bayesian Skygrid (Gill *et al.* 2013). In most cases for the within host model, the constant population model was selected by a Bayes Factor criterion and was therefore applied across all analyses. In BEAST, a Bayes factor test compares the marginal likelihoods of different coalescent models to objectively select the best-fitting demographic model (Baele et al, 2013). Bayes factor is calculated as a likelihood ratio, with values above 10 indicating strong support for one model over another

BEAST implementation

Time trees were inferred using BEAST with the default parameter settings, unless stated otherwise. For each window of alignments, BEAST analyses were run for 100 million MCMC generations, with parameters and trees sampled every 10,000 steps. A codon-partitioned model was implemented to account for the different evolutionary dynamics between codon positions. The first two codon positions were combined into a single partition due to their similar selective constraints, while the third codon position was treated as a separate partition, allowing for different substitution parameters, rate heterogeneity, and relative rates between partitions.

Convergence was assessed using LogAnalyser from the BEAST package, with analyses retained only if effective sample sizes (ESS) exceeded 200 for all parameters after discarding the first 10% of samples as burn-in.

Evolutionary rate estimation

For all successfully analysed windows, I report the mean value of the rateMean parameter across all MCMC samples as the overall evolutionary rate, after discarding the first 10% of samples as burn in. Additionally, for each analysis, I extract the posterior distribution of codon-specific rate variations.

I used BEAST TreeAnnotator to identify the maximum clade credibility (MCC) tree from the posterior distribution of trees generated by BEAST. This MCC tree represents the best supported tree topology from our analysis. Using this annotated tree, I then employed a custom R script to process the BEAST output and determine rates by branch time (external, internal, or backbone). External branch rates were calculated as a weighted mean of the rates for terminal branches, using branch lengths as weights as described in (Drummond *et al.*, 2006). For the backbone rate, I identified the monophyletic clade of the most recent time point and traced the path from its most recent common ancestor to the tree root, excluding branches that are common ancestors for only sequences sampled at the last time point. If the final time point was not monophyletic, the backbone rate was not calculated. I then calculated a weighted mean of the rates for these backbone branches. Internal branch rates were computed as the mean rate of all remaining internal branches not classified as backbone. Branch-averaged rates were typically higher than the rateMean, however the two rates were strongly positively correlated and inferences between the internal branch rate and variable (viral load, sampling time) apply to both measures.

I conducted BEAST analyses on 500bp windows rather than whole genes, focusing only on regions with clear molecular clock signal. To derive gene-level rates for each individual, I calculated weighted means and standard errors of the window-specific rates. I used inverse-variance weighting, with variances taken from the posterior distributions of the BEAST rateMean parameter, giving more weight to more precisely estimated rates. In any comparison featuring both divergence and BEAST evolutionary rates, the same subset of windows was included. Similarly, the average relative rate of the codon partitions was determined by invariance weighting.

5.3.5 Between-host dataset and analysis

A between-host dataset was generated from the Los Alamos National Laboratory sequence database (https://www.hiv.lanl.gov/, accessed November 2024). All

available sequences dated up until 2015 were collated from the database of sequence alignments for the entire genome, and for GAG, POL and ENV (GP120, GP41) genes separately, with GP120 and GP41 analysed separately. Each set of sequences was filtered by subtype to generate alignments of subtype A1, C and D. To increase the molecular clock signal in the dataset, I subsampled data to ensure an approximately even distribution of samples over calendar time. A maximum likelihood tree was fitted to the sampled sequences using IQ-TREE (Minh *et al.*, 2020) which was then re-rooted to maximise the correlation between sampling data and distance from the root. After re-rooting, any outlier sequences in the root-to-tip analysis (residual greater than 3 times the mean) were removed. The root of the re-rooted tree was then inferred with IQ-TREE to generate a reference sequence from which between-host evolution was measured from, where divergence was defined as the hamming distance (number of differences) between the inferred root sequence and the subtype reference sequence. Divergence rates were then derived by a linear regression of the divergence measure over calendar time.

For BEAST analyses, the best substitution model was chosen by IQ-TREE's ModelFinder algorithm, where a GTR+F+4+I was the selected model for all subtypes and for GAG, POL and ENV (GP120, GP41) genes. As with the within-host dataset, a codon partitioned model was chosen, as well as a relaxed log-normal molecular clock, where the CoV posterior distribution showed supported for the relaxed clock by not infringing on zero. Model selection found the Bayesian skygrid to be the best substitution model, as is typically the choice in between-host datasets to account changes in epidemiological spread over time. Analyses were run for 500 million MCMC generations, with parameters sampled every 10,000 steps. Convergence was defined as the ESS exceeding 200. I report the mean of the rateMean posterior distribution as the evolutionary rate.

5.4 Results

5.4.1 Genome-wide patterns of genetic divergence within and between host

To characterise the mode and tempo of HIV evolution within infected individuals, I first examined patterns of sequence divergence across the viral genome for the 89 individuals included in the study dataset. Using the overlapping sliding windows, I. quantified divergence rates across the genome for each infection. Averaging these

window-specific rates across individuals revealed substantial variation in the rate of divergence from the early virus population across different genomic regions (Figure 5.1). As expected from the known functional constraints on viral proteins, the regions with the lowest rates were located in the POL gene, which encodes essential viral enzymes. Specifically, the slowest-evolving regions were centred on HXB2 coordinates 4400-4900 (encompassing the integrase coding region) exhibiting a median rate of 6.3×10^{-4} substitutions per site per year (substitutions/site/year). There is also a pronounced dip in divergence rates centred at the window with HXB2 coordinates 1250-1750, within p24 protein in GAG, with a median rate of 1.3×10^{-3} substitutions/site/year. The highest rates were over 25 times greater (1.7×10^{-2} substitutions/site/year) and concentrated in windows featuring the variable loops within GP120, the part of the envelope (ENV) protein that is the primary target for neutralising antibodies and evolves under strong selection pressure. The patterns across the genome, as well as the magnitude of the divergence rate, are closely aligned to sequence divergence described in Zanini et al. (2017) for a cohort of eight individuals with longitudinal sampling covering many years into infection with less frequent sampling than in our study.

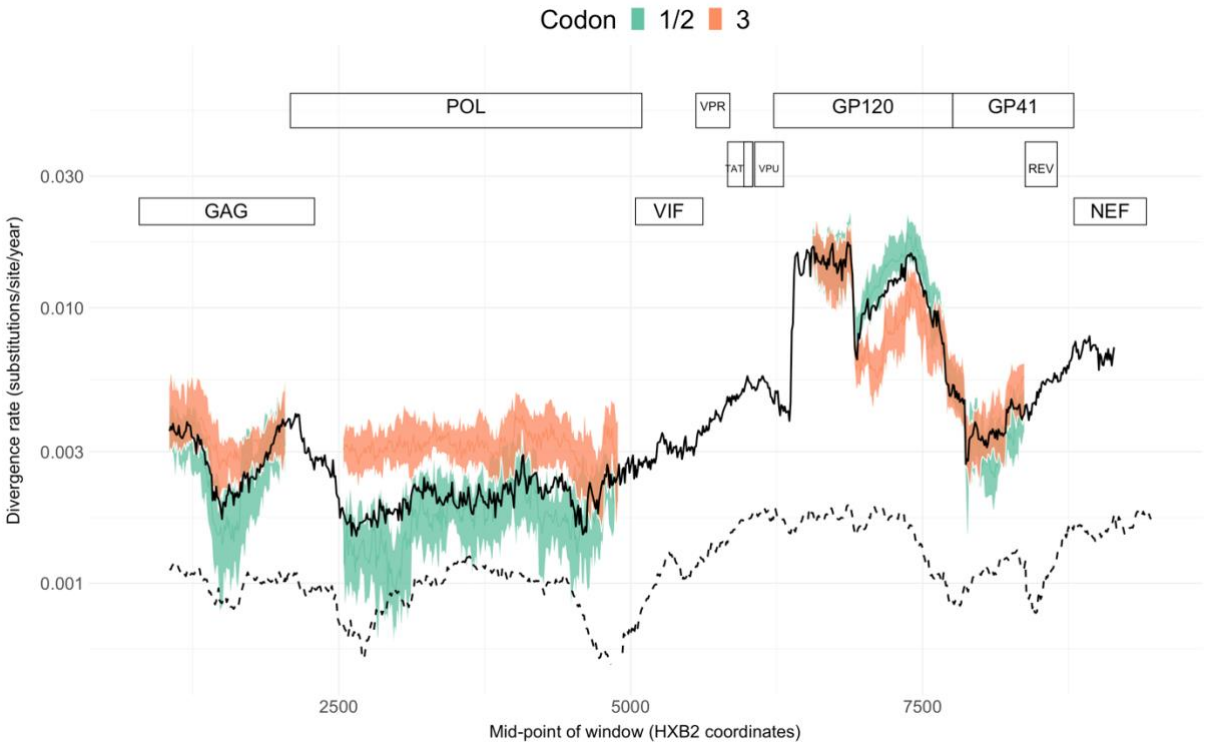

Figure 5.1 Divergence rates across the HIV genome reveal distinct patterns between within-host and between-host evolution. Within-host divergence rates (solid black line) were calculated using overlapping 500bp windows with 490bp overlap, showing the median rate across all individuals for each window. Codon-specific rates were determined for non-overlapping coding regions, with codon positions 1 and 2 (green) and 3 (orange) showing distinct patterns across genes. Shaded regions represent 95% confidence intervals derived from 1000 bootstrap replicates. For each replicate, I randomly sampled N individuals with replacement (N=89). For each resampled set, I calculated the median divergence rate within each time window by first computing the average rate per individual and then taking the median across all individuals in that bootstrap sample. The 95% confidence intervals were then calculated from the distribution of these bootstrap means, taking the 2.5th and 97.5th percentiles. The between-host divergence rate (dashed black line) represents the mean rate across subtypes A, C, and D, revealing consistently lower rates compared to within-host evolution. At the within-host level, codon 3 positions showed significantly higher rates in GAG and POL regions – hypothesised to be driven by purifying selection - while codons 1/2 exhibited elevated rates in parts of GP120. In comparing within- and between-host evolution, within-host evolution is consistently faster, with the largest discrepancy in GP120. The discrepancy in rates is lowest in POL, specifically in the p31 int protein.

The similarity in divergence between the two datasets suggests consistency in divergence patterns across individuals and subtypes that are relatively constant over time and are fundamental to within-host viral evolution.

While overall rate variation across genes reflects different selective constraints, examining rates at different codon positions can provide more detailed insight into the nature of selection acting on these regions. Specifically, I determined the divergence rate by codon position within the non-overlapping regions of three largest genes: GAG, POL and ENV. Given the variation observed across ENV, as well as their distinct

functions, I considered the two proteins of ENV – GP120 and GP41 – separately for the remainder of analyses. Across GAG and POL, I observed higher divergence rates at third codon positions consistently across the regions. Within both ENV proteins, rates across codon positions were similarly elevated, with greater divergence at codon positions 1 and 2 approximately covering the regions with V3-5, likely a consequence of positive selection in response to immune pressure.

To measure divergence rates at the between-host level, I analysed HIV-1 sequences from subtypes A1, C, and D sampled across sub-Saharan Africa over the past four decades. To allow for comparison with the within-host sliding window rates, I determined the rate at which the hamming distance from the subtype phylogenetic tree root increased over time in windows matching in co-ordinates to the within-host analysis, and then averaged rates across the three subtypes to generate a single rate estimate (figure 1, dashed line). Rates across the genome were broadly consistent across subtypes and were highly correlated (Pearson correlation A vs C: 0.39, C vs D 0.46, A vs D: 0.7, $p<0.001$ is each case). As has been previously reported, the rate of divergence at the between-host level is substantially lower than at the within-host scale across the genome, with the greatest discrepancy observed in GP120. However, window-specific rates averaged across all individuals at the within-host level were significantly correlated with the corresponding between-host rate (Pearson correlation coefficient 0.4, $p<0.001$), suggesting similar selective constraints and drivers shaping evolution over both within and between-host scales, as well as over short and long-term time scales, and provides further support of a ‘universal fitness’ landscape of mutations (Figure 5.2).

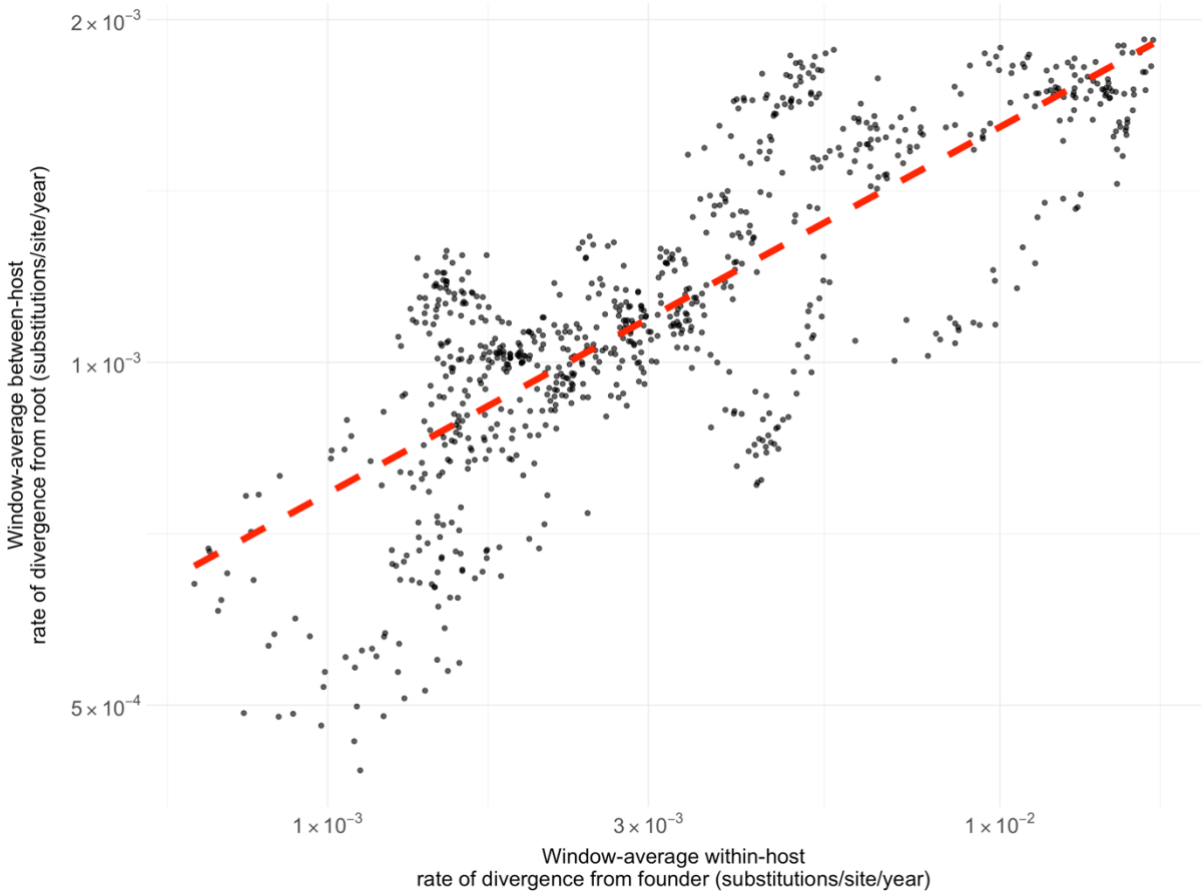

Figure 5.2 Correlation between windows-specific rates at the within and between host scales. Each point represents a single 500bp genomic window. Window-level divergence rates at the within-host level were strongly correlated to the window-level between-host divergence rate. To account for the overlaps between windows, I iteratively sampled all possible sets of independent non-overlapping windows to calculate correlations while maintaining statistical independence and calculated the Pearson correlation coefficient and associated P-value, and found all P-values to be lower than $\mathbf{10^{-3}}$. The red dashed line represents the fitted linear regression between window-specific within-host and between-host evolutionary rates on a log-log scale. A cluster of outlier points in the lower left quadrant reflect the strong functional constraints within-host in the integrase coding region in POL such that within-host and between-host rates are comparable.

5.4.2 Evolutionary rate analysis with BEAST

Molecular clock signal

While measuring divergence provides insight into viral evolution, this approach has important limitations - specifically, the inability to account for multiple substitutions at the same site, phylogenetic relationships between sequences, and variation in rates across lineages. To better capture these complexities of viral evolutionary dynamics, I performed Bayesian evolutionary analysis using BEAST. Since BEAST will always produce a clock rate estimate regardless of whether the data exhibits temporal signal, I first established which genomic regions showed evidence of clock-like evolution,

defined as a positive correlation between genetic distances and sampling times (adjusted R-squared > 0.8) and sequence divergence exceeding background variation and sequencing error noise.

To ensure independence between regions while maximising genome coverage, I selected windows with the strongest clock signal (highest adjusted R-squared), allowing no more than 10% overlap between adjacent windows. Using these criteria, 67% of the genome (30%-90% across individuals) showed evidence of clock-like evolution and was included in the BEAST analysis. Coverage varied by gene, with POL showing the lowest average coverage at 57% (5%-95% range: 30%-80%) and GP120 the highest at 70% (5%-95% range: 33%-95%). While this approach means that not all genomic regions were analysed for every individual, the large sample size of 89 individuals, combined with windows distributed across each gene, provides substantial data for comparing evolutionary patterns with between-host rates when we pool individuals.

Evolutionary rates vary substantially across infections and scales

Evolutionary rates were estimated for each window using a relaxed molecular clock model, with data partitioned by codon position (positions 1 and 2 versus position 3). To derive gene-specific estimates, windows across each gene were pooled, and an average rate was calculated by weighting the clock rate according to the variance of the posterior distribution.

Across all genes, BEAST evolutionary rates exhibited substantial variation, ranging from rates comparable to between-host evolution up to rates nearly 20 times higher (Figure 5.3). These rates were generally higher than those reported in previous studies, which may reflect our study design: more frequent sampling captured transient mutations, the shorter observation period emphasised early infection dynamics, and rate estimates were derived from smaller genomic windows. As the analysis was performed on windows of 500bps in regions with molecular clock signal, the entire gene is not necessarily analysed for each individual, which is necessary for the BEAST analysis but will bias our analysis towards finding faster clock rates.

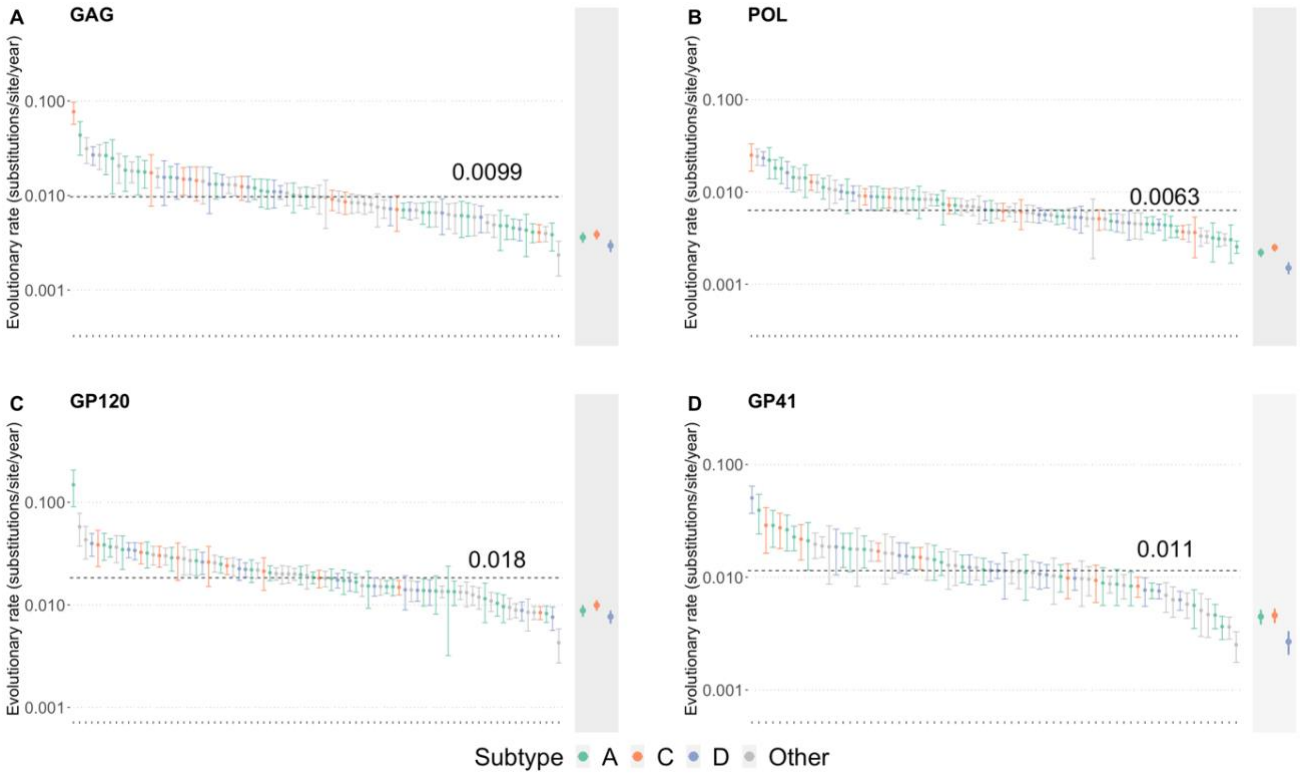

Figure 5.3 Within-host evolutionary rates inferred by BEAST by gene. BEAST evolutionary rates per individual (x-axis), with errors representing one standard error, pooled from the variance estimates for windows within a gene (when more than one window was included). Individual rates are coloured by subtype. The between-host evolutionary rate is denoted in the grey shaded bar, coloured by subtype. The median rate is denoted on the dashed line. An important caveat is that the rates for each individual are not determined from the same set of sites within the gene, as only regions with evidence of the molecular clock were included in the BEAST analysis dataset, which differed across individuals.

To investigate the role of selection in elevating within-host rates and in rate variation across individuals, I analysed codon-partitioned rates (Figure 5.4). Given the codon-specific differences observed across the genome, it is surprising that rates differed minimally between partitions. This pattern also contrasts with previous studies that found elevated rates at first and second codon positions within ENV, which again may be a consequence of study design. At the between-host level, I observed significant variation in evolutionary rates among viral subtypes, with subtype C evolving fastest across the genome and subtype D showing the slowest rates. However, these subtype-specific differences disappeared at the within-host level, with no significant difference between gene-specific within-host rates across subtypes. Similarly, I found no evidence for sex-based differences in evolutionary rates.

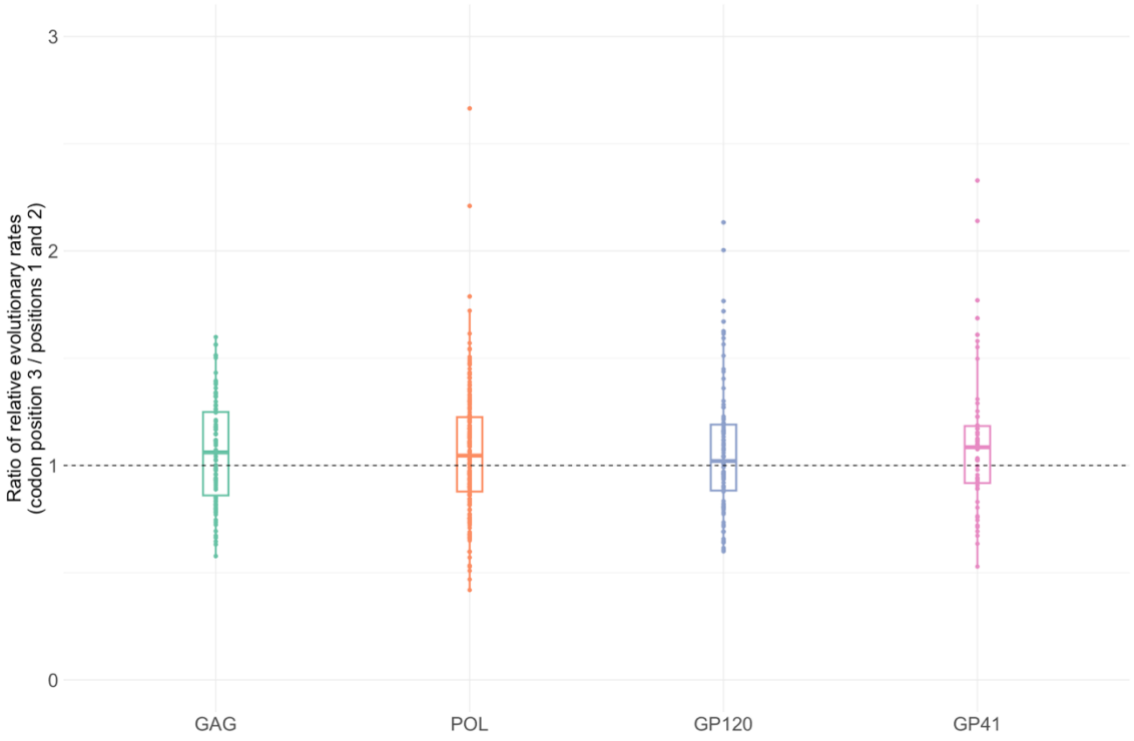

Figure 5.4 Evolutionary rate variation across HIV codon positions differs between genes.

The plot shows the ratio of evolutionary rates between codon position 3 versus positions 1 and 2 across different HIV genes, derived from BEAST analyses with codon-partitioned data. Each point represents a genomic window, with rates combined using inverse-variance weighting from posterior distributions. The dashed line at ratio = 1 indicates equal evolutionary rates between codon positions. While GAG, GP120, and GP41 showed no significant deviation from equal rates (Wilcoxon and t-tests, $p > 0.05$), POL contained three extreme values (ratios 4-12, not shown) that were excluded from visualisation. After removing these outliers, POL also showed no significant difference from equal evolutionary rates between codon positions.

Elevated within-host rates on external branches

In a relaxed molecular clock framework, evolutionary rates are allowed to vary across branches of the phylogenetic tree. Specifically, we can compare rates for external, internal, and backbone branches, where backbone branches are defined as those leading to the tips representing the viral population sampled at the final time point, external branches are the terminal branches of the tree, and internal branches represent all other branches. For each genomic window analysed in a region of interest, we calculated the ratios between the average rates of external and internal branches, external and backbone branches, and internal and backbone branches.

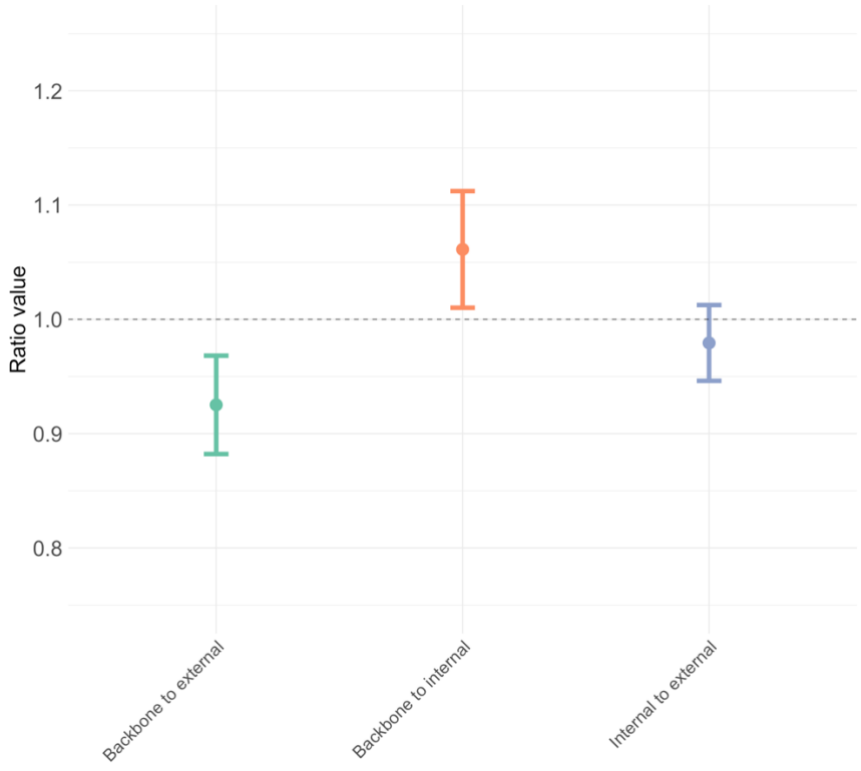

Figure 5.5 Ratios between branch-specific evolutionary rates. Points show ratios of evolutionary rates between different branch types inferred from the best tree from the BEAST posterior distribution, with error bars representing 95% confidence intervals. Rates were weighted by branch length to account for the amount of evolutionary time represented by each branch, ensuring longer branches contribute proportionally more to the rate estimates. Weighting by branch was applied to follow the methods in Lemey et al (2007) when accounting for the effect of transient mutations, which itself was based on Drummond et al (2006). I compared three branch categories: backbone branches (those leading to the final timepoint's monophyletic clade), internal branches (all other internal branches), and external (terminal) branches. The backbone-to-external ratio and internal-to-external highlight the elevated rates on the terminal branches that carry greater mutational burden from sequencing error or deleterious mutations. The backbone-to-internal ratio is significantly greater than 1, suggesting these branches capture adaptations that may later revert (transient mutations), while backbone branches might represent more stable evolutionary trajectories.

As expected, external branches generally exhibited the highest evolutionary rates (Figure 5.5). This is likely due to the additional mutational burden introduced by sequencing errors and deleterious mutations, which are rapidly purged. The backbone-to-internal branch ratio is significantly greater than 1, suggesting that internal branches may capture transient mutations that eventually revert, while backbone branches may reflect more stable evolutionary trajectories over time.

Evolutionary rate is strongly correlated across the genome

By examining correlations in evolutionary rates across different regions of the genome with distinct functions, we can disentangle host factors that influence evolution genome-wide from those with gene-specific effects. To assess these correlations, I focused on internal branches, including backbone branches, to minimise the impact of sequencing errors or recent deleterious mutations on inflating rates, and to more accurately capture the evolutionary dynamics that shape the viral population over longer time scales.

Despite substantial variation in absolute rates of evolution across genes, I observed strong correlations in internal branch rates across most genomic regions with Pearson correlation coefficients ranging from r=0.12 to r=0.56 ($p<0.001$) (Figure 5.6). The strongest correlations are observed between GAG-POL (r=0.56, $p<0.001$) and GAG-GP120 (r=0.52, $p<0.001$), suggesting shared factors influence early evolution across structurally and functionally distinct proteins. GP120-GP41 show strong (r=0.49, $p<0.001$), possibly due to their physical interaction in the envelope complex, while GAG-GP41 show no significant correlation (r=0.12, p=0.437).

The broad correlation patterns observed across genes may be driven by multiple mechanisms. Shared host factors, such as those influencing viral replication, population size and mutation rates, may drive genome-wide patterns. Similarly, comparable selection pressures across genes within individual hosts could contribute to this coordination, for example CTLs can target epitopes within GAG, POL and ENV, whereas antibodies target mostly ENV. The action of different arms of the immune system on different genes may also explain why there is no correlation between GP41 (ENV) and GAG. Epistatic interactions, which maintain co-evolving genomic regions despite high recombination rates, may play a role as well. Additionally, host-specific factors, such as strength of innate and acquired immune response, may globally influence evolutionary rates across the genome.

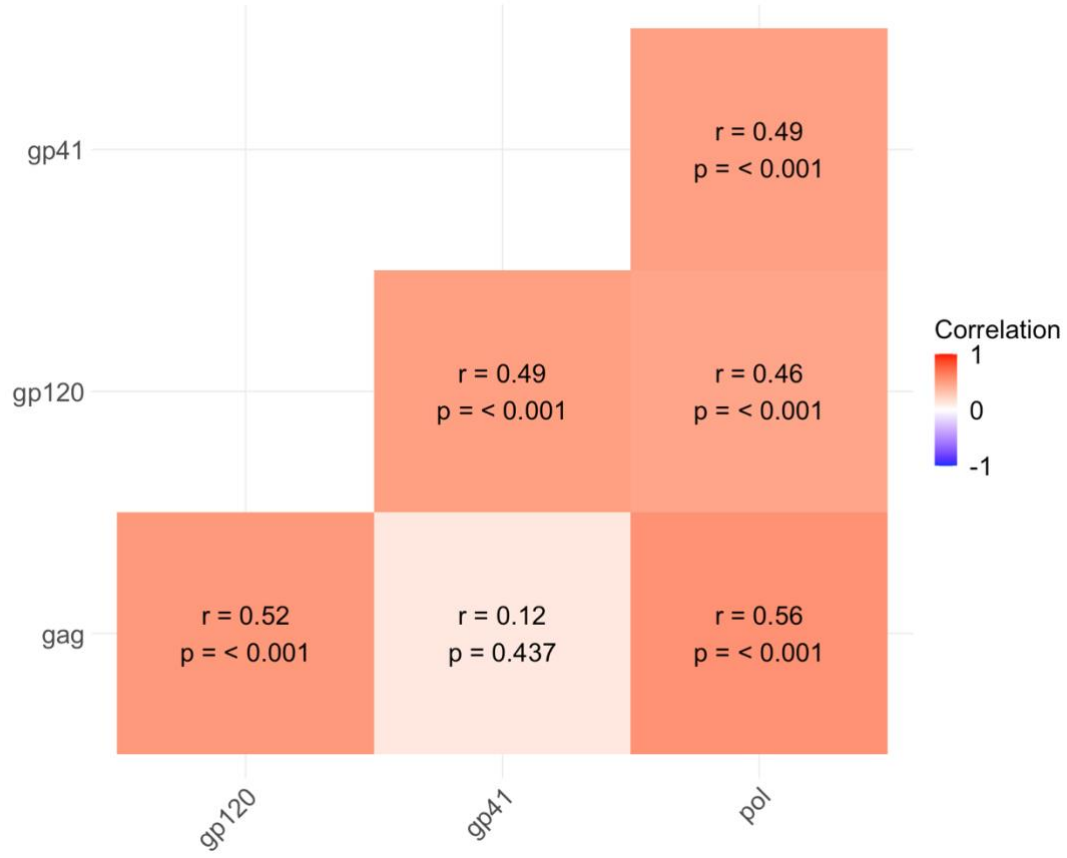

Figure 5.6 Evolutionary rates on internal branches are correlated between viral genes within individuals. Heat map shows pairwise correlations of internal branch rates between genes, with correlation coefficients (r) and p-values shown for each comparison. Stronger correlations (darker red) were observed between POL and both GAG and GP120 as well as between GAG and GP120. GP41 showed strong correlation with POL and GP120, but no significant correlation with GAG. These (almost) consistent within-host correlations across most gene pairs suggest that some individuals maintain higher or lower evolutionary rates across multiple genes, potentially reflecting host-specific factors affecting viral evolution.

The evolutionary rate is negatively correlated with viral load

A previous study examining viral evolution over decades of infection found that set-point viral load (SPVL) - a proxy for disease progression - positively correlates with synonymous evolutionary rates (Lemey *et al.*, 2007). The analysis, again focusing on internal branch rate, reveals a contrasting pattern: a negative association between evolutionary rate and SPVL that is consistent in three of the four major regions examined (GAG, POL, GP120; figure 5.7), and also observed in GP41 albeit with lower statistical support and weaker strength. This consistent signal across functionally distinct proteins suggests a host-level mechanism affecting viral evolution genome-wide. Why we find an opposing result to previous studies is not clear but may be linked

to distinct processes affecting early versus late infection. Early infection may experience stronger immune pressure and higher effective APOBEC activity when viral loads are lower, while studies of chronic infection capture a different dynamic where higher viral loads enable faster evolution through increased replication.

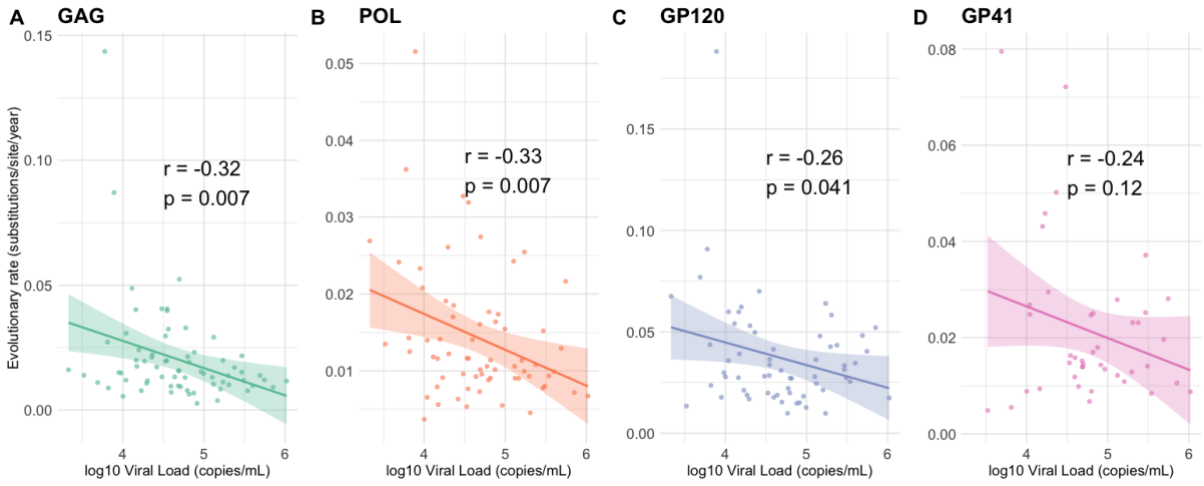

Figure 5.7 Within-host HIV evolutionary rates decrease with higher viral loads across major genes. Scatter plots show the relationship between log10 viral load (copies/mL) and internal branch evolutionary rates (substitutions/site/year) for (A) GAG, (B) POL, (C) GP120, and (D) GP41. The y-axis scale differs across genes due to variation in the magnitude and range of rates. Pearson's correlation test showed negative correlations were observed for all genes, with significant associations in GAG, POL and GP120 and trending negative correlations in GP41. Lines show linear regression fits with 95% confidence intervals (shaded regions). These consistent negative associations suggest that viral populations under lower immune pressure (indicated by higher viral loads) may experience reduced selective pressure for adaptation.

5.4.3 Tempo of evolution differs across methodologies and scales

Understanding how different methodological approaches affect evolutionary rate estimates and the magnitude of the evolutionary rate mismatch is both insightful for interpreting viral dynamics and for comparison across studies. We compared rates estimated using BEAST, which explicitly models the evolutionary process, with those calculated from sequence divergence for both within and between-host scales.

First, I compared evolutionary rate estimates from BEAST to divergence rates to validate our phylogenetic approach by performing a window-to-window comparison. Reassuringly, BEAST evolutionary rates strongly correlate with divergence rates. However, evolutionary rates were elevated, with an average (median) inflation of a factor of 2.4 (5%-95% range: 0.95 – 7.5) (Figure 5.8). This systematic difference likely

reflects BEAST's ability to capture transient mutations that are unobserved when divergence alone is measured. When measuring divergence through linear regression across multiple timepoints, transient mutations that later revert appear as deviations from the trend line rather than contributing substantially to the rate estimate. Ultimately, the regression primarily captures the sustained accumulation of mutations over time, while BEAST's phylogenetic approach explicitly counts both the appearance and reversion of mutations along branches of the tree, as well as variation in evolutionary rates over time, leading to higher rate estimates. This fundamental difference in methodology also may also explain why, despite significant differences in divergence, the evolutionary rate by codon position did not substantially differ. Overall, the strong correlation between methods suggests that while absolute rate estimates differ, both approaches capture similar patterns of rate variation across individuals and the genome.

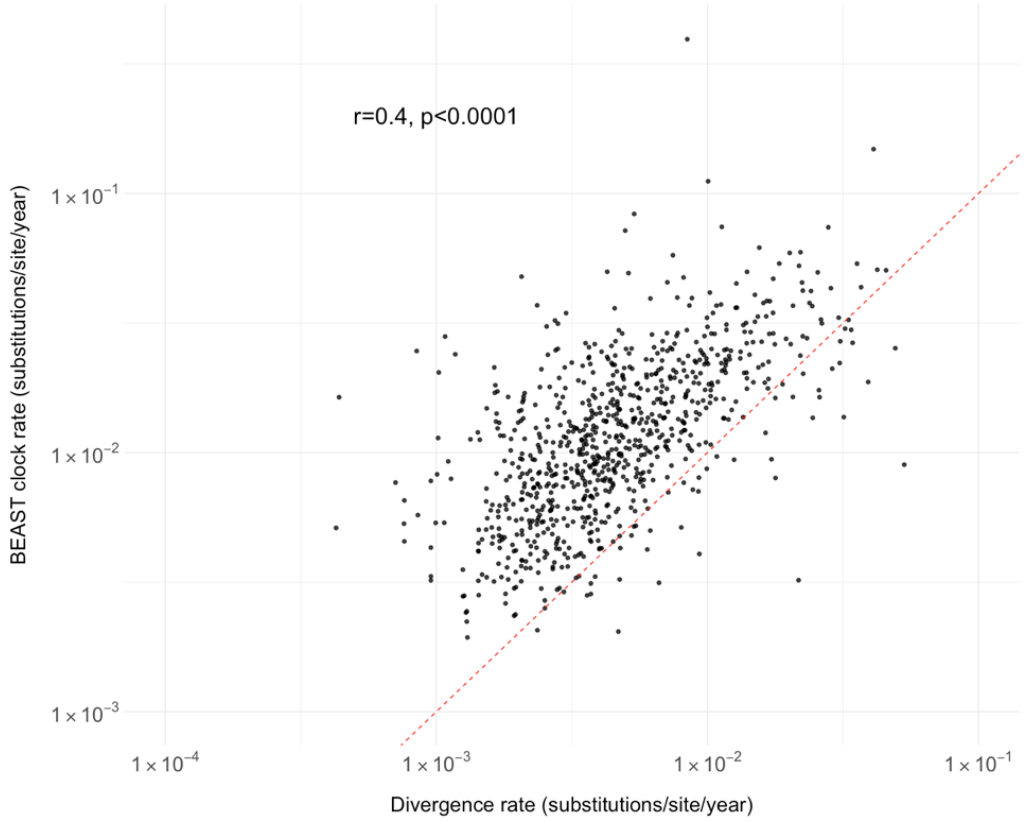

Figure 5.8 BEAST within-host rates systematically greater than divergence rates. Each point represents a 500bp genomic window from one individual, and the red dashed line represents x=y. The correlation in rates is high (r=0.4, p<0.0001), however BEAST rates are systematically higher, with an average inflation of 2.4.

Next, I compared within-host evolutionary (BEAST) and divergence (Hamming distance) rates across the four key regions by methodology and scale (Figure 5.9, Table 5.1). Gene-level within-host divergence rates for each individual were calculated by averaging across genes the same windows as chosen for the BEAST analysis. The pattern I observed within-host is even more pronounced at the between-host level, where the rate of genetic divergence from the subtype-specific root sequence is substantially lower than the BEAST evolutionary rate. Between-host BEAST estimates were 4-5 times higher than between-host divergence rates in GAG, GP120 and GP41, and 4 times higher in POL. This larger difference at the between-host scale likely reflects the accumulation of multiple substitutions at the same genomic positions over longer timescales, particularly given frequent reversions of host-specific adaptations upon transmission.

The disparity between within-host and between-host evolutionary rates was substantial. BEAST clock rates were 2-3 times higher at the within-host level (Table 5.1), consistent with previous estimates. The difference was even more pronounced in divergence-based estimates, ranging from over 4-fold higher in POL to over 7-fold higher in GP120.

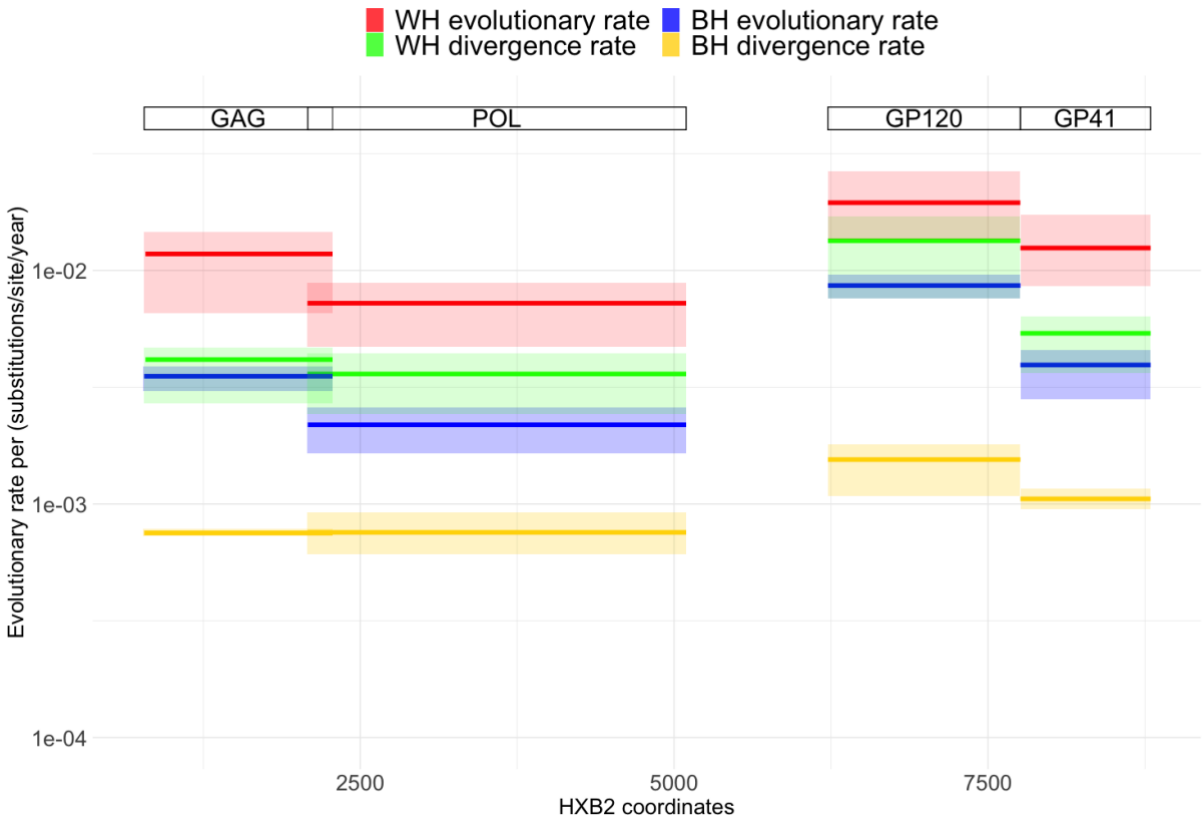

Figure 5.9 Comparing evolutionary rates across methods and scales Lines show evolutionary rates (substitutions/site/year) estimated using both divergence measurements (net sequence change) and BEAST analysis (phylogenetic inference) at the within-host scale (red: BEAST, green: divergence) and between-host scale (blue: BEAST, yellow: divergence). At the within-host scale, shaded regions denote the 25%-75% IQR across individuals. As individuals are pooled, this does not account for the uncertainty in the rate estimates and the variation in the BEAST posterior distributions. At the between-host scale, shaded regions denote the minimum and maximum rates across the three subtypes. For divergence, rates were very similar across subtypes and so the variation is not visible on the plot. Between-host BEAST rates exceed divergence estimates, capturing the impact of sites that repeatedly escape and revert during transmission - a pattern previously documented in between-host HIV evolution. Within-host divergence rates align more closely with between-host BEAST estimates, suggesting that net sequence changes better reflect long-term evolutionary dynamics by excluding transient mutations that inflate within-host BEAST estimates but do not persist at the between-host scale.

Table 5.1 Evolutionary rate mismatch by scale and methodology. The mismatch is defined as the ratio between the evolutionary rate (substitutions/site/year) by a given method/measurement scales and the evolutionary rate by either a different method or measurement scale. For a fair comparison across methodologies, for each individual only the windows included in the BEAST analysis contribute to the within-host divergence rate. As a result, the difference in divergence between and within hosts is larger than shown in figure 1, with the main difference in POL.

Gene	WH: BEAST to Divergence ratio	BH: BEAST to divergence ratio	WH BEAST rate to between BEAST rate ratio	WH divergence rate to BH divergence rate ratio
GAG	2.8	4.9	3.2	5.6
POL	2.01	3.1	3.1	4.6
ENV: GP120	1.45	4.8	2.3	7.5
ENV: GP41	2.32	4.2	2.8	5.9

5.4.4 Elevated within-host rates across shorter time scales

In our dataset, the number of days between the first and last sampling point varied between 150 and 1200 days, with a median observation period of 357 days. I observed that evolutionary rates decrease with longer observation periods in both BEAST and divergence-based analyses, though BEAST consistently estimates higher rates (Figure 11). Correlation tests were performed on log transformed data (time in days and rates), however the significance of the result did not change when the test was performed on untransformed values. The strength of the negative correlation varies across genes between R=-0.26 in GP120, R= -0.28 in GP41, R=-0.32 in POL, and R=-0.38 in GAG for divergence data. For evolutionary rates from BEAST, R=-0.26 in GP120, R= -0.29 in GP41, R=-0.22 in POL, and R=-0.29 in GAG. In all cases, the p-value was less than 0.05. The strength of the relationship between evolutionary rate and time interval is stronger when using divergence-based methods (with the exception of GP41). This is because divergence calculations do not account for repeated substitutions at the same site, which are increasingly likely over longer time intervals. In contrast, BEAST incorporates these repeated substitutions into its estimates, dampening the apparent effect of time interval on evolutionary rates. In our data, individuals with longer total observation periods (time between first and last sample) typically had larger intervals between sequential samples. This means that rapid evolutionary dynamics occurring over short timescales were less likely to be

captured in these individuals compared to those with more frequent sampling, which may contribute to the time-dependence we observe in BEAST.

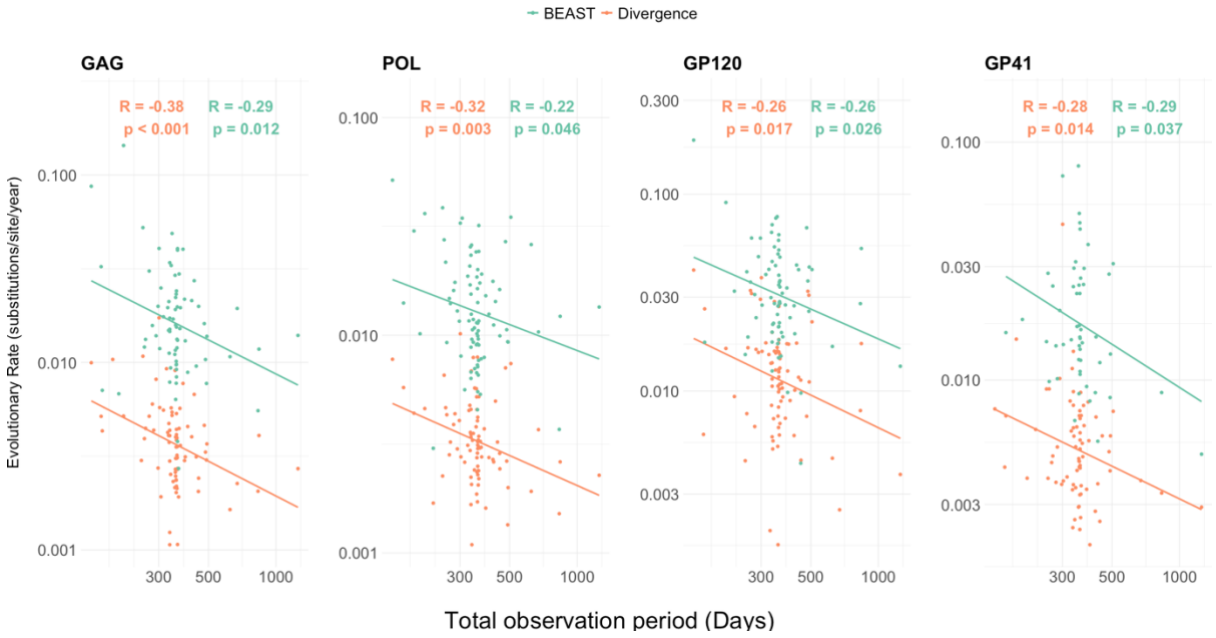

Figure 5.10 Within-host divergence and evolutionary rates decease over longer time scales. The number of days between the first and last sample varied from 150 to 1200 across individuals, with an average of 357 days. Gene-averaged rates slowed down in individuals for whom a greater number of days separated the first and last sample. Each point represents the evolutionary/divergence rate of individual, with BEAST in green and divergence in green. The output of a correlation test between days between first and last sampling point and evolutionary/divergence rate are annotated in text. Time (in days) and rates have been log-transformed, however the correlation size and significance value did not significantly change when tests were performed on untransformed values.

The higher rates in early infection may also reflect genuine biological processes—strong initial immune pressures, adaptation to the new host environment, and less constrained viral populations, which supports our earlier viral load finding that implied distinct processes during early and late infection.

In summary, divergence better captures net genetic change, while BEAST's phylogenetic reconstruction provides the most complete picture of evolutionary dynamics. Crucially, BEAST can detect mutations that temporarily rise to high frequency before declining - a phenomenon known as "within-host toggling." To explore the extent of within-host toggling, I next examined the temporal trajectories of specific mutations within individuals to directly observe these frequency fluctuations and provide empirical support for this mechanism.

5.4.5 Extensive within-host toggling for synonymous and non-synonymous mutations

To investigate the potential role of toggling, I tracked the trajectories of all observed alleles at polymorphic sites across the genome for each individual. Specifically, we looked at de novo mutations, defined as a mutation with an initial frequency lower than 5%. Using codon-level data, mutations were classified as either synonymous or non-synonymous (see methods). The analysis revealed two distinct patterns: selective sweeps occurring at specific loci across the genome ("not reverting") and a high prevalence of minor variants that temporarily rise to substantial frequencies (>50%) before "reverting" to their original state. I term this dynamic "toggling," reflecting the oscillation of alleles between states over time. Trajectory data is visualised for two individuals in figure 5.11A-B, with the remaining individuals shown in supplementary data (Figure 9.6), where specifically we look at the trajectories of alleles initially at lower than 5% frequency, before reaching consensus-level frequency (>50%). For person A, 34% of the synonymous mutations are no longer the major allele by the final time point, with 14% of the non-synonymous mutations. For person B, 27% synonymous mutations toggle, with 55% for non-synonymous.

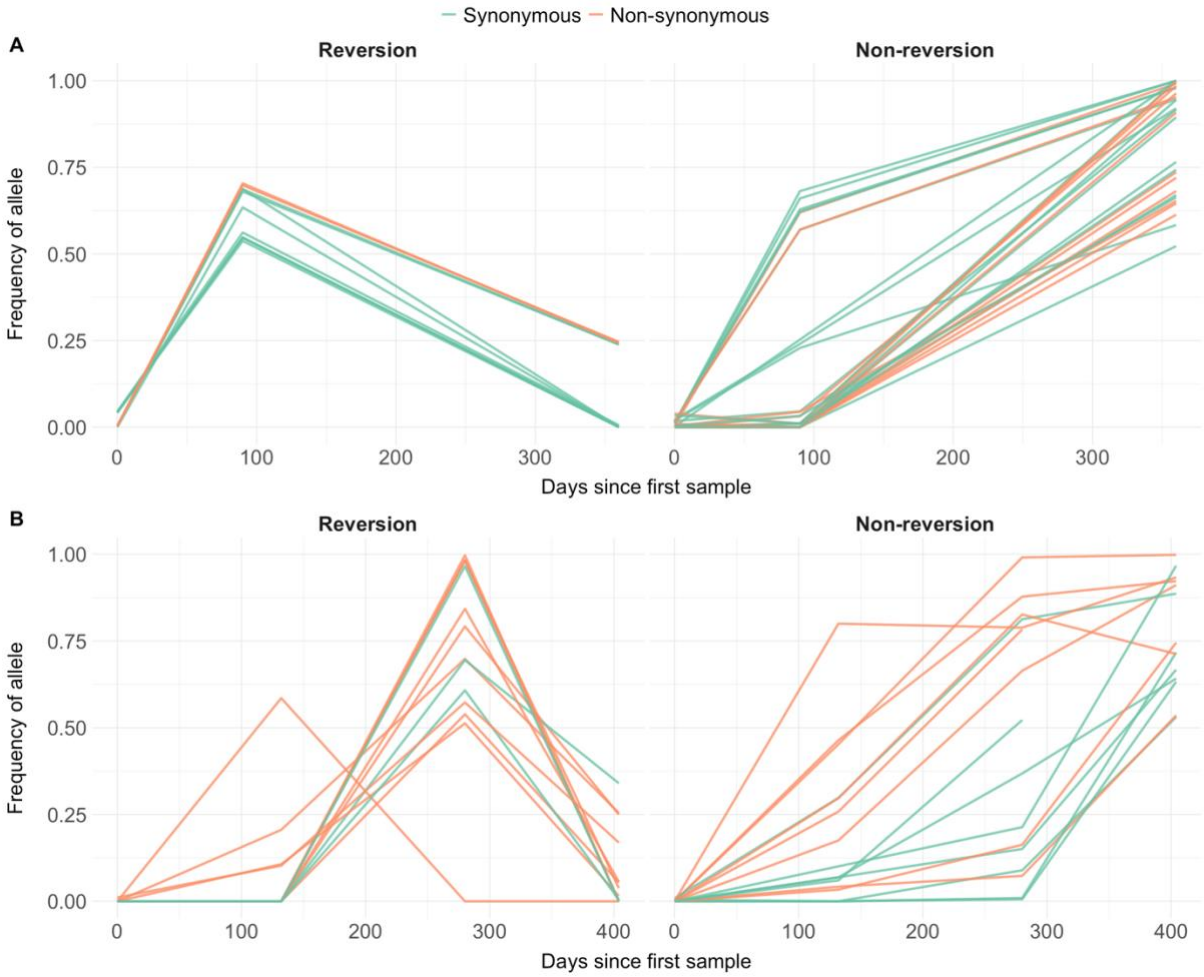

Figure 5.11 Trajectory of minor variants at genomic sites that "toggle" or sweep over time A) Trajectories of (initially) minor alleles that reach consensus frequency (>50%) at a later time during infection, separated by "reversion" – the mutation toggles- and – "non-reversion" - the mutation persists at high frequency, for a single individual (person 'A') in the study. Trajectories are coloured by whether they cause a change at the amino acid level (non-synonymous: orange) or do not cause a change (green: synonymous). For both types of mutations, we observe a substantial proportion of reversions, with 34% of synonymous mutations reverting and 14% of non-synonymous. B) Analogous to figure A with a data representative of a viral population of a different individual (person 'B') in the study. Here, 27% of non-synonymous mutations that reach consensus frequency toggle/revert, and 55% non-synonymous, highlighting that the dynamic is not exclusive to neutral mutations.

To quantify the prevalence of toggling in the entire dataset, I analysed *de novo* mutation trajectories by tracking alleles that reached specific frequency thresholds before the final sampling point. I examined synonymous and non-synonymous mutations separately, pooling trajectories across all individuals. I found that as frequency thresholds increased (i.e. maximum frequency prior to final time point), the proportion of alleles that maintained their high-frequency status increases (Figure 5.12). This pattern makes biological sense - mutations that achieve higher frequencies

are more likely to be maintained. Synonymous mutations (orange line) showed higher maintenance rates compared to non-synonymous mutations (green line) across all thresholds. For instance, at a frequency threshold of 0.75, ~70% of synonymous mutations maintained their frequency, while only about 45% of non-synonymous mutations did so. The former result aligns with expectations - synonymous mutations are generally neutral, and once they reach high frequencies through hitchhiking with beneficial mutations, there is less opportunity for recombination to break the linkage. Given the expected small fitness cost associated with the synonymous mutations, the time to decline frequency will be slower than for non-synonymous mutations. The lower maintenance of non-synonymous mutations suggests many are either deleterious and eventually purged, or experience dynamic selection pressures within the host, potentially due to changing immune responses, leading to frequency fluctuations over time. Toggling of non-synonymous mutations also likely explains why we observe differences in divergence rates by codon position in a large proportion of the genome, but we do not observe differences in evolutionary rates.

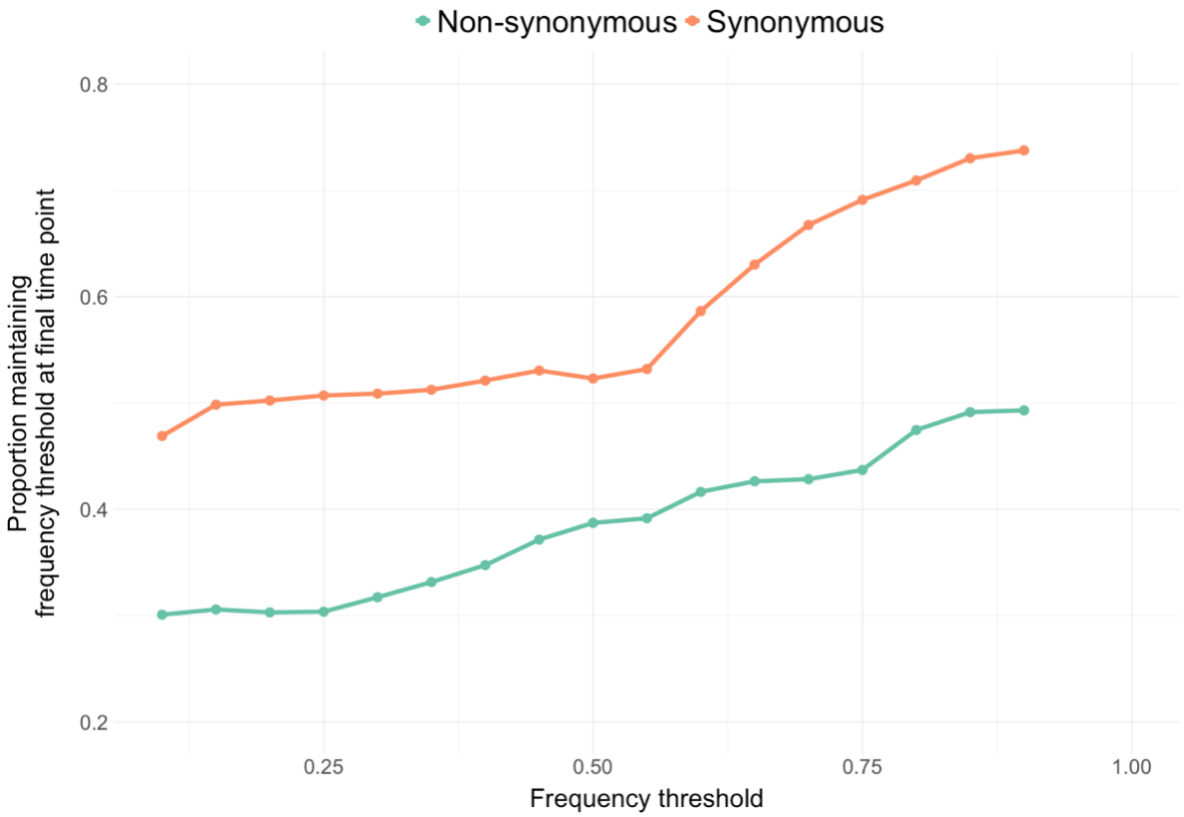

Figure 5.12 Persistence of mutations above frequency thresholds. For each frequency threshold (0.05-0.9), I identified all *de novo* mutations that exceeded this threshold prior to the final sampling point, with mutations pooled across individuals and a *de novo* mutation defined as a minor allele with initial frequency less than 5%. At the final sampling point, we determined the proportion of mutations that maintained or exceeded their threshold frequency. Synonymous (green) and non-synonymous (orange) mutations were analysed separately. In both cases, as the threshold increases, the proportion that maintain their frequency also increases. The higher maintenance rate of synonymous mutations is expected given their neutral nature - once they reach high frequencies through hitchhiking with beneficial mutations, they are unlikely to be selected against. In contrast, non-synonymous mutations show lower maintenance rates, suggesting more complex selection dynamics. While some non-synonymous mutations are likely purged due to being deleterious, the substantial decline in frequency of mutations that reached high thresholds (>75%) may also reflect changing selection pressures within the host, particularly from dynamic immune responses that create time-varying selective environments.

5.5 Discussion

Through longitudinal sampling of a large cohort of untreated HIV-positive individuals during early infection, I quantified viral evolution using two distinct methodological approaches (BEAST and divergence) to characterise evolutional dynamics across genomic regions, individual hosts and ecological scales. The findings suggest host-specific or viral factors may be influencing the rate of evolution across the entire genome, and that evolutionary processes during early and late infection may differ. By incorporating population-level consensus sequences sampled in sub-Saharan Africa

over the last four decades, I directly compare the evolutionary and divergence rate across within and between scales. Although evolutionary rates differed substantially between scales, I observed strong correlation in patterns of sequence divergence across the genome at both within-host and between-host levels. This concordance suggests that consistent evolutionary constraints operate across biological scales and over short and long-time scales, from early infection to decades-long epidemic spread.

Using a BEAST analysis, I find that within-host evolutionary rates are noticeably higher than in past studies (Lemey *et al.*, 2007; Alizon and Fraser, 2013; Vrancken *et al.*, 2014; Raghwani *et al.*, 2018), and I also do not observe significant differences by codon-partition as may be expected from regions evolving under purifying (POL) or positive (ENV) selection (Novitsky *et al.*, 2013) . Most likely this is a consequence of study design, particularly the use of denser sampling, with sampling every three months. In the study of rapidly evolving viruses, more frequent sampling has been linked to evolutionary rate increases due to the effect of transient mutations that do not fix (Duffy, Shackelton and Holmes, 2008). Comparing BEAST estimates with divergence rates at the within-host scale reveals that BEAST rates are approximately 2.4-fold higher, reflecting detection of toggling mutations that divergence measures do not capture.

Both evolutionary and divergence rates decrease over longer observation periods in our study, correlating with reduced sampling frequency. While the decay of evolutionary rates of viral pathogens over increasing timescales (from decades to millennia) is well-documented and typically attributed to the purging of slightly deleterious mutations and sequence saturation (Aiewsakun and Katzourakis, 2016; Ghafari *et al.*, 2021), I observe this phenomenon occurring within months. This rapid decay may also suggest that evolutionary rates vary throughout infection, possibly reflecting temporal shifts in immune-mediated selection pressures.

The disparity between within-host evolutionary and divergence rates, primarily driven by toggling mutations, appears to be a key factor explaining the evolutionary rate mismatch across scales. This is evidenced by within-host divergence rates that closely approximate, though slightly exceed, between-host evolutionary rates throughout the genome. However, the difference in evolutionary and divergence rates is likely to also stem from other methodological differences, including the fundamental approaches used by BEAST versus direct measures of divergence. While BEAST incorporates a

full probabilistic model of sequence evolution and accounts for phylogenetic uncertainty, divergence calculations provide a more direct measure of accumulated genetic change.

At the between-host scale, we observe a significantly reduced divergence rate compared to the evolutionary rate, consistent with an "escape and revert" model (Herbeck *et al.*, 2006; Zanini *et al.*, 2015; Illingworth *et al.*, 2020; Druelle and Neher, 2023). In this process, specific sites undergo repeated cycles of selection in hosts with similar immunological backgrounds, followed by reversion—a pattern that limits the accumulation of long-term mutations.

The choice of methodology for studying pathogen evolution is consequential and should be carefully aligned with the specific research objectives. For investigating fine-scale within-host evolutionary dynamics, including temporal changes in viral population composition and shifting selection pressures, sophisticated phylogenetic approaches like BEAST provide greater insight. However, when examining long-term evolutionary trajectories or transmission patterns, simpler divergence-based methods may be more appropriate, as toggling mutations can obscure transmission signals and complicate estimates of time to most recent common ancestor. The impact of the toggling dynamics on phylogenetic inference and transmission reconstruction needs further investigation, particularly in the context of densely sampled longitudinal studies.

A key strength of this study is its comprehensive genomic coverage across multiple HIV-infected individuals. However, our need to minimise recombination effects necessitated the use of 500bp windows, constraining our ability to analyse complete genes in each individual and potentially biasing our analysis toward faster clock rates—likely contributing to our elevated evolutionary rate estimates. While pooling data across individuals enabled robust gene-level analysis, this approach prevented us from incorporating the uncertainty inherent in BEAST-derived posterior distributions of clock rates. Determining optimal sequence lengths that balance recombination detection with phylogenetic accuracy remains a critical research challenge, particularly as new sequencing technologies enable analysis of longer genomic regions.

In tracking minor variant allele trajectories over time, I find extensive toggling of both synonymous and non-synonymous mutations. While previous studies in the C2-V5 region during chronic infection demonstrated that synonymous mutations can reach

high frequency through hitchhiking before being removed (Zanini and Neher, 2013), we show this dynamic occurs genome-wide on much shorter timescales. Our observation of widespread non-synonymous mutation toggling aligns with previous findings in C2-V5, where non-synonymous toggling was linked to delayed antibody responses (Hedskog *et al.*, 2010; Bonsignori *et al.*, 2017), but we demonstrate this occurs across the genome and over months rather than years. Since non-synonymous mutations are rarely neutral, this rapid genome-wide toggling may reflect either HIV exploring its fitness landscape as fitter variants emerge, or competition between epitope variants during CTL response adaptation (Chapter 6). Understanding these dynamics is crucial for choosing appropriate analytical approaches, as methods that do not account for such rapid reversals may miss important evolutionary patterns.

We find substantial variation in genome-wide evolutionary rates across individuals. This heterogeneity likely stems from both host-specific factors, such as immune response strength, and viral factors including replication dynamics and population size. The observed correlation in evolutionary rates across different genomic regions supports the existence of a genome-wide mechanism modulating evolutionary rates. Intriguingly, we found that lower set-point viral loads (spVLs) were associated with increased evolutionary rates, a pattern that contrasts with previous studies of chronic infection (Mikhail *et al.*, 2005; Lemey *et al.*, 2007). This unexpected relationship may reflect distinct evolutionary dynamics during early infection, potentially driven by robust early immune responses in individuals with lower viral loads. This hypothesis aligns with observations that early immune escape is associated with specific HLA profiles characteristic of long-term progressors, where strong immune pressure drives rapid viral adaptation through CTL escape mutations (Leslie *et al.*, 2004; Boutwell *et al.*, 2010; Goulder and Walker, 2012).

In summary, this study has quantified evolutionary rates across the HIV genome in a large cohort of untreated infections, revealing substantial heterogeneity both in the pace of evolution and methodological approaches. These findings suggest that frequent mutational toggling at both nucleotide and amino acid levels is a primary driver of elevated within-host evolutionary rates. These results have important practical implications for phylogenetic inference, as sampling strategies and measurement approaches significantly influence our understanding of viral evolution. Further investigation is needed to determine how transient mutations affect estimates

of transmission timing and the identification of transmission pairs. The impact of toggling mutations on evolutionary rate mismatches may extend beyond HIV to other chronic viral infections with rapid evolution, such as HIV-2, HCV and HBV (Lemey *et al.*, 2003; Gray *et al.*, 2011; Rocha *et al.*, 2013; Raghwani *et al.*, 2016; Vrancken, Suchard and Lemey, 2017).

6 Direct evidence of CTL escape and reversion across transmission Pairs

6.1 Abstract

Cytotoxic T lymphocytes (CTLs) shape HIV evolution by driving viral adaptation to the host's immune system. Tailored to the host's HLA profile, CTL responses often lead to viral escape mutations that enable immune evasion but impose fitness costs. Upon transmission to a new host, where immune pressures differ, these mutations may revert, potentially slowing viral evolution at the population level. Using viral populations sampled from 62 transmission pairs and their HLA genotypes, I identified hundreds of CTL escape mutations and tracked their trajectories across transmission events. The data revealed that certain viral regions and epitopes are consistently targeted across hosts with similar immunological profiles. At the same time, significant variation in escape mutations highlights HIV's adaptability. Evidence of negative selection on escape mutations following transmission, likely driven by HLA differences between source and recipient, provides direct evidence of the “escape and revert” dynamic**.** Notably, CTL escape contributed more to total genetic divergence than other sources during the first year of infection.

6.2 Introduction

Cytotoxic T lymphocytes (CTLs), also known as CD8+ T cells, are a key component of the adaptive immune response to viral infections (Plata *et al.*, 1987; Walker *et al.*, 1987). HIV infects CD4+ T cells and the genetic material of the virus is integrated into the host genome, after which viral proteins are processed and displayed on the surface of the cell as short peptides known as epitopes (McMichael and Rowland-Jones, 2001). Epitopes are presented by Human Leukocyte Antigens (HLAs), which are encoded by Major Histocompatibility Complex (MHC) class I genes (Townsend *et al.*, 1989; Yewdell and Bennink, 1992). CTLs recognise epitopes bound to HLA molecules via their T-cell receptors (TCRs), enabling them to identify and bind to the infected cells (Wange and Samelson, 1996). Once CTLs recognise the infected cells, they are activated to kill them, preventing further viral replication. Each HLA allele presents a unique set of viral epitopes to CTLs, of which some are more effective at presenting critical, conserved regions of the HIV genome, leading to stronger and more effective CTL responses (Kaslow *et al.*, 1996; Migueles *et al.*, 2000; O’Brien, Gao and

Carrington, 2001; Tang *et al.*, 2002; Altfeld *et al.*, 2006; Brumme *et al.*, 2009; Goulder and Walker, 2012).

The CTL response places strong evolutionary pressure on the virus, making the selection of escape mutations a major driver of evolution. Viral adaptations specific to CTL evasion have been identified across the HIV genome (Los Alamos National Laboratory, 2024), with over 50% of Amino Acid mutations outside of the Envelope gene (ENV) linked to the CTL response (Allen *et al.*, 2005). The virus evades immune recognition by altering the amino acid sequence in viral proteins, preventing CTLs from recognising or binding to their target epitopes (Yokomaku *et al.*, 2004; Carlson *et al.*, 2012; Kløverpris, Leslie and Goulder, 2016).

Adaptation to the CTL response is a continuous and dynamic process, with escape mutations emerging during both acute and chronic stages of HIV infection (Friedrich *et al.*, 2004; Leslie *et al.*, 2005; Troyer *et al.*, 2009). The selection pressure exerted by CTLs is particularly intense during the first 12 months of infection, where a study of acute infection in one individual showed that the majority of early viral mutations are driven by immune evasion (Henn *et al.*, 2012). However, defining CTL escape mutations is complicated by factors such as founder effects and the challenge of identifying associations between HLA alleles and escape mutations, particularly in population-level studies where shared ancestry can obscure these relationships (Goulder and Walker, 2012; Roberts *et al.*, 2015). Additionally, the timing of escape varies widely, and considerable variation is observed across different epitopes (Goulder and Watkins, 2004; Klenerman and Hill, 2005; Brumme *et al.*, 2008). In one large cohort study, only one-third of infections developed HLA-specific adapted epitopes within the first two years of infection (Roberts *et al.*, 2015). It is challenging to distinguish between transmitted and *de novo* escape, and the transmission of virus that is pre-adapted to the HLA profile of the recipient can also influence the occurrence of escape (Avila-Rios *et al.*, 2019). The extent of pre-adaptation is expected to shape the early trajectory of immune escape, as well as the long-term course of viral evolution within the host (Carlson *et al.*, 2016).

Escape pathways of the virus are considered to be HLA-specific, and genetic variation studies have identified profound differences in viraemic control and disease outcome across HLA alleles (Altfeld *et al.*, 2006; Carlson *et al.*, 2012; Goulder and Walker, 2012). In particular, individuals possessing HLA-B*57, HLA-B*58 or HLA-B*27 alleles

have repeatedly been identified as long-term nonprogressors (Kaslow *et al.*, 1996; Migueles *et al.*, 2000; O'Brien, Gao and Carrington, 2001; Navis *et al.*, 2007; Ramírez De Arellano *et al.*, 2019). Protective HLA alleles are linked to epitopes in highly conserved regions in which mutations carry considerable cost to viral fitness, specifically by reducing viral replicative capacity (Liu *et al.*, 2006; Crawford *et al.*, 2009; Miura *et al.*, 2009). Following transmission to a host with a distinct set of HLA alleles, which recognise and target different viral epitopes, an escape mutation may no longer provide an advantage and is often selected against, favouring a reversion to the ancestral state at the same genomic position. This phenomenon, known as reversion, has been repeatedly observed in HIV, as well as in SIV and HCV infections (Friedrich *et al.*, 2004; Leslie *et al.*, 2004; Timm *et al.*, 2004; Li *et al.*, 2007; Davenport *et al.*, 2008). A study of ENV sequence divergence in a small cohort of transmission pairs identified a general trend for the recovery of ancestral features during the first 2-3 months of infection, suggesting the sequence constraints that exist across individuals even in the most diverse genome regions (Herbeck *et al.*, 2006). Indeed, a study of seven individuals during the first year of infection reported reversion events to be more frequent than *de novo* escapes, signifying the substantial role that the reversion contributes to viral evolution during early infection (Li *et al.*, 2007). HLA-restricted mutations have also been shown to be remarkably predictable, with the same amino acid switches observed to emerge in infections across sets of individuals carrying the same HLA allele (Allen *et al.*, 2005; Carlson *et al.*, 2008; Brumme *et al.*, 2009).

Building on well-documented reversion events, the theory of a universal fitness landscape for HIV was proposed, suggesting that the subtype-specific consensus represents universally beneficial mutations that provide a general advantage to the virus, independent of immune pressure (Zanini et al 2017). Nevertheless, some escape mutations remain stable in the absence of the selecting HLA allele, with the most well-studied example being the KK10 epitope in GAG (Kløverpris, Leslie and Goulder, 2016), raising critical and unanswered questions on the variability of fitness costs across different epitopes and HLA alleles. Compensatory mutations have also been reported that mitigate fitness costs, consequently allowing intrinsically deleterious mutations to propagate at a population level and posing potential hurdles for treatments and vaccines that aim to exploit weak points in the adaptability of the

virus to immune pressure (Leslie *et al.*, 2004; Brockman *et al.*, 2007; Crawford *et al.*, 2007; Schneidewind *et al.*, 2009; Liu *et al.*, 2014).

The observed selection pressure acting on a CTL-linked mutation following transmission to a new host has been shown to determine the degree of conservation of the residue at the population level (Allen *et al.*, 2004; Leslie *et al.*, 2004), and the substantial diversity across subtypes has been said to be partly a reflection of geographical variation in HLA profiles of the infected population (Kist *et al.*, 2020), highlighting the footprint effect of immune-pressure selection on the circuiting virus population and spread of CTL-linked mutations (Matthews *et al.*, 2009) . By quantifying the fates of a broad spectrum of mutations and HLA profiles, we can gain deeper insights into the key drivers of both within and between evolution and immune escape, and more fully understand the balance between diversifying selection of immune-driven adaptations and purifying selection under functional and protein constraints in shaping the viral genome.

The repeated process of escape mutations reverting upon transmission implies that evolution at the between-host scale is not necessarily progressing in one single direction with the virus population continually diverging, despite extensive within-host divergence. As a result, the ‘escape and revert’ hypothesis has been favoured as a leading theory in explaining why HIV has been observed to evolve up to six times faster over calendar time within-host than between-host. Evidence of a bias towards evolution towards the population consensus provides strong support (Herbeck *et al.*, 2006; Raghwani *et al.*, 2018; Druelle and Neher, 2023), however studies specifically evaluating the relative importance of the ‘escape and revert’ dynamic to the rate mismatch have been limited by lack of availability of viral sequences from the source infection or HLA genotype data, and consequently it is not possible to accurately classify transmitted mutations linked to the host CTL response. Additionally, the degree of pre-adaptation to the most common HLA alleles in a population may lessen the impact of the ‘escape and revert’ dynamic on to the tempo of multi-scale evolution.

Studying the complex dynamics of virus evolution under the CTL immune response has long been a challenge due to the limited availability of longitudinal data capturing both source and recipient infections, leading to difficulty in distinguishing between transmitted variants and early escapes. Additionally, detailed HLA profiles of study participants are not always available, and studies have had to rely on broader

demographic data to infer HLA-restricted mutations (Illingworth *et al.*, 2020). When longitudinal sequencing is available, only a small number of infections are included, and larger population-level studies that have detected associations between HLA and viral polymorphisms are affected by shared ancestry between sequences (Carlson *et al.*, 2008). Much of the research in this area is biased towards Europe and the Americas, as reflected by the overrepresentation of subtype-B studies in the LANL database of CTL variants, and research of escape dynamics have been heavily focussed on the narrow group of HLA alleles considered protective. With the current burden of infection heavily placed on sub-Saharan Africa, we need to develop an understanding as to whether the observations and trends previously described are generalisable.

In this study, I have brought together whole-genome longitudinal amino acid sequences of viral populations from over sixty transmission pairs, the HLA profiles of the individuals, and an immunological database of known CTL escape variants. With this comprehensive approach, I identify hundreds of HLA-restricted escape mutations—some previously known and some novel; quantify their fitness following transmission to a new host environment; and determine the relative roles of diversifying selection in response to CTL pressure and purifying selection for viral fitness. I tracked trajectories of escape alleles across source and recipient transmission pairs, from which I observe the 'escape and revert' process in action. Finally, I measured the relative contribution of CTL-driven adaptation to within-host evolution, situating this in the broader context of the evolutionary rate mismatch.

6.3 Methods

6.3.1 Study datasets

All individuals whose samples were analysed were enrolled in the Partners in Prevention PrEP study between July 2008 and November 201. I define "source" individuals as those that were HIV-1 positive when recruited, and "recipient" individuals as those that were initially HIV negative and seroconverted at some point during the study. A total of 1408 serodiscordant pairs were recruited in the study, of which 138 seronegative individuals seroconverted during the study.

Details on the preparation and sequencing of samples can be found in chapter 2, in addition to HLA genotyping information. Here, I only consider long-read PacBio

sequences. While long reads were not necessary for the analysis, I found PacBio provided greater depth.

As this analysis considers protein changes and is performed at the amino acid level, it was necessary to consider sequencing reads, rather than nucleotide base frequencies. The software Phyloscanner was used to generate short sliding windows nucleotide alignments for each transmission pair, with each window 90 base pairs in length and an overlap of 80 nucleotides with the neighbouring window. Windows covered the whole genome from position 520 (with respect to HXB2 sequence). The reads in every window were aligned to HXB2 positions, and so insertions and deleterious relative to HXB2 were excluded from the analysis. The generation of windows was performed for each gene individually to avoid frame shifts within the window. The alignments of every window were then translated to amino acids, considering the specific gene of interest and relevant reading frame. Due to substantial variation across individuals and, as a result, pool alignments between hosts, the variable loops are excluded from the analysis. A read depth cut off of N>10 reads was applied to each alignment window.

Of the 138 transmission pairs in which the sero-negative individual seroconverted during the study, 124 had HLA genotype data for both the recipient and source individual. Of the 248 individuals included in the 124 transmission pairs with HLA data, 226 had at least one sample sequenced on the PacBio platform. For each of the 226 individuals, a consensus sequence was determined for the final sampled time point in order to generate a population-level dataset of viral sequences. For the main analysis, I included all transmission pairs where at least one sample was sequenced for both source and recipient with a minimum depth of 10 reads at each sampling time point, and a source sequenced sample was available that represents a sampling point prior to transmission and no earlier than 6 months before transmission. After this filtering, the main analysis dataset includes 61 transmission pairs.

The frequency of all observed alleles at each position for every time point is calculated from the sequences, resulting in a trajectory for each allele at every time point for both source and recipient. The majority of genome positions will feature in multiple windows, and so only the frequency for the window with the highest number of reads is included in the analysis.

The HIV-1 subtype was determined by the subtype of the best reference selected by Shiver during the sequence processing stage. If more than one subtype was selected for the best reference across different samples of the same individual, the most common best reference was used.

HLA alleles were grouped into supergroups based upon standard grouping definitions (Wang and Claesson, 2014).

6.3.2 Subtype-specific consensus

Consensus protein sequences for every subtype represented in our dataset were retrieved from LANL sequencing database (Los Alamos National Laboratory, 2024b). The consensus sequence was determined from at least 3 sequences representative of the subtype, however most subtype sequences were determined by a substantially larger pool of representative sequences. The subtype consensus sequences were then realigned with MAFFT , maintaining the HXB2 positional references to ensure consistent position mapping across all sequences.

6.3.3 Identifying candidate CTL-escape codon positions

I used the LANL CTL variant and CTL summary databases, downloaded on 18th September 2024, to identify candidate CTL-escape codon positions (Los Alamos National Laboratory, 2024). Some of the common HLA alleles in the study population appear underrepresented in the LANL epitope database, for example the fourth most common group A allele - A*66:01 - has no reported HLA-restricted epitopes. Due to this low representativeness, I identified candidate CTL-escape codon positions in a given individual as all codons that are in an epitope targeted by any of their six HLA alleles (I call these ‘restricted epitope-associated escapes’), or positions that have been previously associated with CTL-immune pressure driven escape, regardless of the associated HLA allele (we call these ‘documented escapes’). Documented escapes were filtered to those classed as Escape (E), Calculated Escape (CE), Observed Escape (OE) and Inferred Escape (IE), and an escape site was defined as any variant allele within a variant epitope in the filtered LANL database, described as a lower-case letter in the variant sequence. After filtering, the documented variant dataset includes a total of 71 positions in *GAG*; 56 in *ENV*; 50 in *POL;* and 49 in *NEF* that are classified as documented escape sites.

6.3.4 Estimating the strength of selection

To determine the presence and strength of selection of an escape allele, a likelihood approach was applied. The method distinguishes selection from noise from sampling and mutations, and is a single-locus version of a method previously described in Illingworth et al 2014 to quantify the selection of haplotypes (Illingworth, Fischer and Mustonen, 2014). Briefly, for each variant, a deterministic model of evolution was applied to describe the frequency of variant, $q(t_{k+1})_i$, at time t_{k+1} given the mutation rate per site per generation $\mu = 3 \times 10^{-5}$ (Mansky, 1996a), selection strength σ_i and the frequency at the previous time point, $q(t_k)_i$. The relationship between $q(t_{k+1})_i$ and $q(t_k)_i$ is described as:

$$q(t_{k+1})_i = \frac{(q(t_k)_i(1-\mu) + \mu(1-q(t_k)_i))\exp(\sigma_i \Delta_{k,k+1})}{\sum_{\boldsymbol{a}}(q^{\boldsymbol{a}}(t_k)_i(1-\mu) + \mu(1-q^{\boldsymbol{a}}(t_k)_i))\exp(\sigma_i \Delta_{k,k+1})}$$

Where $\boldsymbol{a}$ is the set of alleles observed at position *i* and the variant allele is evolving neutrally σ_i=0 or under constant selection $\sigma_i = s$. The likelihood of the observed frequencies given the parameters is given by:

$$\mathcal{L}(\sigma_i, q_0) = \sum_k \left[\log \frac{N_k!}{\prod_a n^a(t_k)!} \prod_a q(t_k)_i^{\,n^a(t_k)}\right]$$

Where N_k is the sample read depth at time point k and $n^a(t_k)$ is the number of reads of allele a at time point k. To prevent numeric overflow, the maximum value of N is fixed at 100. For recipient allele trajectories, the transmitted allele is fixed at frequency 1 at the estimated date of transmission. If an allele initially has a frequency of zero (or one in the case of selection against transmitted escapes), it is assumed that a single mutation emerged at the time point of the last zero sampling, and as such the calculated selection coefficient will be at the lowest end of the possible values. Due to evidence of changes in selection direction within a single infection, I only considered selection in one direction when inferring the selection strength by only considering the period until the maximum observed frequency was reached. A likelihood ratio test was then applied to determine the best model – $\sigma_i = 0$ vs $\sigma_i = s$ - accounting for the number of parameters, and a Bernoulli correction was applied to the p-value threshold to account for the number of tests.

6.4 Results

6.4.1 High HLA diversity among transmission pairs

After initial filtering (see methods) a total of 62 treatment-naïve transmission pairs were included in the analysis dataset, of which 55% were classified as subtype A, 10% as C, 20% as D and the remaining 15% were recombinant forms. The average number of days between the last available sample for a source prior to the first recipient sample was 51 days, however for 28% of the transmission pairs a sequenced sample was available for the same calendar date as the earliest recipient sample.

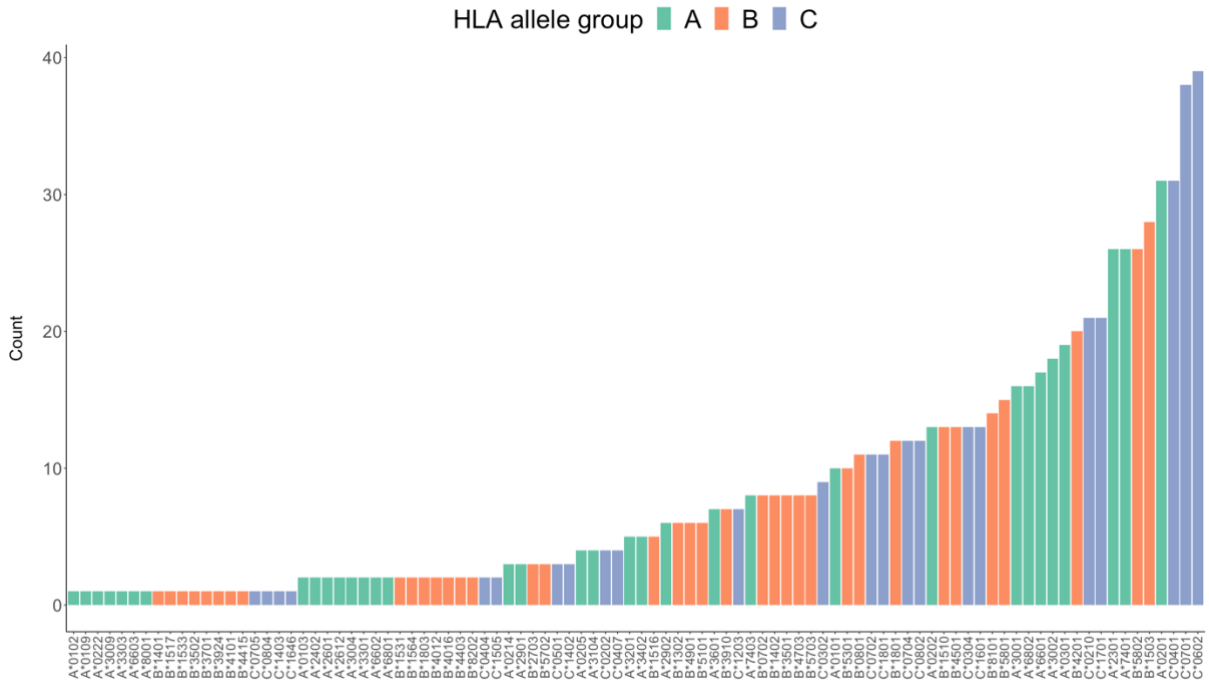

Figure 6.1 Summary of HLA groups A, B and C in the study cohort. Histogram of the count of each HLA-A, B and C allele present in our dataset. The total number of individuals is 122, and so no allele is shared by more than a third of the study participants included in the main analysis dataset. Several high frequency alleles (A*74:01, A*23:01, B*58:02) are poorly represented in the LANL database of CTL variants, showcasing the need for studies of CTL immune escape in sub-Saharan Africa.

The four-digit HLA class I alleles (A, B, and C) were determined for each participant (figure 6.1), with individuals being either heterozygous or homozygous for an HLA group allele. Transmission pairs were considered HLA-matched for a particular HLA group if they shared at least one allele. Among the 62 transmission pairs, 15 matched at HLA-A, 19 at HLA-B, and 25 at HLA-C. The population is diverse, with no allele

shared by more than a third of other individuals and most alleles have substantially lower prevalence. I also find that some of the most common alleles (A*74:01, A*23:01, B*58:02) are poorly represented in the LANL database of CTL epitopes, highlighting the need for improved knowledge on the CTL response and virus evolution in sub-Saharan Africa. To address this limitation and avoid missing important escape mutations, I expanded our search beyond HLA-matched epitopes. Specifically, I identified escape mutations both within known HLA-restricted epitopes and at any position previously documented as an escape site, regardless of HLA association. This dual approach allowed us to capture escape mutations even in understudied HLA contexts.

6.4.2 CTL Escape occurs at all stages of infection with substantial variation in selection strength across escapes

Alleles that differed to the subtype specific consensus – variant alleles - in our sequenced samples at the set of candidate CTL-escape positions were identified, and those that exceeded 5% frequency or N=2 reads within the sample (whichever is larger) were then considered as candidate CTL-escape mutations. These candidate CTL-escape mutations were determined to be CTL-escape mutations if they were either observed to be under positive selection during the study period or inferred to have been under positive selection because they were unlikely to have been transmitted to the individual. Specifically, these were: (1) alleles in source or recipient that demonstrate evidence of evolving under positive selection pressure during the sampling period, and (2) variants present at consensus (>50%) frequency at the earliest sampling timepoint that are assumed to have emerged after transmission, defined for recipient individuals as variants not present in any sample in the corresponding source and for source individuals as variant alleles that are unique to the individual (at consensus level).

Escapes defined in the first category are termed *observed escapes* and in the second *inferred escapes.* An advantage of this approach is that I utilise the evolutionary dynamics to infer selection, as well as a definition of escape that incorporates HLA information and previously verified escape mutations, and therefore we have reasonable confidence that the escape emerged under CTL immune pressure specific to the HLA profile of the host rather than via transmission.

Among recipient individuals, I identified 292 CTL-escape mutations (232 observed and 60 inferred). Of the 62 individuals included in the analysis, 50 individuals drove at least one escape during the first year, with an average of 4 escapes across all individuals, and varying from zero to 24 escapes (figure 6.2A).

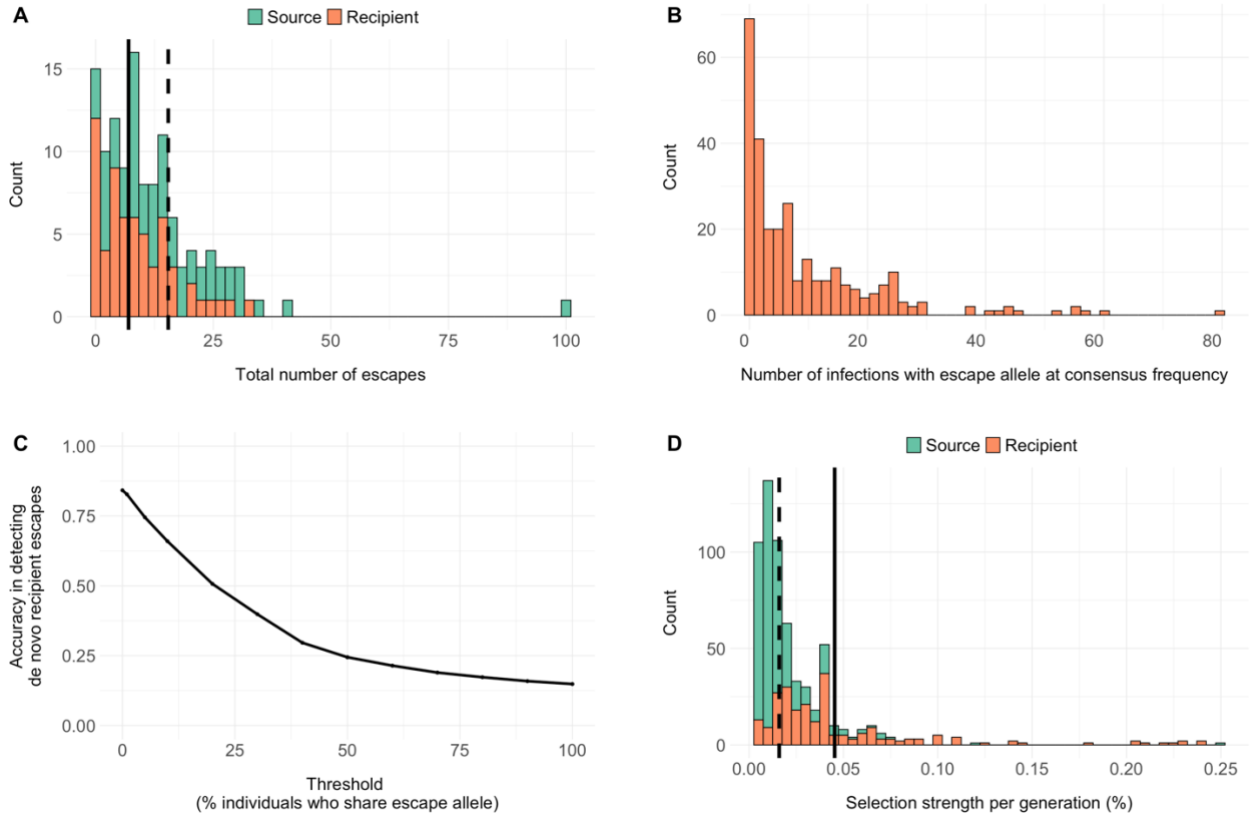

Figure 6.2 CTL escape in recipient and source individuals A) Histogram of the number of escapes in source (green) and recipient (orange) individuals. The average number of escapes identified in a source individual is 14 (dashed black line) and 3 (full black line) in a recipient. B) Histogram of the number of other infections in which an identified escape allele is found at consensus frequency (>50%). For 75% of escape alleles, no other infections in the between-host dataset shared the escape allele, where comparisons were performed at a subtype level if a sufficient number of subtype matching infections (N>10) were available, otherwise the entire dataset was considered. C) Accuracy (ACC) of detecting de novo escape alleles from transmission escape alleles based upon the number of other individuals who share the escape allele at consensus (>50%) frequency, not including the source infection. By applying an increasingly lower threshold to the number of infections of the same subtype outside of the transmission pair that have the same allele at consensus level, we find that we more accurately distinguish de novo escapes to those that were transmitted. D) Distribution of selection strength per generation for each escape identified. The mean strength of escape identified in the recipient group (orange bar, full black line) is substantially higher than for escapes identified in the source group (green bar, dotted black line).

I used the CTL-escape results in recipients to justify our criteria for inferred escapes in source individuals. Specifically, I observed that most CTL escape mutations in recipients are rare at the population level (figure 6.2B), with 75% of escape alleles unique to the individual, and that we could therefore identify inferred escape mutations with 85% accuracy if I only considered unique variants (figure 6.2C). I identified all variants at over 50% frequency within documented CTL sites and HLA restricted epitopes in recipients, disregarding source infection data, and the accuracy of distinguishing *de novo* from transmitted escapes improved as we restricted the set of escapes to variants found in fewer other infections of the same subtype, excluding the corresponding source. I therefore used the criteria that a mutation in a source had to be over 50% frequency and unique to that individual to be an inferred CTL-escape mutation. While the approach will likely exclude some genuine host-driven escapes, we prioritise specificity over sensitivity.

Among source individuals we identified 859 CTL-escape mutations (395 observed and 464 inferred). The number of escape mutations varied dramatically among sources (range: 0-97), likely reflecting both heterogeneity in immune pressure and in time since infection (figure 6. 2A).

The strength of selection acting upon escapes was calculated for both observed and inferred CTL-escape mutations in recipients, but only for observed escape mutations in source individuals. For inferred CTL-escapes in recipient individuals, I assumed the mutation occurred at the estimated time of transmission, and hence the calculated strength of selection for these mutations is likely to be an underestimate. I could not calculate the strength of selection for inferred escapes in source individuals as I do not have information on when the source individual was infected. The estimated strength of selection varies by orders of magnitude in both source and recipient individuals (figure 6.2D), however the majority of escapes are evolving under relatively weak selection, with an average of 4.5% selection strength per generation in the recipient group and 1.6% selection strength per generation in the source group, meaning a variant initially at 5% frequency will reach 95% frequency after 8 months and 21 months respectively, assuming no change in selection or interference effects.

The escapes evolving under the strongest selection were detected in recipient individuals, it is however challenging to make a direct comparison in the fitness landscapes across the groups due to less frequent sampling of source individuals. By

simulating trajectories with known selection strength and observation times sampled from the source and recipient sampling frequencies, we found that the higher rate in the recipient group cannot be explained by differences in sampling frequency alone. To establish whether the selection acting on CTL escapes is stronger than other evolutionary forces, we compared the distribution of selection strengths at CTL sites to other polymorphic sites (MAF greater than 5%) for both source and recipient groups, and found CTL escape mutations to be evolving under significantly stronger selection (Mann-Whitney test $P<10^{-5}$).

6.4.3 Shared Selection Patterns Among HLA-Matched Individuals

CTL escape mutations in HIV can often be predicted based on the host's HLA profile, with certain escape mutations repeatedly emerging in individuals who share specific HLA alleles. While these associations between HLA alleles and escape mutations are well documented, we lack a complete understanding of escape pathway patterns and their consistency across individuals. Here, while 75% of escape mutations are unique to individual patients when compared to consensus-level data of our study population, suggesting substantial heterogeneity in escape pathways, the codon positions where these escapes occur show significant consistency. I find that 50% of all sites where escapes are observed experience selective pressure in multiple individuals (Figure 6.3A). This pattern suggests that while the specific mutations may vary, the locations where HIV can successfully escape immune pressure are constrained, likely reflecting a balance between immune evasion and viral fitness.

Next, I considered whether the same epitopes were under selection for escape in individuals with similar HLA profiles. When I examined individuals sharing HLA supergroups - sets of HLA alleles with similar peptide-binding properties - I found some shared escape pathways, for example a documented variant of the epitope located at HXB2 position 530-538 in POL was inferred or observed to be under selection in 9 of the 20 individuals in the clinically relevant B57 supergroup (B*57:01, B*57:02, B*57:03, B*58:01)(Miura *et al.*, 2009) , in addition to the well-studied escape variant epitope in GAG at HXB2 position 240-249 (TW10) that was under selection in 7 of the 20 B57 supergroup individuals. However, I do not uncover any escapes under selection in the majority of the individuals within an HLA supergroup, highlighting that not all potential epitopes are targeted and that the timing and strength of the CTL response may differ across individuals. This variation in escape patterns, even within well-characterised

HLA-epitope combinations like B57-TW10, highlights the complexity of host-virus interactions and suggests that HLA type alone may not be sufficient to predict viral evolution.

To investigate the broader relationship between HLA profiles and escape mutations, I examined pairs of individuals who developed identical escape variants (the same amino acid change at the same position). For each such pair, I calculated their HLA similarity by counting their shared alleles, both overall and specifically within HLA-A, B, and C groups. To determine if this degree of similarity in HLA profiles was greater than expected by chance, I implemented a randomisation approach. I repeatedly sampled random pairs of individuals from our dataset and calculated their HLA similarity, generating a null distribution of expected HLA sharing. Specifically, I sampled 100 random pairs to calculate a mean sharing value and repeated this process 1,000 times to generate a distribution of expected means. I then compared the observed HLA sharing among individuals with identical escapes to this null distribution and found that pairs of individuals who developed identical escape mutations shared significantly more HLA alleles than expected by chance (figure 6.3C). To ensure this association was not simply an artifact of our escape mutation identification method, which included searching known HLA-restricted epitopes, I repeated the comparison using only escape mutations identified based on their location at documented escape sites, without considering HLA information. The significant enrichment of HLA sharing was replicated in the supporting analysis, suggesting a biological association between shared HLA alleles and convergent escape evolution (supp. Figure 9.7). When considering pairs of individuals at which the same site is under selection but the allele may differ, there was no significant difference in the proportion of shared alleles with the null distribution, suggesting that while specific escape mutations are associated with shared HLA alleles, the position of escape mutations alone is not predictive of HLA sharing.

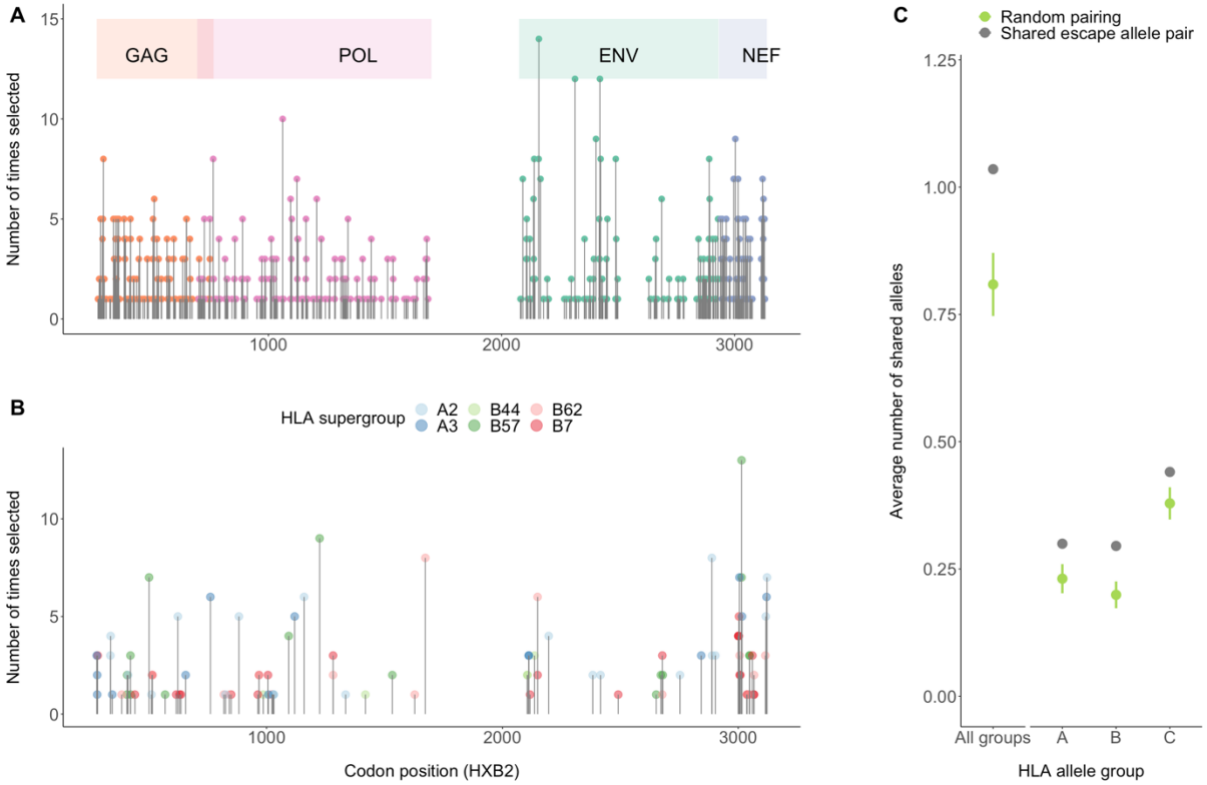

Figure 6.3 Recurrent escape among individuals. A) The number of times each genomic position was inferred to be under selection for CTL escape during the observation period or inferred to have been under selection prior to the first sample. B) The number of times a documented epitope associated with one of 6 HLA supergroups was associated to be under selection in an individual carrying an allele within the relevant supergroup. C) The average number of shared alleles between a pair sharing an escape allele for all HLA groups and for HLA-A, B and C specifically. Green bars indicate confidence intervals of the mean number of shared alleles in 1,000 sets of 100 randomly selected pairs of individuals from the study (NB: not equivalent to transmission pair), representing a null distribution. The average number of shared alleles in pairs that share an allele is significantly higher than for the null comparator data.

6.4.4 Evidence of shifting selection pressures between and within hosts

CTL escape mutations can revert upon transmission to new hosts, but the impact of this mechanism on long-term evolutionary dynamics partly depends on the transmission chain length between escape and reversion events. In other words, if reversions occur in the subsequent infection, then CTL escape will have minimal effect on between-host evolution, while if many transmission events have occurred between escape and reversion of the escape, then escape will have greater influence on the accumulation of mutations over longer time scales at the population level.

Our dataset provides direct evidence of escape-reversion dynamics occurring across single transmission events. I identified multiple cases where a CTL-associated mutation under positive selection in the source individual was present in the

transmitted virus, but subsequently reverted to the subtype consensus in the recipient (supp. Table 9.1). Among these cases, several illustrative examples of this "toggling" pattern include: the T303I mutation in GAG, associated with the HXB2 epitope (positions 298-306) targeted by HLA-B*14:02, expressed by the source but not their transmission recipient; the K347E mutation in ENV, within the HXB2 epitope (positions 341-349) restricted by HLA-A*02:01, carried by the source but not their recipient; and the A83V mutation in GAG, within the HXB2 epitope (positions 77-85) associated with HLA-A*02, again present in the source but absent in their paired recipient (figure 6.4A-C). These and other cases provide compelling evidence for host-specific fitness trade-offs, demonstrating how CTL escape mutations beneficial in one host can be rapidly selected against when transmitted to a recipient lacking the relevant HLA allele.

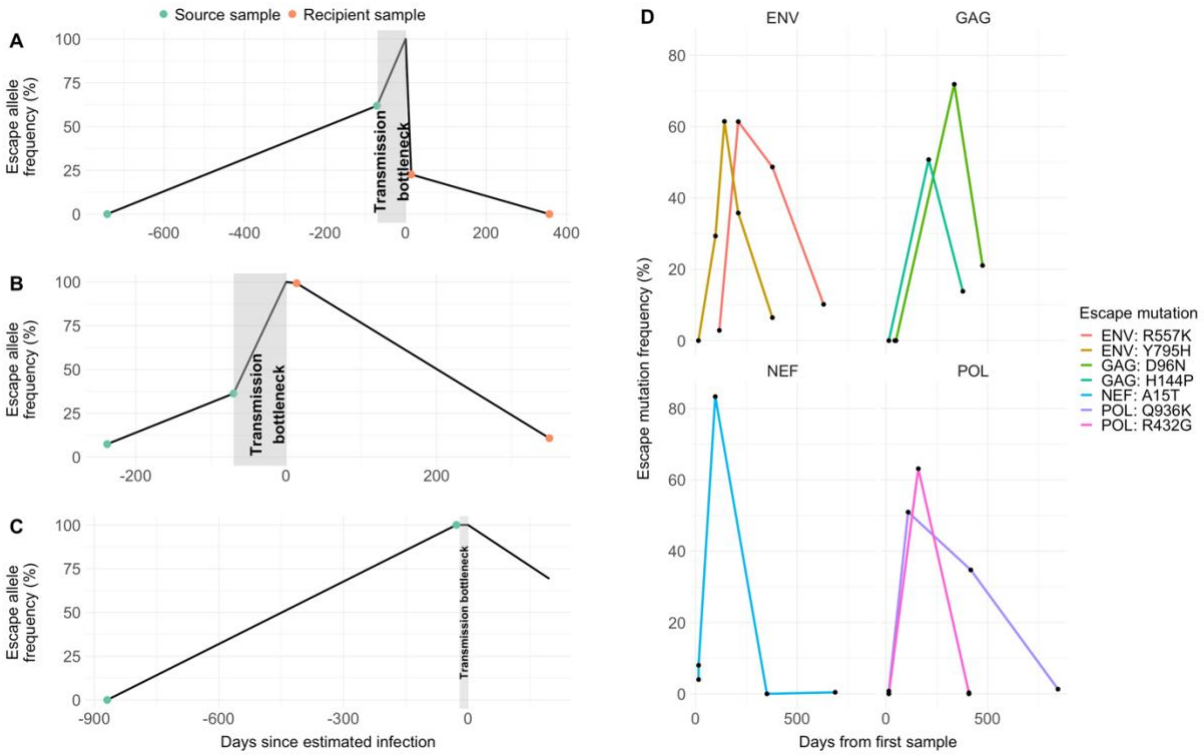

Figure 6.4 Change in selection pressure between and within host. A-C) Examples to the 'escape and revert' process across a single transmission event. Three escapes (A: GAG T303I HLA-B*14:02, B: ENV K347E HLA-A*02:01C: GAG A83V HLA-A*02) are under selection in the source individual, pass through the transmission bottleneck (shaded grey) in which their frequency is assumed to reach 100%. In the recipient who does not share the corresponding HLA allele, the mutation is under negative selection in favour the subtype-specific allele. D) The within-host trajectories of 7 escape mutations that are initially selected for but are later selected against, often in favour of a mutation within the same epitope.

In addition to the toggling of mutations across transmission events, we also observe multiple instances of within-host toggling in both source and recipient individuals (figure 6.4D). Such mutations appear to be only transiently beneficial, and we hypothesise that a change in the selective advantage is driven by either the emergence of a fitter escape in the same or different epitope, or a change in the host immune response. We detected 7 *de novo* escapes that are later selected against in favour of the initial consensus allele, observed across 7 recipient individuals (figure 6.4D), in addition to 8 escapes in source individuals (supp. Figure 9.8). Two recipient escapes – Y795H in ENV and R432Q in POL – were selected against in favour of a mutation within the same epitope, and in both cases the two variants have previously been identified as CTL escapes in individuals possessing an allele carried by the individual we here observe as driving the escape, providing biological support that the toggling is indeed a consequence of adaptation to the CTL immune response of the host. A further two escapes – H144P in GAG and Q936K in POL – were selected against at the same time as an escape emerged within close genomic proximity, specifically position 145 (GAG) and 934 (POL) respectively. The observed toggling behaviour appears to be influenced, at least in part, by competition between mutations within the same epitope, where fitter escape variants can emerge and outcompete prior ones, and overall depicts a highly dynamic landscape of immune selection that shifts within short time scales

6.4.5 Broad distribution of fitness costs of transmitted escape mutations

A major advantage of the study is our ability to track escape variants across the HIV transmission interface. An escape variant detected in the source individual was classified as transmitted if it was present (the larger of $\geq 5\%$ frequency or ≥ 2 reads) prior to transmission in a sample taken within 6 months prior to the estimated transmission date and was observed at greater than 5% frequency or $N \geq 2$ reads (whichever is larger) in the sample corresponding to the earliest time point in the recipient. In total, I found 315 source escape mutations with evidence of transmission.

Reversions of escape mutations are relatively rare, with only 10% having evidence of being selected against following transmission. This may be an underestimate if mutations are under extremely strong negative selection immediately after transmission, however given the short time window between transmission and the first sample, it is unlikely to lead to a significant difference. As a comparison and to check

whether unidentified escape mutations would increase the reversion rate, we determined the selection strength of all variant alleles (differing from subtype) that are transmitted and found that a significantly lower proportion are under negative selection (5%, Mann-Whitney P<10^{-3}), and the selection strength is significantly lower (Mann-Whitney P<10^{-3}).

The fitness costs of transmitted escapes showed clear relationships with HLA matching between source and recipient. Escape mutations transmitted between individuals mismatched for HLA-B alleles showed significantly higher fitness costs (Mann Whitney P=0.03) (Figure 6.5B). Similarly, HLA-C mismatches were associated with higher fitness costs (P=0.05) (6.5C). However, HLA-A matching status showed no significant effect on fitness costs (Figure 6.5A). These patterns are consistent with the hypothesis that escape mutations optimised for one HLA context may be costly when transmitted to hosts with different HLA profiles, particularly for variants associated with HLA-B and -C group alleles.

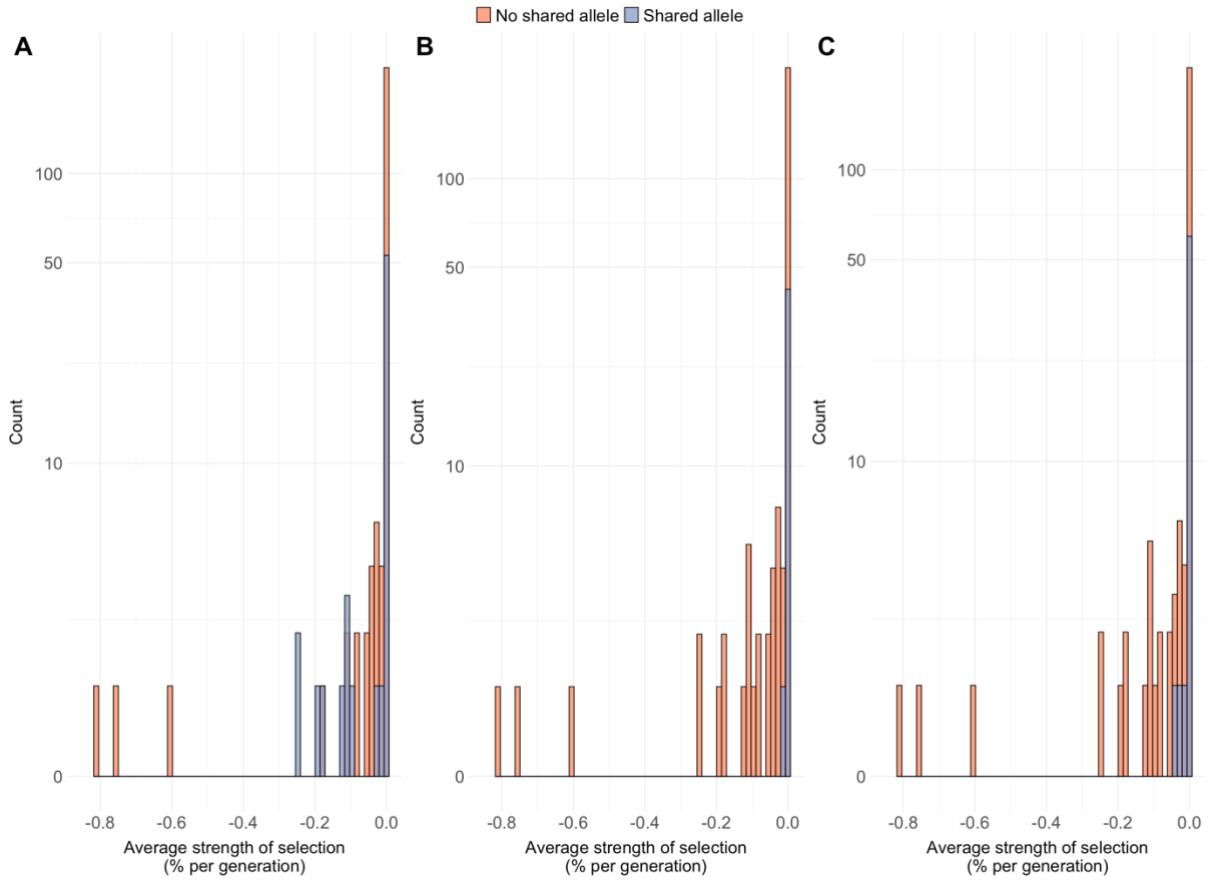

Figure 6.5 The distribution of fitness costs of transmitted escape mutations. For each transmitted escape mutation evidence of selection was determined via the likelihood approach applied to inferring escapes. For transmitted escapes inferred to be evolving under selection, the selection strength was inferred from the same model, where an initial mutation was assumed to have occurred immediately after transmission. We find the majority (90%) of mutations to show no evidence of negative selection. Escapes transmitted between individuals with matching HLA-A alleles did not show a significant difference in fitness to those mismatched in HLA-A (A), however selection is significantly greater in escapes transmitted within individuals who do not share an HLA-B allele (B), compared to those who do (Mann-Whitney P=0.03), as escapes are more likely to continue to be beneficial across individuals who share HLA alleles as they present the same viral epitopes. We also see a significantly higher fitness costs in HLA-C mismatched individuals (C) (Mann-Whitney P=0.05).

6.4.6 More frequent and costlier escape associated with protective HLA alleles

Protective HLA alleles typically present conserved viral epitopes and generate robust CTL immune responses, suggesting associated escape mutations may carry larger fitness costs and emerge earlier in infection. I investigated whether escape pathways differ in individuals carrying protective alleles, focusing on three well-characterized protective HLA alleles associated with delayed disease progression in Sub-Saharan African populations: B*57:03, B*58:01, and B*81:01(Miura *et al.*, 2009). In the recipient group, 15 of the 61 individuals carried a protective allele, and 19 of the 61 individuals in the source group.

I first analysed escape dynamics in the recipient group using survival analysis to account for variable follow-up periods. I defined escape as the time until the first escape mutation reached consensus frequency (>50%) and applied right censoring when no escape was observed. The mean time to first escape across all individuals was 270 days. Stratifying by protective allele status revealed significantly accelerated escape in individuals carrying at least one protective allele (log rank test p=0.04; Figure 6.6A).

Since infection timing was unknown for the source group, I analysed escape burden instead of timing. Source individuals with protective HLA alleles harboured significantly more escape mutations (Mann-Whitney P=0.05). Notably, three individuals with protective HLA alleles maintained typical viral loads (10^4-10^5 viral copies per mL) despite accumulating >30 escapes, suggesting these variants successfully evaded CTL pressure while maintaining viral fitness through either intrinsically low fitness costs or compensatory mutations.

To assess fitness costs of escape mutations, I examined selection against transmitted escapes that emerged in source individuals with protective HLA alleles. I defined transmitted variants as those present at >5% frequency or above sequencing noise threshold N at the recipient's first sampling timepoint. Escapes transmitted from individuals with protective alleles showed significantly stronger negative selection compared to those from individuals without protective alleles (Mann-whitney test P=1.25×10^{-6}, mean selection strength -2.4×10^{-2} and -7.8×10^{-3} per generation respectively).

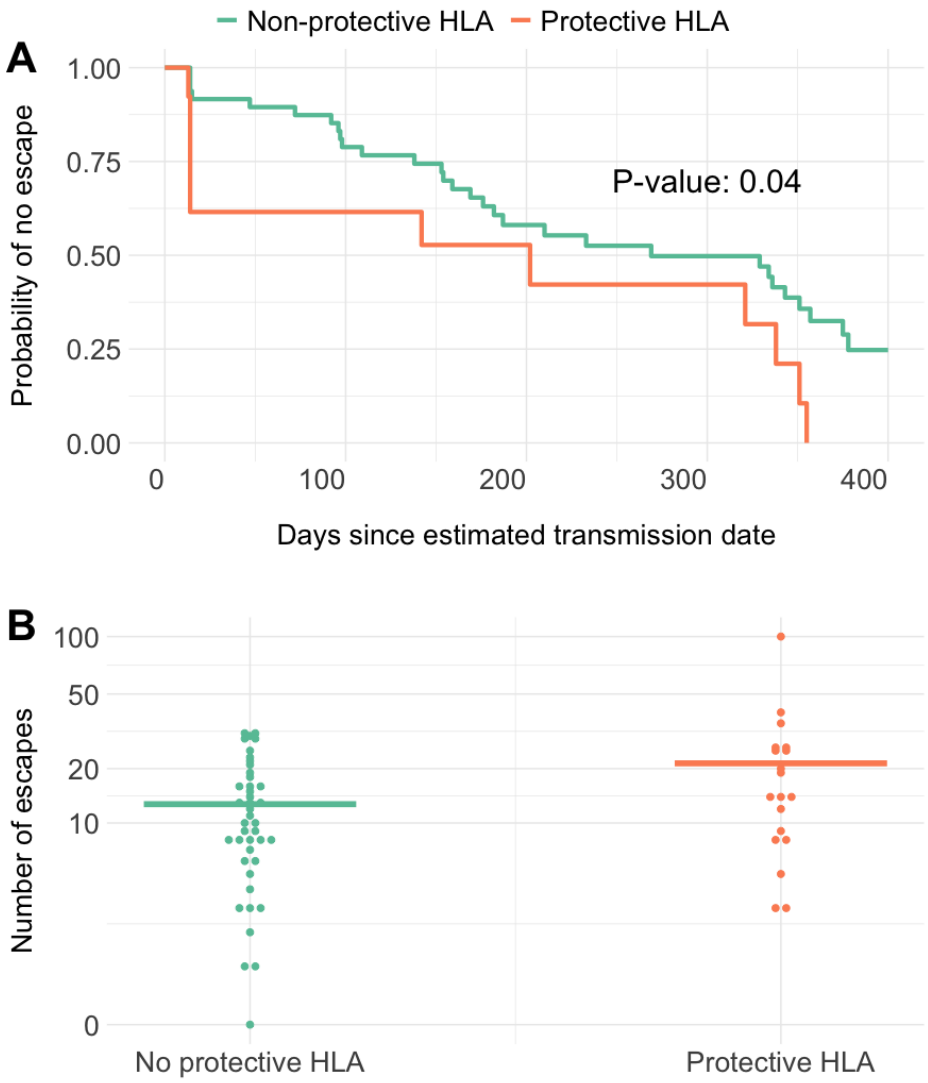

Figure 6.6 Faster and more frequent escape in infections of individuals with protective HLA-alleles. A) Kaplan Meier estimators for time until first escape for recipient individuals grouped by HLA-protective status. Protective HLAs are B*57:03, B*81:01, B*58:01 (). A log rank test found the time until escape to be significantly lower in the HLA-protective group. B) The number of inferred and observed escapes in source individuals stratified by HLA-protective status. The average (mean) is indicated by full lines (no protective HLA: 13 escapes, protective escapes: 21 escape, Mann-Whitney P=0.02).

6.4.7 CTL escape is the dominant force of evolution

While the analysis of reversion events has been conservative, focusing only on transmitted escapes during early infection, the broader evolutionary implications of CTL escape and reversion are likely substantial. Two key observations suggest that most CTL escapes must eventually revert or be lost after transmission through HLA-mismatched hosts. First, I find that most escape mutations are unique to individual hosts when considering the population-level diversity. Second, any escape mutations

that do not carry fitness costs would be expected to have reached high frequency in the circulating viral population and would therefore not be identified in our analysis. To understand how this escape-reversion cycle contributes to viral evolution, I quantified the relative importance of CTL-mediated evolution compared to other evolutionary forces.

I measured the total amount of evolution during early infection (~first 12 months) across all recipient infections, comparing sites associated with CTL escape (by definition also includes all identified reversions) to other polymorphic sites, termed the "background". Specifically, I looked at all *de novo* mutations, identified as those that emerged within the recipient and exceeded 5% frequency, and identified those with evidence of evolving under selection. To comprehensively capture CTL-associated evolution, I defined CTL-associated sites as any genomic position where an escape mutation was identified in our study, allowing me to detect potential reversions of escape mutations that were transmitted from earlier in the infection chain.

For each group of sites – CTL versus non-CTL sites - I summed the number of sites identified as under selection, weighted by the final observed frequency of the allele. I found that CTL escape accounts for approximately 1.5-3 times more genomic change than other drivers of evolution, with the exception of ENV where we expect escape from the antibody response to also be an important evolutionary driver (figure 6.7). For GAG, POL and NEF, if we make the simplifying assumption that all escapes eventually revert and that between-host evolution only consists of background evolution, this would imply an approximately 2.5-4 times higher within-host compared to between-host rate of evolution.

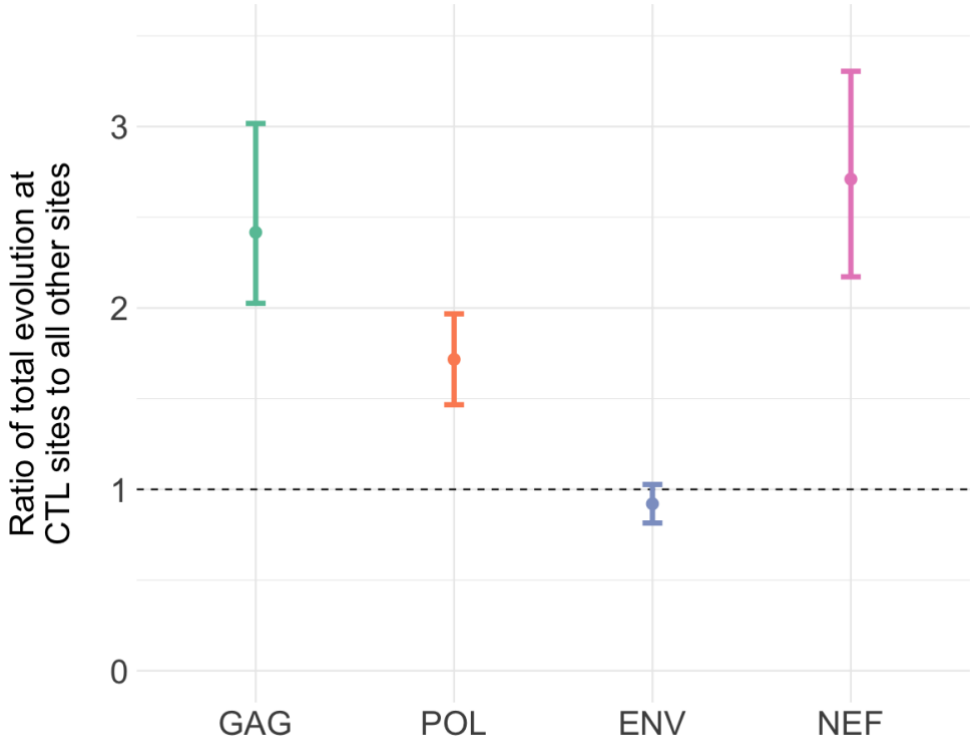

Figure 6.7 Total contribution of CTL evolution compared to "background" evolution at the end of ~ 1 year of infection. De novo recipient mutations identified as evolving under selection were classified as linked to CTL escape if they were located at a codon position at which an escape was identified in any individual in the study, to account for transmitted escapes that fixed further back in the infection of the source or in the transmission change. Other sites were classified as "background". Total evolution was quantified by summing the number of sites, with each site weighted by the frequency of the variant allele at the final time point. 1000 Bootstrap replicates of the datasets were generated, and the total evolution quantified in each replicate to generate confidence intervals, denote by bars. In GAG, POL and NEF, CTL contributed the majority of the genomic change at the amino acid level. In ENV, the contribution is largely balanced in the two groups, likely due to antibody escape.

6.5 Discussion

Viral adaptation to host CTL pressure is a major driver of HIV within-host evolution, yet our understanding of the balance between viral fitness and immune escape and the consequences for viral evolution at the population level remains incomplete. Previous studies have been constrained by small study cohorts, no HLA data or source sampling, or relying on consensus sequencing (Brockman *et al.*, 2007; Li *et al.*, 2007, 2007; Brumme *et al.*, 2009; Boutwell *et al.*, 2010; Roberts *et al.*, 2015; Illingworth *et al.*, 2020). This study addresses these limitations by examining CTL escape in a large, genetically diverse cohort of transmission pairs from an understudied population. By tracking viral evolution both within hosts and across transmission events, I demonstrate that CTL escape remains important throughout infection, not just in early

stages. Moreover, I described the distribution of fitness costs for transmitted mutations during early infection, linking this specifically to HLA profiles of the source and recipients, providing direct evidence for how immune-driven adaptations impact viral fitness in different host environments.

This analysis revealed widespread CTL escape across individuals during both early and late infection. Across studies, consensus on the timing and prevalence of CTL escape has not been reached, which is most likely a reflection of significant heterogeneity across individuals, challenges in detangling T/F effects with *de novo* escape and a large bias towards a small set of epitopes and protective HLA alleles (Goulder and Watkins, 2004; Brumme *et al.*, 2008; Fryer *et al.*, 2010; Roberts *et al.*, 2015). Rapid escape during the first weeks and months of infection has been reported, with escape linked to the decline in viral load at the end of acute infection (Wood *et al.*, 2009; Henn *et al.*, 2012). Other studies have revealed the escape process is complex, often multiple mutations on an epitope are required before it is selected for (Oxenius *et al.*, 2004; Goonetilleke *et al.*, 2009), and in accounting for transmitted escapes one study showed escape occurred in only a minority of individuals during the first few years of infection (Roberts *et al.*, 2015). By comparing viral populations between source and recipient, our study provides a unique window into escape dynamics, allowing clear discrimination between transmitted and *de novo* mutations. I found that CTL escape was the dominant evolutionary force, with selection strength for escape particularly pronounced in early infection.

A 2010 study by Fryer *et al*. examined the dynamics of CTL escape and reversion across 28 epitopes in the GAG, POL (RT), and NEF proteins. The study concluded that both escape and reversion generally occur slowly. Rates were estimated using a model informed by cross-sectional population data from subtype B infections and were further validated using an independent longitudinal dataset. The median time to escape in HLA-matched individuals was approximately 8 years, with an interquartile range of 1.8 to 34 years across the epitopes studied.

While my findings show that escape can occur rapidly—sometimes reaching fixation within weeks or months after seroconversion—this does not necessarily contradict the conclusions of Fryer *et al.* Their reported rates represent averages across individuals whose HLA types match the restricting epitope, and thus reflect both the frequency and timing of specific escape events. If escape does not occur in every HLA-matched

host, the average rate will appear slower. Indeed, my results also show that many escape events are rare; when considering escape rates across all individuals with the relevant HLA type, the average time to escape is therefore expected to be long. Notably, Fryer *et al.* also observed early escape in several individuals within their longitudinal dataset, consistent with the early escape dynamics I report.

Comparing the frequency and timing of escape across studies requires consistency in how escape is defined and measured—whether it is the time for an individual to fix an escape mutation, or the population-level rate at which a specific escape arises across HLA-matched individuals. Differences in genome regions analysed, cohort composition, and geographic location (due to HLA allele frequencies) further complicate direct comparisons between studies.

It is important to note that the escapes in this study have not been verified with functional evidence, such as CTL recognition assays. While computational identification of escape mutations has inherent uncertainty, multiple lines of evidence support the biological relevance of our findings. The observed relationships between HLA profiles and escape patterns - including escape timing, frequency per individual, and consistency across HLA supergroups - indicate that we have captured the fundamental dynamics of CTL-driven viral evolution.

I found that the same codons were repeatedly under selection across multiple individuals, confirming observations from smaller cohort studies (Illingworth *et al.*, 2020). This pattern aligns with an "escape and revert" dynamic where the same codons - and therefore epitopes - come under selection in individuals with similar HLA profiles. The previously observed predictability of CTL escape (Moore *et al.*, 2002; Leslie *et al.*, 2004; Carlson *et al.*, 2008, 2012) was further supported by selection on shared epitopes within HLA supergroups, and individuals with shared HLA alleles driving identical escape mutations more frequently than expected by chance. However, escape was largely host-specific - 75% of escape mutations were unique to individual hosts when examining consensus sequences from our cohort. Overall, while certain CTL escape pathways appear predictable and shared across hosts, the majority of escape mutations represent individual-specific solutions to immune pressure.

This pattern of shared selection sites but unique escape mutations reveals two important insights. First, the rarity of identical transmitted escape mutations suggests they generally do not persist long enough in the population to be repeatedly transmitted, supporting a model where most escapes ultimately revert. Second, the diversity of mutations at repeatedly targeted sites demonstrates HIV's remarkable plasticity - even when constrained to specific genomic regions under immune pressure, the virus can explore multiple mutational pathways to achieve escape. This flexibility in escape routes likely contributes to HIV's success in adapting to diverse host immune environments.

While I detected widespread CTL escape, understanding the fitness implications of these adaptations and long-term escape trajectories remains challenging. The hypothesis that frequent reversion of host-specific adaptations constrains between-host evolution has been built primarily on evidence from a limited number of well-studied epitopes, particularly in GAG, where fitness has been demonstrated *in-vivo* and experimentally (Loh *et al.*, 2007; Miura *et al.*, 2009; Troyer *et al.*, 2009). Most notably, the T242N mutation in TW10 is selected for in the presence of B*57/5801 allele and reverts in the absence of the host allele (Leslie *et al.*, 2004; Chopera *et al.*, 2008), however reversion of an escape in the KK10 epitope associated with HLA-B*27 has been shown to be rare (Goulder *et al.*, 2001). Previous work has also shown that networks of compensatory mutations can preserve viral fitness while maintaining immune escape, particularly in GAG where secondary mutations can restore protein folding and function (Brockman *et al.*, 2007; Crawford, Prado, Leslie, Hué, Honeyborne, Reddy, Van Der Stok, *et al.*, 2007; Liu *et al.*, 2014).

Approximately 90% of transmitted escape mutations persisted in recipients throughout the first year of infection without evidence of negative selection. Notably, the strongest signals of fitness costs were observed for escapes transmitted between HLA-B and C mismatched pairs and for escapes from donors with protective HLA alleles, suggesting these alleles may drive particularly costly escape pathways. The persistence of many escape mutations may also reflect their potential benefit across multiple HLA backgrounds, particularly within HLA supergroups that share similar peptide binding properties. However, the limited reversion we observed within the first year indicates that the process of restoring optimal viral fitness following transmission may occur over longer timescales than previously appreciated, possibly across multiple transmission

events. Longer-term studies of chronic infections are needed to fully understand these dynamics.

By analysing changes in allele frequency over time, it was possible to estimate the strength of selection acting on escape mutations in both source and recipient individuals. Overall, selection was stronger in recipients, likely due to differences in host immune response during the first year of infection, although the majority of mutations in both groups evolved under relatively weak selection pressure. The distribution of inferred selection coefficients was broadly consistent with previous estimates reported by Illingworth et al. (2020), who applied a similar likelihood-based inference approach. In their analysis of CTL and antibody escape mutations in *p24* (Gag) and *gp41* (Env), 65% of variants had selection coefficients below 5% per viral generation. In comparison, I found an even higher proportion of variants under this threshold—approximately 77% in recipients and 96% in sources. This difference is not unexpected, given the broader genomic scope of my analysis, which included additional, more functionally constrained genes, and focused solely on CTL escape rather than both CTL and antibody-mediated escape. Furthermore, I used a slightly shorter generation time (1.8 days) than the 2-day estimate employed by Illingworth *et al.*, which may also contribute to differences in inferred selection strengths.

It is important to note that direct comparisons across studies are complicated by differences in sampling frequency and resolution limits. The sampling intervals in this dataset—averaging approximately three months—reduce the ability to detect rapid allele frequency shifts, particularly those driven by strong selection. As a result, the strength of selection at fast-evolving sites is likely to be underestimated. This limitation is shared by other studies. For instance, an investigation of non-synonymous mutations in the *C2–V5* region of *env* reported substantially lower selection coefficients (all below 1% per generation) (Neher and Leitner, 2010), although this likely reflects the low temporal resolution of their sampling strategy, with sampling only every 6 months.

Other approaches, such as the 2017 genome-wide study of fitness costs (Zanini *et al*, 2017), have circumvented this issue by inferring selection under the assumption of a mutation–selection balance. That study found that nearly half of all non-synonymous mutations across the genome carried fitness costs exceeding 10%. While a direct comparison is difficult—given their focus on purifying rather than positive selection and

distinct approach that is not limited by temporal resolution —these findings suggest that current estimates of selection strength, including those reported here, may substantially underestimate the true in vivo dynamics of selection at certain loci.

A key finding of our study is direct evidence for "escape and revert" dynamics within transmission pairs. I observed cases where an escape mutation associated with a source's HLA allele emerged under positive selection, was transmitted to the recipient who did not share the HLA allele, and then experienced negative selection in the new host. We term this pattern "between-host toggling," providing direct empirical support for a process previously inferred primarily from population-level trends and theoretical arguments about evolution toward subtype-specific optimal sequences (Zanini *et al.*, 2015; Druelle and Neher, 2023). We quantified CTL-associated evolution during early infection and found it accounts for 1.5-3 times more genomic change than background evolution across most viral genes. This substantial contribution of immune-driven toggling to within-host evolution, combined with the likelihood of eventual reversion, helps explain the longstanding puzzle of higher evolutionary rates observed within hosts compared to between hosts.

Complementing this between-host pattern, I also identified "within-host toggling," where escape mutations fluctuated in frequency within individual hosts over time, occurring in both early and chronic infection. These within-host dynamics appear driven by competition between different escape variants both within and across epitopes, highlighting the complexity of HIV's fitness landscape and the dynamic nature of immune escape.

The conclusions regarding fitness costs are limited by the challenge of accounting for compensatory mutations. Identifying interacting mutations, both within and outside of variant epitopes, remains technically challenging for a study cohort of this size. However, emerging machine learning approaches are beginning to provide new tools for understanding these complex mutational interactions (Haddox *et al.*, 2018; Blassel *et al.*, 2021). Future work combining these computational approaches with longitudinal data will be crucial for mapping HIV's complex fitness landscape and understanding the interaction between escape and compensatory mutations, which will be fundamental for leveraging fitness costs for CTL-based vaccine designs (Goulder and Watkins, 2004).

By tracking viral evolution across transmission pairs, our study sheds light on the intricate balance between immune escape and viral fitness in HIV infection, while highlighting the role of host-specific selection in shaping population-level evolution. These findings carry implications for vaccine design. Specifically, targeting epitopes where escape mutations incur substantial fitness costs—particularly those associated with protective HLA alleles—may offer the most effective strategy. Notably, a large proportion of escape mutations appear to have minimal fitness costs in the short term. Future research combining longitudinal deep sequencing with advanced computational approaches will be critical for fully mapping HIV's fitness landscape and unravelling compensatory mutation networks. Such insights will be instrumental in developing vaccines that leverage fitness constraints to control HIV and could also inform therapeutic strategies for other rapidly evolving pathogens.

7 Discussion

7.1 Summary of findings

The within-host evolutionary dynamics of HIV represent a complex interplay of frequent recombination, navigation of vast fitness landscapes, and continuous adaptation to multifaceted immune pressures. Understanding how these factors shape viral populations and influence HIV's long-term persistence at pandemic scale is crucial for developing effective public health interventions, therapeutics, and vaccines. In this thesis, I explore the evolutionary forces driving within-host viral dynamics, examining how selection and recombination shape within-host viral populations and their relevance in adaptation at the population level over decades. The analyses I have presented demonstrate how key features of viral sequence data - including sampling frequency and read length - as well as the choice of analysis methodology approach can fundamentally affect our interpretation of viral evolutionary dynamics, highlighting the importance of careful methodological consideration in studying viral evolution.

In chapter 3, I introduced the concept of multi-scale evolution with a well-studied example of where short-term selection pressures within-host conflict with between-host transmission fitness: the evolution of virulence. Using an adapted nested model framework, I examined quasispecies evolution by modelling multiple genomic sites affecting viral load, where the fitness impact of mutations decreased as the number of segregating sites increased. At any moment during infection, viral load was determined by the fitness of the quasispecies. This framework provided a parsimonious mechanistic explanation for several established observations in HIV research: the emergence of set-point viral load, the heritability of viral characteristics across transmission pairs, and the challenge of identifying viral mutations associated with spVL in genome-wide association studies.

Moving beyond mutation as the only source of genetic diversity, I focussed on the role of recombination in chapter 4, with a data-driven approach leveraging longitudinal sequencing data from hundreds of untreated infections captured during early or late infection. Using both simulated and real sequencing data, I demonstrated that the accuracy of recombination rate estimates is influenced by the interaction between sampling intervals, read lengths, and the true underlying recombination rate. I observed substantial differences in the effective recombination rate across infection

stages, viral subtypes, and viral loads. By estimating recombination in sliding windows across the genome, I found hot and cold spots for recombination that were consistent across with previously identified hot and cold spots at the inter-subtype level. Additionally, by analysing codon positions separately, I found significant variation in effective recombination rates across different codon positions in ENV and POL. This observation underscores the potential role of selection pressures in shaping recombination dynamics, potentially driven by epistatic interactions.

In chapter 5, I focussed on the rate of evolution across scales. Using two approaches (BEAST coalescent time trees vs genetic divergence), I quantified evolutionary rates for GAG, POL and ENV. BEAST consistently yielded higher rate estimates than divergence-based approaches. Notably, evolutionary rates were correlated across genes within individuals, suggesting host-specific or viral factors that modulate genome-wide evolutionary rates. By comparing within-host rates to between-host evolutionary rates, I confirmed and extended previous evidence of elevated evolutionary rates at the within-host scale, a pattern consistent across the genome and across codon positions. I proposed that mutational toggling—where variants repeatedly rise to high frequency before declining—drives the elevated within-host evolutionary rates and helps explain the evolutionary rate mismatch between scales and between methods. Supporting this hypothesis, I demonstrate frequent *de novo* mutation toggling, at both synonymous and non-synonymous sites.

In addition to the contributions of within-host toggling, I explored the extent of between-host toggling across transmission events in chapter 6. By incorporating HLA genotypes, sequencing data from source and recipient in a transmission pair, and existing literature on CTL escape, I identified hundreds of likely CTL escapes across 62 transmission pairs. By tracking mutation trajectories across transmission events, I quantified the fitness costs of escape mutations following a change in the host immunological profile, with higher fitness costs linked to escape driven in individuals with protective HLA alleles, and selection against transmitted escapes more likely when source and recipient differed in their HLA-B and C alleles. Additionally, I documented clear examples of the "escape and revert" process within single transmission events, where escape mutations selected in response to one individual's HLA profile were subsequently selected against in a recipient lacking that HLA type. Finally, I demonstrated that within-host toggling can be driven by competition within or

between epitopes, providing a mechanistic explanation for the within-host toggling of non-synonymous mutations in chapter 5.

7.2 Key themes

7.2.1 Multi-scale selection

An important question driving much of the analyses presented here is how selection pressures operating within individual hosts—including viral replication fitness and immune escape—influence the accumulation of mutations at the population level and shape viral evolution across decades of the pandemic. I first focussed on the deleterious component of the fitness landscape and how weak selection can be balanced by the high influx of mutations, constraining short-term evolution and providing the opportunity for between-host evolution to select for fitness.

Analysis of viral populations within transmission pairs revealed HIV's capacity to explore extensive mutational space, evidenced by the sequential selection and replacement of CTL escape variants as new epitope mutations emerge. Immune-escape mutations that are only transiently beneficial due to adaptations within the host response have also been reported in the context of antibody escape (Hedskog *et al.*, 2010). I found CTL escape to be largely host-specific, with *de novo* escapes in recipients showing little overlap with population-level variants. This pattern reveals both the virus's remarkable ability to evolve novel, functional escape routes and the tendency for most escapes to be purged at the between-host level, suggesting evolution toward a "universally fit" viral genome. While I do not explicitly look for compensatory mutations, previous studies have demonstrated that the virus by-passes fitness costs by selecting for mutations outside of epitopes that make up for fitness costs (Brockman *et al.*, 2007; Crawford, Prado, Leslie, Hué, Honeyborne, Reddy, Van Der Stok, *et al.*, 2007; Kløverpris, Leslie and Goulder, 2016), which may be why most transmitted escapes persist during the first year of infection and again is evidence of HIV’s ability to adapt and continue to spread despite enormous heterogeneity in host immunological responses.

An important finding of this study is that despite expected differences in selection pressures across scales and markedly different rates of evolution, we find evidence that there are general rules and constraints by which the virus evolves. While CTL escape mutations were often unique, we found the same codons positions were

repeatedly under selection, as has been reported for both antibody and CTL escape sites in a smaller cohort of individuals (Illingworth *et al.*, 2020). CTL escape has been shown to a certain extent be predictable during early infection, with the ordered pattern of escape mutations particularly well documented in infection individuals with protective HLA alleles B57 and B27 (Leslie *et al.*, 2004; Crawford, Prado, Leslie, Hué, Honeyborne, Reddy, Van Der Stok, *et al.*, 2007; Carlson *et al.*, 2008).

Hot spots and cold spots for recombination across the genome were also remarkably consistent not only across subtypes, but also at both the within and between host scales. Recombination patterns are therefore likely to be intrinsic to the genome structure and governed by underlying mechanisms. Consistency has been demonstrated across datasets at an inter-subtype level (Fan, Negroni and Robertson, 2007; Jia *et al.*, 2016; Tongo, De Oliveira and Martin, 2018; Grant *et al.*, 2020), yet here I show these correlations exist in infections of a single subtype with an entirely distinct methodological approach. Similarly, when measuring the rate of divergence across the genome, I found consistency across scales, further supporting the existence of fundamental evolutionary constraints. This finding aligns with the documented tendency of HIV to evolve toward a subtype-specific consensus sequence, representing an optimally fit virus in the absence of host-specific immune pressures.

7.2.2 The evolutionary rate mismatch: within and between host toggling

Several hypotheses have been proposed that explain the observed mismatch in the evolutionary rates across the within and between-host scales (Lythgoe and Fraser, 2012; Alizon and Fraser, 2013). A leading theory is the "escape and revert" hypothesis, proposing that host-specific adaptations, particularly CTL escape mutations, carry fitness costs that lead to their reversion upon transmission to hosts with different HLA types (Leslie *et al.*, 2004; Herbeck *et al.*, 2006; Boutwell *et al.*, 2010; Zanini *et al.*, 2015; Raghwani *et al.*, 2018). This cyclical pattern of adaptation and reversion means that many mutations that contribute to within-host evolution do not persist at the population level, resulting in slower apparent evolutionary rates between hosts. In chapter 6, I presented examples of this process across single transmission events, and highlighted how the uniqueness of the escape mutations I identified at a population level suggests escapes typically carry a fitness cost. The relatively low number of transmitted escapes selected against suggests this process is slow.

Within-host toggling emerged as a prominent feature of early infection dynamics in Chapter 5, and I argue this mechanism significantly contributes to elevated evolutionary rates due to the repeated detection of mutations that rise to high frequency but fail to fix in the population. These transient mutations are captured by BEAST analyses as evolutionary events, leading to longer branch lengths and higher rate estimates, while divergence-based methods measuring only net genetic change do not detect them. This explains both the disparity between BEAST and divergence rate estimates at the within-host scale and contributes to the evolutionary rate mismatch between within-host and between-host scales.

The high recombination rates documented in Chapter 3 facilitate the breakdown of linkage between beneficial mutations and hitchhiking synonymous variants, promoting within-host toggling. Following a recombination or mutation event on a hitchhiking synonymous mutation, the mutation is selected against and declines in frequency. Synonymous mutations have been shown to be weakly deleterious in C2-V5 (Zanini and Neher, 2013), and the toggling of synonymous mutations in this genome-wide study suggests this likely holds across the whole genome. Toggling as the dominant mechanism driving the evolutionary rate mismatch is also likely a more parsimonious explanation of the rate mismatch for synonymous evolution than the “store and retrieve” mechanism, while “escape and revert” specifically explains non-synonymous evolution.

Notably, we observed toggling of non-synonymous mutations, a phenomenon that can be driven by multiple mechanisms. It has been previously suggested that both epitope competition and time-dependent selection are potential drivers (Zanini and Neher, 2013), and my analysis of transmission pairs provided direct evidence for competition between epitope variants as a mechanism for non-synonymous toggling. The impact of toggling mutations on evolutionary rate disparities across scales may extend beyond HIV to other chronic viral infections where similar evolutionary rate mismatches have occurred.

A key takeaway from the analyses is that methodological choice and study design are fundamental considerations when researching the evolutionary dynamics of rapidly evolving viruses. In measuring recombination, I showed with both simulated and empirical data that the inferred rate of recombination is influenced by both the frequency of sampling and read lengths. To improve accuracy, it may be necessary to

account for study design before putting resources into generating longer reads, with longer reads more problematic when sampling is less dense, and linkage is more likely to have broken down across sampling time points. Similarly, when measuring evolutionary rates, more frequent sampling captures a greater proportion of transient dynamics, thus elevating rates, with this effect significantly more impactful in time-tree analyses including BEAST. When choosing a methodological approach, it is important to consider the research question at hand, and caution should be taken when comparing results across studies with differences in sampling strategy.

7.3 Strengths and Limitations

A major strength of this thesis lies in its comprehensive dataset, encompassing hundreds of individuals across both early and late stages of infection, multiple viral subtypes, and diverse HLA backgrounds. This breadth enabled robust analysis of within-host viral evolution, including CTL-driven selection, recombination, and fitness costs. Notably, inclusion of HLA data for both transmission partners allowed direct mapping of within-host mutations to known CTL escape motifs and facilitated assessment of fitness effects following shifts in host immune pressure. Longitudinal sampling during early infection captured key evolutionary dynamics, while sampling of chronically infected source partners provided the genetic diversity necessary for reliable recombination analyses. Furthermore, this genome-wide approach contrasts with prior studies focused on individual viral proteins, enabling broader insight into HIV adaptation.

Despite the statistical power of the large sample size, several methodological limitations merit careful consideration. First, the scale of the dataset complicated detection of subtle evolutionary patterns such as compensatory mutations following CTL escape. These may be better resolved in future studies using machine learning approaches (Haddox et al., 2018; Blassel et al., 2021). Additionally, the absence of experimental validation means that some reported escape mutations could reflect hitchhiking variants or mutations conferring selective advantages unrelated to CTL pressure, such as antibody escape.

Sampling biases also constrain interpretation. The three- to six-month intervals between samples likely underestimate selection coefficients at rapidly evolving sites, as meaningful allele frequency shifts may occur between timepoints. These intervals

likely also bias recombination rate estimates. Furthermore, reliance on peripheral blood sampling excludes viral populations in other anatomical compartments, where divergent selective pressures may drive distinct evolutionary trajectories.

Sequencing-related limitations may have introduced additional bias into the analyses. While PacBio long-read sequencing enhances haplotype resolution and reduces alignment artifacts common to short-read platforms, it is not entirely free from distortion. Pre-sequencing PCR amplification can introduce template bias, potentially skewing the representation of viral variants, and the lower sequencing depth relative to short-read methods can limit the detection of rare variants (Ross et al 2013). This may lead to underrepresentation of low-frequency CTL escape mutations or early-stage toggling dynamics, thereby affecting estimates of within-host diversity, selection pressures, and apparent convergent evolution. Further investigation into potential sequencing errors on the PacBio platform—such as by resequencing samples—could help assess these limitations more precisely, though this lies beyond the scope and available resources for this thesis.

Analytical constraints arose from the need to aggregate data across the cohort, which limited the integration of uncertainty in evolutionary rates inferred via BEAST into cross-scale comparisons. Model selection was only performed on a subset of the data, restricting exploration of alternative evolutionary frameworks. Different assumptions—such as substitution models (e.g., JC69, K80, GTR), molecular clock models (strict vs. relaxed), and coalescent priors—could yield divergent phylogenetic outcomes. A more systematic comparison would better establish robustness but was hindered by the substantial computational burden of the full dataset.

A more fundamental limitation stems from HIV's exceptionally high recombination rate, which rivals its mutation rate and challenges the core assumption of a single, tree-like evolutionary history underlying standard phylogenetic tools. While restricting analyses to 500 bp windows mitigated this to some extent, recombination likely continues to bias evolutionary rate estimates and may generate mosaic genomes that artificially inflate signals of convergent evolution.

Given additional resources, I would prioritise implementation of recombination-aware evolutionary models, complemented by simulation-based approaches where the true evolutionary history is known. These could help distinguish biological signal from

methodological artifact and noise, and clarify how mutation and recombination jointly shape HIV's evolutionary landscape. Simulations could also test the impact of toggling dynamics on phylogenetic inferences, including tree topology and estimates of time to the most recent common ancestor. Finally, given more time, I would integrate findings from CTL escape dynamics and associated fitness costs into the set-point viral load problem, assessing how immune escape influences viral load and embedding these dynamics into mathematical models for a more comprehensive understanding of host–virus interactions.

7.4 Future Outlook

The findings of this thesis open several promising research directions. Detailed characterisation of evolutionary dynamics—including weak purifying selection, immune escape selection patterns, and genome-wide recombination rates—provides a foundation for developing biologically informed simulations. Such simulations could illuminate how within-host selection shapes long-term viral evolution and improve our understanding of how these processes affect evolutionary inference across scales, particularly in reconstructing outbreak timing and transmission networks.

In understanding the key drivers of within-host evolution, and how these impact between-host evolution is key in the continued search of an effective cure or vaccine. In particular, understanding the general rules that govern the effective recombination rate and evolutionary change, we are better prepared in harnessing the weak points in the virus's ability to adapt.

7.5 Concluding remarks

The extraordinary evolutionary capacity of HIV, driven by its rapid mutation rate and ability to adapt, remains a fundamental challenge in global public health. This thesis examines the interplay between within-host adaptation and population-level viral diversification, revealing conserved features that govern HIV evolution - from molecular constraints on recombination to consistent signatures of selection. We demonstrate that methodological choices significantly impact evolutionary inference, highlighting the importance of careful interpretation when contextualising findings within existing literature.

8 Bibliography

Abecasis, A.B., Vandamme, A.-M. and Lemey, P. (2009) 'Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution', *Journal of Virology*, 83(24), pp. 12917–12924. Available at: https://doi.org/10.1128/JVI.01022-09.

Aiewsakun, P. and Katzourakis, A. (2016) 'Time-Dependent Rate Phenomenon in Viruses', *Journal of Virology*. Edited by S.R. Ross, 90(16), pp. 7184–7195. Available at: https://doi.org/10.1128/JVI.00593-16.

Aiken, C. and Rousso, I. (2021) 'The HIV-1 capsid and reverse transcription', *Retrovirology*, 18(1), p. 29. Available at: https://doi.org/10.1186/s12977-021-00566-0.

Alizon, S. *et al.* (2009) 'Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future', *Journal of Evolutionary Biology*, 22(2), pp. 245–259. Available at: https://doi.org/10.1111/j.1420-9101.2008.01658.x.

Alizon, S. *et al.* (2010) 'Phylogenetic Approach Reveals That Virus Genotype Largely Determines HIV Set-Point Viral Load', *PLoS Pathogens*. Edited by C.O. Wilke, 6(9), p. e1001123. Available at: https://doi.org/10.1371/journal.ppat.1001123.

Alizon, S. and Fraser, C. (2013) 'Within-host and between-host evolutionary rates across the HIV-1 genome', *Retrovirology*, 10(1), p. 49. Available at: https://doi.org/10.1186/1742-4690-10-49.

Allen, T.M. *et al.* (2004) 'Selection, Transmission, and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection', *Journal of Virology*, 78(13), pp. 7069–7078. Available at: https://doi.org/10.1128/JVI.78.13.7069-7078.2004.

Allen, T.M. *et al.* (2005) 'Selective Escape from $CD8^+$ T-Cell Responses Represents a Major Driving Force of Human Immunodeficiency Virus Type 1 (HIV-1) Sequence Diversity and Reveals Constraints on HIV-1 Evolution', *Journal of Virology*, 79(21), pp. 13239–13249. Available at: https://doi.org/10.1128/JVI.79.21.13239-13249.2005.

Altfeld, M. *et al.* (2006) 'HLA Alleles Associated with Delayed Progression to AIDS Contribute Strongly to the Initial CD8(+) T Cell Response against HIV-1', *PLoS medicine*, 3(10), p. e403. Available at: https://doi.org/10.1371/journal.pmed.0030403.

An. W., and Telesnitsky, A. (2002). 'E↵ects of varying sequence similarity on the frequency of repeat deletion during reverse transcription of a Human Immunodeficiency Virus Type 1 vector', *J Virol*, 76(15):7897–7902 DOI: 10.1128/jvi.76.15.7897-7902.2002

Anderson, R.M. and May, R.M. (1982) 'Coevolution of hosts and parasites', *Parasitology*, 85(2), pp. 411–426. Available at: https://doi.org/10.1017/S0031182000055360.

Archer, J. *et al.* (2008) ‘Identifying the Important HIV-1 Recombination Breakpoints’, *PLoS Computational Biology*. Edited by S. Bonhoeffer, 4(9), p. e1000178. Available at: https://doi.org/10.1371/journal.pcbi.1000178.

Ash, M.K., Al-Harthi, L. and Schneider, J.R. (2021) ‘HIV in the Brain: Identifying Viral Reservoirs and Addressing the Challenges of an HIV Cure’, *Vaccines*, 9(8), p. 867. Available at: https://doi.org/10.3390/vaccines9080867.

Avila-Rios, S. *et al.* (2019) ‘Clinical and evolutionary consequences of HIV adaptation to HLA: implications for vaccine and cure’, *Current Opinion in HIV and AIDS*, 14(3), pp. 194–204. Available at: https://doi.org/10.1097/COH.0000000000000541.

Baeten, J.M. *et al.* (2012) ‘Antiretroviral Prophylaxis for HIV Prevention in Heterosexual Men and Women’, *New England Journal of Medicine*, 367(5), pp. 399–410. Available at: https://doi.org/10.1056/NEJMoa1108524.

Baele, G, Li W. L., S, Drummond, A.J., et al. (2013), ’Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics’, *Molecular Biology and Evolution*, 30(2), pp 239–243, https://doi.org/10.1093/molbev/mss243

Bagaya, B.S. *et al.* (2015) ‘Functional bottlenecks for generation of HIV-1 intersubtype Env recombinants’, *Retrovirology*, 12(1), p. 44. Available at: https://doi.org/10.1186/s12977-015-0170-8.

Baird, H.A. *et al.* (2006) ‘Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination’, *Retrovirology*, 3(1), p. 91. Available at: https://doi.org/10.1186/1742-4690-3-91.

Balakrishnan, M. *et al.* (2003) ‘Template Dimerization Promotes an Acceptor Invasion-Induced Transfer Mechanism during Human Immunodeficiency Virus Type 1 Minus-Strand Synthesis’, *Journal of Virology*, 77(8), pp. 4710–4721. Available at: https://doi.org/10.1128/JVI.77.8.4710-4721.2003.

Balakrishnan, M., Fay, P.J. and Bambara, R.A. (2001) ‘The Kissing Hairpin Sequence Promotes Recombination within the HIV-I 5′ Leader Region’, *Journal of Biological Chemistry*, 276(39), pp. 36482–36492. Available at: https://doi.org/10.1074/jbc.M102860200.

Bar, K.J. *et al.* (2010) ‘Wide Variation in the Multiplicity of HIV-1 Infection among Injection Drug Users’, *Journal of Virology*, 84(12), pp. 6241–6247. Available at: https://doi.org/10.1128/JVI.00077-10.

Barré-Sinoussi, F. *et al.* (1983) ‘Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS)’, *Science*, 220(4599), pp. 868–871. Available at: https://doi.org/10.1126/science.6189183.

Bartha, I. *et al.* (2013) ‘A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control’, *eLife*, 2, p. e01123. Available at: https://doi.org/10.7554/eLife.01123.

Bartha, I. *et al.* (2017) ‘Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control’, *PLOS Computational Biology*. Edited by V. Müller, 13(2), p. e1005339. Available at: https://doi.org/10.1371/journal.pcbi.1005339.

Batorsky, R. *et al.* (2011) ‘Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection’, *Proceedings of the National Academy of Sciences*, 108(14), pp. 5661–5666. Available at: https://doi.org/10.1073/pnas.1102036108.

Bbosa, N., Kaleebu, P. and Ssemwanga, D. (2019) ‘HIV subtype diversity worldwide’, *Current Opinion in HIV and AIDS*, 14(3), pp. 153–160. Available at: https://doi.org/10.1097/COH.0000000000000534.

Bernardin, F. *et al.* (2005) ‘Human Immunodeficiency Virus Mutations during the First Month of Infection Are Preferentially Found in Known Cytotoxic T-Lymphocyte Epitopes’, *Journal of Virology*, 79(17), pp. 11523–11528. Available at: https://doi.org/10.1128/JVI.79.17.11523-11528.2005.

Bertels, F. *et al.* (2018) ‘Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T-Cell Decline, and Per-Parasite Pathogenicity’, *Molecular Biology and Evolution*, 35(1), pp. 27–37. Available at: https://doi.org/10.1093/molbev/msx246.

Blackard, J.T., Cohen, D.E. and Mayer, K.H. (2002) ‘Human Immunodeficiency Virus Superinfection and Recombination: Current State of Knowledge and Potential Clinical Consequences’, *Clinical Infectious Diseases*, 34(8), pp. 1108–1114. Available at: https://doi.org/10.1086/339547.

Blanquart, F. *et al.* (2016) ‘A transmission-virulence evolutionary trade-off explains attenuation of HIV-1 in Uganda’, *eLife*, 5, p. e20492. Available at: https://doi.org/10.7554/eLife.20492.

Blanquart, F. *et al.* (2017) ‘Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe’, *PLOS Biology*. Edited by R. Sanjuan, 15(6), p. e2001855. Available at: https://doi.org/10.1371/journal.pbio.2001855.

Blassel, L. *et al.* (2021) ‘Using machine learning and big data to explore the drug resistance landscape in HIV’, *PLOS Computational Biology*. Edited by A.L. Hill, 17(8), p. e1008873. Available at: https://doi.org/10.1371/journal.pcbi.1008873.

Bleul, C.C. *et al.* (1996) ‘The lymphocyte chemoattractant SDF-1 is a ligand for LESTR/fusin and blocks HIV-1 entry’, *Nature*, 382(6594), pp. 829–833. Available at: https://doi.org/10.1038/382829a0.

Bonhoeffer, S., Fraser, C. and Leventhal, G.E. (2015) ‘High Heritability Is Compatible with the Broad Distribution of Set Point Viral Load in HIV Carriers’, *PLOS Pathogens*. Edited by R. Swanstrom, 11(2), p. e1004634. Available at: https://doi.org/10.1371/journal.ppat.1004634.

Bonsall, D. *et al.* (2020) ‘A Comprehensive Genomics Solution for HIV Surveillance and Clinical Monitoring in Low-Income Settings’, *Journal of Clinical Microbiology*. Edited by A.M. Caliendo, 58(10), pp. e00382-20. Available at: https://doi.org/10.1128/JCM.00382-20.

Bonsignori, M. *et al.* (2017) 'Antibody-virus co-evolution in HIV infection: paths for HIV vaccine development', *Immunological Reviews*, 275(1), pp. 145–160. Available at: https://doi.org/10.1111/imr.12509.

Borrell, M. *et al.* (2021) 'High rates of long-term progression in HIV-1-positive elite controllers', *Journal of the International AIDS Society*, 24(2), p. e25675. Available at: https://doi.org/10.1002/jia2.25675.

Boutwell, C.L. *et al.* (2010) 'Viral Evolution and Escape during Acute HIV-1 Infection', *The Journal of Infectious Diseases*, 202(S2), pp. S309–S314. Available at: https://doi.org/10.1086/655653.

Brockman, M.A. *et al.* (2007) 'Escape and compensation from early HLA-B57-mediated cytotoxic T-lymphocyte pressure on human immunodeficiency virus type 1 Gag alter capsid interactions with cyclophilin A', *Journal of Virology*, 81(22), pp. 12608–12618. Available at: https://doi.org/10.1128/JVI.01369-07.

Brown, A.J.L. (1997) 'Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population', *Proceedings of the National Academy of Sciences*, 94(5), pp. 1862–1865. Available at: https://doi.org/10.1073/pnas.94.5.1862.

Brumme, Z.L. *et al.* (2008) 'Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection', *Journal of Virology*, 82(18), pp. 9216–9227. Available at: https://doi.org/10.1128/JVI.01041-08.

Brumme, Z.L. *et al.* (2009) 'HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins', *PLoS ONE*. Edited by D.F. Nixon, 4(8), p. e6687. Available at: https://doi.org/10.1371/journal.pone.0006687.

Burton, D.R. and Hangartner, L. (2016) 'Broadly Neutralizing Antibodies to HIV and Their Role in Vaccine Design', *Annual Review of Immunology*, 34(1), pp. 635–659. Available at: https://doi.org/10.1146/annurev-immunol-041015-055515.

Buzon, M.J. *et al.* (2014) 'HIV-1 persistence in CD4+ T cells with stem cell–like properties', *Nature Medicine*, 20(2), pp. 139–142. Available at: https://doi.org/10.1038/nm.3445.

Caillat, C. *et al.* (2021) 'Structure of HIV-1 gp41 with its membrane anchors targeted by neutralizing antibodies', *eLife*, 10, p. e65005. Available at: https://doi.org/10.7554/eLife.65005.

Carlson, J.M. *et al.* (2008) 'Phylogenetic Dependency Networks: Inferring Patterns of CTL Escape and Codon Covariation in HIV-1 Gag', *PLoS Computational Biology*. Edited by R.J. De Boer, 4(11), p. e1000225. Available at: https://doi.org/10.1371/journal.pcbi.1000225.

Carlson, J.M. *et al.* (2012) 'Widespread Impact of HLA Restriction on Immune Control and Escape Pathways of HIV-1', *Journal of Virology*, 86(9), pp. 5230–5243. Available at: https://doi.org/10.1128/JVI.06728-11.

Carlson, J.M. *et al.* (2014) ‘Selection bias at the heterosexual HIV-1 transmission bottleneck’, *Science*, 345(6193), p. 1254031. Available at: https://doi.org/10.1126/science.1254031.

Carlson, J.M. *et al.* (2015) ‘HIV-1 adaptation to HLA: a window into virus–host immune interactions’, *Trends in Microbiology*, 23(4), pp. 212–224. Available at: https://doi.org/10.1016/j.tim.2014.12.008.

Carlson, J.M. *et al.* (2016) ‘Impact of pre-adapted HIV transmission’, *Nature Medicine*, 22(6), pp. 606–613. Available at: https://doi.org/10.1038/nm.4100.

Carlson, J.M. and Brumme, Z.L. (2008) ‘HIV evolution in response to HLA-restricted CTL selection pressures: a population-based perspective’, *Microbes and Infection*, 10(5), pp. 455–461. Available at: https://doi.org/10.1016/j.micinf.2008.01.013.

Carter, A. *et al.* (2024) ‘Global, regional, and national burden of HIV/AIDS, 1990–2021, and forecasts to 2050, for 204 countries and territories: the Global Burden of Disease Study 2021’, *The Lancet HIV*, 11(12), pp. e807–e822. Available at: https://doi.org/10.1016/S2352-3018(24)00212-1.

Caskey, M. (2020) ‘Broadly neutralizing antibodies for the treatment and prevention of HIV infection’, *Current opinion in HIV and AIDS*, 15(1), pp. 49–55. Available at: https://doi.org/10.1097/COH.0000000000000600.

Celum, C. *et al.* (2010) ‘Acyclovir and Transmission of HIV-1 from Persons Infected with HIV-1 and HSV-2’, *New England Journal of Medicine*, 362(5), pp. 427–439. Available at: https://doi.org/10.1056/NEJMoa0904849.

Chan, D.C. and Kim, P.S. (1998) ‘HIV Entry and Its Inhibition’, *Cell*, 93(5), pp. 681–684. Available at: https://doi.org/10.1016/S0092-8674(00)81430-0.

Checkley, M.A., Luttge, B.G. and Freed, E.O. (2011) ‘HIV-1 Envelope Glycoprotein Biosynthesis, Trafficking, and Incorporation’, *Journal of Molecular Biology*, 410(4), pp. 582–608. Available at: https://doi.org/10.1016/j.jmb.2011.04.042.

Chen, J. *et al.* (2009) ‘High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis’, *Proceedings of the National Academy of Sciences*, 106(32), pp. 13535–13540. Available at: https://doi.org/10.1073/pnas.0906822106.

Chen, J. *et al.* (2022) ‘The reservoir of latent HIV’, *Frontiers in Cellular and Infection Microbiology*, 12, p. 945956. Available at: https://doi.org/10.3389/fcimb.2022.945956.

Chen, J., Powell, D. and Hu, W.-S. (2006) ‘High Frequency of Genetic Recombination Is a Common Feature of Primate Lentivirus Replication’, *Journal of Virology*, 80(19), pp. 9651–9658. Available at: https://doi.org/10.1128/JVI.00936-06.

Chen, J., Rhodes, T.D. and Hu, W.-S. (2005) ‘Comparison of the Genetic Recombination Rates of Human Immunodeficiency Virus Type 1 in Macrophages and T Cells’, *Journal of Virology*, 79(14), pp. 9337–9340. Available at: https://doi.org/10.1128/JVI.79.14.9337-9340.2005.

Childs, L.M. *et al.* (2019) ‘Linked within-host and between-host models and data for infectious diseases: a systematic review’, *PeerJ*, 7, p. e7057. Available at: https://doi.org/10.7717/peerj.7057.

Chohan, B. *et al.* (2005) ‘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’, *Journal of Virology*, 79(10), pp. 6528–6531. Available at: https://doi.org/10.1128/JVI.79.10.6528-6531.2005.

Chomont, N. *et al.* (2009) ‘HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation’, *Nature Medicine*, 15(8), pp. 893–900. Available at: https://doi.org/10.1038/nm.1972.

Chopera, D.R. *et al.* (2008) ‘Transmission of HIV-1 CTL Escape Variants Provides HLA-Mismatched Recipients with a Survival Advantage’, *PLoS Pathogens*. Edited by R.A. Koup, 4(3), p. e1000033. Available at: https://doi.org/10.1371/journal.ppat.1000033.

Chun, T.W. *et al.* (1997) ‘Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection’, *Nature*, 387(6629), pp. 183–188. Available at: https://doi.org/10.1038/387183a0.

Clark, S.J. *et al.* (1991) ‘High Titers of Cytopathic Virus in Plasma of Patients with Symptomatic Primary HIV-1 Infection’, *New England Journal of Medicine*, 324(14), pp. 954–960. Available at: https://doi.org/10.1056/NEJM199104043241404.

Coffin, J.M. (1995) ‘HIV Population Dynamics in Vivo: Implications for Genetic Variation, Pathogenesis, and Therapy’, *Science*, 267(5197), pp. 483–489. Available at: https://doi.org/10.1126/science.7824947.

Connor, R.I. *et al.* (1997) ‘Change in coreceptor use correlates with disease progression in HIV-1--infected individuals’, *The Journal of Experimental Medicine*, 185(4), pp. 621–628. Available at: https://doi.org/10.1084/jem.185.4.621.

Crawford, H., Prado, J.G., Leslie, A., Hué, S., Honeyborne, I., Reddy, S., Van Der Stok, M., *et al.* (2007) ‘Compensatory Mutation Partially Restores Fitness and Delays Reversion of Escape Mutation within the Immunodominant HLA-B*5703-Restricted Gag Epitope in Chronic Human Immunodeficiency Virus Type 1 Infection’, *Journal of Virology*, 81(15), pp. 8346–8351. Available at: https://doi.org/10.1128/JVI.00465-07.

Crawford, H., Prado, J.G., Leslie, A., Hué, S., Honeyborne, I., Reddy, S., van der Stok, M., *et al.* (2007) ‘Compensatory mutation partially restores fitness and delays reversion of escape mutation within the immunodominant HLA-B*5703-restricted Gag epitope in chronic human immunodeficiency virus type 1 infection’, *Journal of Virology*, 81(15), pp. 8346–8351. Available at: https://doi.org/10.1128/JVI.00465-07.

Crawford, H. *et al.* (2009) ‘Evolution of HLA-B*5703 HIV-1 escape mutations in HLA-B*5703–positive individuals and their transmission recipients’, *Journal of Experimental Medicine*, 206(4), pp. 909–921. Available at: https://doi.org/10.1084/jem.20081984.

Cressler, C.E. *et al.* (2016) ‘The adaptive evolution of virulence: a review of theoretical predictions and empirical tests’, *Parasitology*, 143(7), pp. 915–930. Available at: https://doi.org/10.1017/S003118201500092X.

Dahl, V., Josefsson, L. and Palmer, S. (2010) ‘HIV reservoirs, latency, and reactivation: Prospects for eradication’, *Antiviral Research*, 85(1), pp. 286–294. Available at: https://doi.org/10.1016/j.antiviral.2009.09.016.

Davenport, M.P. *et al.* (2008) ‘Rates of HIV immune escape and reversion: implications for vaccination’, *Trends in Microbiology*, 16(12), pp. 561–566. Available at: https://doi.org/10.1016/j.tim.2008.09.001.

Day, C.L. *et al.* (2006) ‘PD-1 expression on HIV-specific T cells is associated with T-cell exhaustion and disease progression’, *Nature*, 443(7109), pp. 350–354. Available at: https://doi.org/10.1038/nature05115.

De Wolf, F. *et al.* (1997) ‘AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion’, *AIDS*, 11(15), pp. 1799–1806. Available at: https://doi.org/10.1097/00002030-199715000-00003.

Deeks, S.G. *et al.* (2015) ‘HIV infection’, *Nature Reviews Disease Primers*, 1(1), p. 15035. Available at: https://doi.org/10.1038/nrdp.2015.35.

Deeks, S.G. and Walker, B.D. (2004) ‘The immune response to AIDS virus infection: good, bad, or both?’, *The Journal of Clinical Investigation*, 113(6), pp. 808–810. Available at: https://doi.org/10.1172/JCI21318.

Didelot X., Wilson D.J. (2015), ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Computational Biology* 11(2): e1004041.https://doi.org/10.1371/journal.pcbi.100

Derdeyn, C.A. *et al.* (2004) ‘Envelope-Constrained Neutralization-Sensitive HIV-1 After Heterosexual Transmission’, *Science*, 303(5666), pp. 2019–2022. Available at: https://doi.org/10.1126/science.1093137.

Diekmann, O. and Heesterbeek, J.A.P. (2000) ‘Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation’, *Wiley Series in Mathematical and Computational Biology, Chichester, Wiley* [Preprint].

Doekes, H.M., Fraser, C. and Lythgoe, K.A. (2017) ‘Effect of the Latent Reservoir on the Evolution of HIV at the Within- and Between-Host Levels’, *PLOS Computational Biology*. Edited by R.R. Regoes, 13(1), p. e1005228. Available at: https://doi.org/10.1371/journal.pcbi.1005228.

Domingo, E., Sheldon, J. and Perales, C. (2012) ‘Viral Quasispecies Evolution’, *Microbiology and Molecular Biology Reviews*, 76(2), pp. 159–216. Available at: https://doi.org/10.1128/MMBR.05023-11.

Doria-Rose, N.A. *et al.* (2009) ‘Frequency and Phenotype of Human Immunodeficiency Virus Envelope-Specific B Cells from Patients with Broadly Cross-

Neutralizing Antibodies', *Journal of Virology*, 83(1), pp. 188–199. Available at: https://doi.org/10.1128/JVI.01583-08.

Dragic, T. *et al.* (1996) 'HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5', *Nature*, 381(6584), pp. 667–673. Available at: https://doi.org/10.1038/381667a0.

Druelle, V. and Neher, R.A. (2023) 'Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates', *Virus Evolution*, 9(1), p. veac118. Available at: https://doi.org/10.1093/ve/veac118.

Drummond, A.J. *et al.* (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*. Edited by D. Penny, 4(5), p. e88. Available at: https://doi.org/10.1371/journal.pbio.0040088.

Duchêne, S., Holmes, E.C. and Ho, S.Y.W. (2014) 'Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates', *Proceedings. Biological Sciences*, 281(1786), p. 20140732. Available at: https://doi.org/10.1098/rspb.2014.0732.

Duffy, S., Shackelton, L.A. and Holmes, E.C. (2008) 'Rates of evolutionary change in viruses: patterns and determinants', *Nature Reviews Genetics*, 9(4), pp. 267–276. Available at: https://doi.org/10.1038/nrg2323.

Eigen, M., Schuster, P., 1979, 'The Hypercycle: A Principle of Natural Self-Organization', Berlin: Springer-Verlag. ISBN 978-0-387-09293-5.

Emerman, M. and Malim, M.H. (1998) 'HIV-1 Regulatory/Accessory Genes: Keys to Unraveling Viral and Host Cell Biology', *Science*, 280(5371), pp. 1880–1884. Available at: https://doi.org/10.1126/science.280.5371.1880.

Ewald, P.W. (1983) 'Host-Parasite Relations, Vectors, and the Evolution of Disease Severity', *Annual Review of Ecology and Systematics*, 14(1), pp. 465–485. Available at: https://doi.org/10.1146/annurev.es.14.110183.002341.

Fan, J., Negroni, M. and Robertson, D.L. (2007) 'The distribution of HIV-1 recombination breakpoints', *Infection, Genetics and Evolution*, 7(6), pp. 717–723. Available at: https://doi.org/10.1016/j.meegid.2007.07.012.

Faria, N.R. *et al.* (2014) 'The early spread and epidemic ignition of HIV-1 in human populations', *Science*, 346(6205), pp. 56–61. Available at: https://doi.org/10.1126/science.1256739.

Fideli, Ü.S. *et al.* (2001) 'Virologic and Immunologic Determinants of Heterosexual Transmission of Human Immunodeficiency Virus Type 1 in Africa', *AIDS Research and Human Retroviruses*, 17(10), pp. 901–910. Available at: https://doi.org/10.1089/088922201750290023.

Finzi, D. *et al.* (1997) 'Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy', *Science*, 278(5341), pp. 1295–1300. Available at: https://doi.org/10.1126/science.278.5341.1295.

Fraser, C. *et al.* (2007) 'Variation in HIV-1 set-point viral load: Epidemiological analysis and an evolutionary hypothesis', *Proceedings of the National Academy of Sciences*, 104(44), pp. 17441–17446. Available at: https://doi.org/10.1073/pnas.0708559104.

Fraser, C. *et al.* (2014) 'Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective', *Science*, 343(6177), p. 1243727. Available at: https://doi.org/10.1126/science.1243727.

Friedrich, T.C. *et al.* (2004) 'Reversion of CTL escape–variant immunodeficiency viruses in vivo', *Nature Medicine*, 10(3), pp. 275–281. Available at: https://doi.org/10.1038/nm998.

Fryer, H.R. *et al.* (2010) 'Modelling the Evolution and Spread of HIV Immune Escape Mutants', *PLoS Pathogens*. Edited by C.O. Wilke, 6(11), p. e1001196. Available at: https://doi.org/10.1371/journal.ppat.1001196.

Gabrielaite, M. *et al.* (2021) 'Human Immunotypes Impose Selection on Viral Genotypes Through Viral Epitope Specificity', *The Journal of Infectious Diseases*, 224(12), pp. 2053–2063. Available at: https://doi.org/10.1093/infdis/jiab253.

Galetto, R. *et al.* (2004) 'The Structure of HIV-1 Genomic RNA in the gp120 Gene Determines a Recombination Hot Spot in Vivo', *Journal of Biological Chemistry*, 279(35), pp. 36625–36632. Available at: https://doi.org/10.1074/jbc.M405476200.

Galli, A., Kearney, M., Nikolaitchik, O. A., et al, (2010), 'Patterns of Human Immunodeficiency Virus Type 1 Recombination Ex Vivo Provide Evidence for Coadaptation of Distant Sites, Resulting in Purifying Selection for Intersubtype Recombinants during Replication', *J. Virol*, 84(15):7651-61. doi: 10.1128/JVI.00276-10.

Ghafari, M. *et al.* (2021) 'Prisoner of War dynamics explains the time-dependent pattern of substitution rates in viruses'. Available at: https://doi.org/10.1101/2021.02.09.430479.

Gill, M.S,, Lemey, P., Faria. N., R. et al, 2013, 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30(3), pp: 713–724, https://doi.org/10.1093/molbev/mss265

Golden, M. *et al.* (2014) 'Patterns of Recombination in HIV-1M Are Influenced by Selection Disfavouring the Survival of Recombinants with Disrupted Genomic RNA and Protein Structures', *PLoS ONE*. Edited by N. Singh, 9(6), p. e100400. Available at: https://doi.org/10.1371/journal.pone.0100400.

Goonetilleke, N. *et al.* (2009) 'The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection', *Journal of Experimental Medicine*, 206(6), pp. 1253–1272. Available at: https://doi.org/10.1084/jem.20090365.

Goulder, P.J.R. *et al.* (2001) 'Evolution and transmission of stable CTL escape mutations in HIV infection', *Nature*, 412(6844), pp. 334–338. Available at: https://doi.org/10.1038/35085576.

Goulder, P.J.R. and Walker, B.D. (2012) ‘HIV and HLA Class I: An Evolving Relationship’, *Immunity*, 37(3), pp. 426–440. Available at: https://doi.org/10.1016/j.immuni.2012.09.005.

Goulder, P.J.R. and Watkins, D.I. (2004) ‘HIV and SIV CTL escape: implications for vaccine design’, *Nature Reviews. Immunology*, 4(8), pp. 630–640. Available at: https://doi.org/10.1038/nri1417.

Grant, H.E. *et al.* (2020) ‘Pervasive and non-random recombination in near full-length HIV genomes from Uganda’, *Virus Evolution*, 6(1), p. veaa004. Available at: https://doi.org/10.1093/ve/veaa004.

Grant, R.M. *et al.* (2010) ‘Preexposure Chemoprophylaxis for HIV Prevention in Men Who Have Sex with Men’, *New England Journal of Medicine*, 363(27), pp. 2587–2599. Available at: https://doi.org/10.1056/NEJMoa1011205.

Gray, R.R. *et al.* (2011) ‘The mode and tempo of hepatitis C virus evolution within and among hosts’, *BMC evolutionary biology*, 11, p. 131. Available at: https://doi.org/10.1186/1471-2148-11-131.

Gulick, R.M. *et al.* (1997) ‘Treatment with Indinavir, Zidovudine, and Lamivudine in Adults with Human Immunodeficiency Virus Infection and Prior Antiretroviral Therapy’, *New England Journal of Medicine*, 337(11), pp. 734–739. Available at: https://doi.org/10.1056/NEJM199709113371102.

Gutiérrez, S., Michalakis, Y. and Blanc, S. (2012) ‘Virus population bottlenecks during within-host progression and host-to-host transmission’, *Current Opinion in Virology*, 2(5), pp. 546–555. Available at: https://doi.org/10.1016/j.coviro.2012.08.001.

Haaland, R.E. *et al.* (2009) ‘Inflammatory Genital Infections Mitigate a Severe Genetic Bottleneck in Heterosexual Transmission of Subtype A and C HIV-1’, *PLoS Pathogens*. Edited by A. Trkola, 5(1), p. e1000274. Available at: https://doi.org/10.1371/journal.ppat.1000274.

Haddox, H.K. *et al.* (2018) ‘Mapping mutational effects along the evolutionary landscape of HIV envelope’, *eLife*, 7, p. e34420. Available at: https://doi.org/10.7554/eLife.34420.

Haller, B.C. and Messer, P.W. (2023) ‘SLiM 4: Multispecies Eco-Evolutionary Modeling’, *The American Naturalist*, 201(5), pp. E127–E139. Available at: https://doi.org/10.1086/723601.

Hedskog, C. *et al.* (2010) ‘Dynamics of HIV-1 Quasispecies during Antiviral Treatment Dissected Using Ultra-Deep Pyrosequencing’, *PLoS ONE*. Edited by D.F. Nixon, 5(7), p. e11345. Available at: https://doi.org/10.1371/journal.pone.0011345.

Hemelaar, J. *et al.* (2019) ‘Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis’, *The Lancet Infectious Diseases*, 19(2), pp. 143–155. Available at: https://doi.org/10.1016/S1473-3099(18)30647-9.

Henn, M.R. *et al.* (2012) 'Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection', *PLoS Pathogens*. Edited by C.M. Walker, 8(3), p. e1002529. Available at: https://doi.org/10.1371/journal.ppat.1002529.

Henrard, D.R. (1995) 'Natural History of HIV-1 Cell-Free Viremia', *JAMA: The Journal of the American Medical Association*, 274(7), p. 554. Available at: https://doi.org/10.1001/jama.1995.03530070052029.

Herbeck, J.T. *et al.* (2006) 'Human Immunodeficiency Virus Type 1 *env* Evolves toward Ancestral States upon Transmission to a New Host', *Journal of Virology*, 80(4), pp. 1637–1644. Available at: https://doi.org/10.1128/JVI.80.4.1637-1644.2006.

Hill, M., Tachedjian, G. and Mak, J. (2005) 'The Packaging and Maturation of the HIV-1 Pol Proteins', *Current HIV Research*, 3(1), pp. 73–85. Available at: https://doi.org/10.2174/1570162052772942.

Ho, S.Y.W. *et al.* (2011) 'Time-dependent rates of molecular evolution: TIME-DEPENDENT RATES OF MOLECULAR EVOLUTION', *Molecular Ecology*, 20(15), pp. 3087–3101. Available at: https://doi.org/10.1111/j.1365-294X.2011.05178.x.

Hodcroft, E., Hadfield, J.D., Fearnhill, E. *et al* (2014), 'The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection', *PLoS Pathog*; 10(5):e1004112. Available at: https://doi.org/10. 1371/journal.ppat.1004112

Hollingsworth, T.D. *et al.* (2010) 'HIV-1 Transmitting Couples Have Similar Viral Load Set-Points in Rakai, Uganda', *PLoS Pathogens*. Edited by E.C. Holmes, 6(5), p. e1000876. Available at: https://doi.org/10.1371/journal.ppat.1000876.

Holmes,E.C., 2010, 'The RNA Virus Quasispecies: Fact or Fiction?', *Journal of Molecular Biology*, 400(3), pp. 271 – 273. Available at https://doi.org/10.1016/j.jmb.2010.05.032.

Hool, A., Leventhal, G.E. and Bonhoeffer, S. (2013) 'Virus-induced target cell activation reconciles set-point viral load heritability and within-host evolution', *Epidemics*, 5(4), pp. 174–180. Available at: https://doi.org/10.1016/j.epidem.2013.09.002.

Hu, W.S. *et al.* (1997) 'Homologous recombination occurs in a distinct retroviral subpopulation and exhibits high negative interference', *Journal of Virology*, 71(8), pp. 6028–6036. Available at: https://doi.org/10.1128/jvi.71.8.6028-6036.1997.

Hubisz M, Siepel A. (2020), 'Inference of Ancestral Recombination Graphs Using ARGweaver', *Methods Mol Biol*. 2090:231-266. doi:10.1007/978-1-0716-0199-0_10

Hübner, W., McNerney G., Chen, P., et al., (2000) Quantitative 3D video microscopy of HIV transfer across T cell virological synapses, *Science*, 323, 1743–1747. Available at https://doi.org/10.1126/science.1167525

Igiraneza, A.B. *et al.* (2024) 'Learning patterns of HIV-1 resistance to broadly neutralizing antibodies with reduced subtype bias using multi-task learning', *PLOS Computational Biology*. Edited by R.R. Regoes, 20(11), p. e1012618. Available at: https://doi.org/10.1371/journal.pcbi.1012618.

Illingworth, C.J.R. *et al.* (2020) ‘A de novo approach to inferring within-host fitness effects during untreated HIV-1 infection’, *PLOS Pathogens*. Edited by A. Stern, 16(6), p. e1008171. Available at: https://doi.org/10.1371/journal.ppat.1008171.

Illingworth, C.J.R., Fischer, A. and Mustonen, V. (2014) ‘Identifying Selection in the Within-Host Evolution of Influenza Using Viral Sequence Data’, *PLoS Computational Biology*. Edited by C.O. Wilke, 10(7), p. e1003755. Available at: https://doi.org/10.1371/journal.pcbi.1003755.

Jaya, F.R., Brito, B.P. and Darling, A.E. (2023) ‘Evaluation of recombination detection methods for viral sequencing’, *Virus Evolution*, 9(2), p. vead066. Available at: https://doi.org/10.1093/ve/vead066.

Jenkins, F. *et al.* (2023) ‘Validation of an HIV whole genome sequencing method for HIV drug resistance testing in an Australian clinical microbiology laboratory’, *Journal of Medical Virology*, 95(12), p. e29273. Available at: https://doi.org/10.1002/jmv.29273.

Jetzt, A.E. *et al.* (2000) ‘High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome’, *Journal of Virology*, 74(3), pp. 1234–1240. Available at: https://doi.org/10.1128/JVI.74.3.1234-1240.2000.

Jia, L. *et al.* (2016) ‘Analysis of HIV-1 intersubtype recombination breakpoints suggests region with high pairing probability may be a more fundamental factor than sequence similarity affecting HIV-1 recombination’, *Virology Journal*, 13(1), p. 156. Available at: https://doi.org/10.1186/s12985-016-0616-1.

Joint United Nations Programme on HIV/AIDS (UNAIDS) (2024) *The urgency of now: AIDS at a crossroads*. Geneva: Joint United Nations Programme on HIV/AIDS.

Jones, N.A. *et al.* (2004) ‘Determinants of Human Immunodeficiency Virus Type 1 Escape from the Primary CD8+ Cytotoxic T Lymphocyte Response’, *The Journal of Experimental Medicine*, 200(10), pp. 1243–1256. Available at: https://doi.org/10.1084/jem.20040511.

Joseph, S.B. *et al.* (2015) ‘Bottlenecks in HIV-1 transmission: insights from the study of founder viruses’, *Nature Reviews. Microbiology*, 13(7), pp. 414–425. Available at: https://doi.org/10.1038/nrmicro3471.

Jung, A., Maier, R., Vartanian, J.P., *et al.* (2002) ‘Recombination: Multiply infected spleen cells in HIV patients’, *Nature*, 418(6894):144. doi:10.1038/418144a

Kalyaanamoorthy, S. *et al.* (2017) ‘ModelFinder: fast model selection for accurate phylogenetic estimates’, *Nature Methods*, 14(6), pp. 587–589. Available at: https://doi.org/10.1038/nmeth.4285.

Kaslow, R.A. *et al.* (1996) ‘Influence of combinations of human major histocompatibility complex genes on the course of HIV–1 infection’, *Nature Medicine*, 2(4), pp. 405–411. Available at: https://doi.org/10.1038/nm0496-405.

Katoh, K. (2002) ‘MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform’, *Nucleic Acids Research*, 30(14), pp. 3059–3066. Available at: https://doi.org/10.1093/nar/gkf436.

Kawashima, Y. *et al.* (2009) ‘Adaptation of HIV-1 to human leukocyte antigen class I’, *Nature*, 458(7238), pp. 641–645. Available at: https://doi.org/10.1038/nature07746.

Keele, B.F. *et al.* (2006) ‘Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1’, *Science*, 313(5786), pp. 523–526. Available at: https://doi.org/10.1126/science.1126531.

Kiepiela, P. *et al.* (2004) ‘Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA’, *Nature*, 432(7018), pp. 769–775. Available at: https://doi.org/10.1038/nature03113.

Kimata, J.T., Rice, A.P. and Wang, J. (2016) ‘Challenges and strategies for the eradication of the HIV reservoir’, *Current Opinion in Immunology*, 42, pp. 65–70. Available at: https://doi.org/10.1016/j.coi.2016.05.015.

Kist, N.C. *et al.* (2020) ‘HIV-1 p24Gag adaptation to modern and archaic HLA-allele frequency differences in ethnic groups contributes to viral subtype diversification’, *Virus Evolution*, 6(2), p. veaa085. Available at: https://doi.org/10.1093/ve/veaa085.

Klenerman, P. and Hill, A. (2005) ‘T cells and viral persistence: lessons from diverse infections’, *Nature Immunology*, 6(9), pp. 873–879. Available at: https://doi.org/10.1038/ni1241.

Klingler, J. *et al.* (2020) ‘How HIV-1 Gag Manipulates Its Host Cell Proteins: A Focus on Interactors of the Nucleocapsid Domain’, *Viruses*, 12(8), p. 888. Available at: https://doi.org/10.3390/v12080888.

Kløverpris, H.N., Leslie, A. and Goulder, P. (2016) ‘Role of HLA Adaptation in HIV Evolution’, *Frontiers in Immunology*, 6. Available at: https://doi.org/10.3389/fimmu.2015.00665.

Komarova, N.L, Wodar, D., (2013), ‘Virus dynamics in the presence of synaptic transmission’, *Math. Biosci*., 242, 161–171. Available at: https://doi.org/10.1016/j.mbs.2013.01.003

Koup, R.A. *et al.* (1994) ‘Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome’, *Journal of Virology*, 68(7), pp. 4650–4655. Available at: https://doi.org/10.1128/jvi.68.7.4650-4655.1994.

Kouyos, R.D. *et al.* (2011) ‘Assessing Predicted HIV-1 Replicative Capacity in a Clinical Setting’, *PLoS Pathogens*. Edited by D.C. Douek, 7(11), p. e1002321. Available at: https://doi.org/10.1371/journal.ppat.1002321.

Kwong, P.D. *et al.* (1998) ‘Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody’, *Nature*, 393(6686), pp. 648–659. Available at: https://doi.org/10.1038/31405.

Lauring, A.S. and Andino, R. (2010) ‘Quasispecies Theory and the Behavior of RNA Viruses’, *PLoS Pathogens*. Edited by M. Manchester, 6(7), p. e1001005. Available at: https://doi.org/10.1371/journal.ppat.1001005.

Leitman, E.M. *et al.* (2016) ‘Lower Viral Loads and Slower $CD4^+$ T-Cell Count Decline in MRKAd5 HIV-1 Vaccinees Expressing Disease-Susceptible HLA-B*58:02’, *Journal of Infectious Diseases*, 214(3), pp. 379–389. Available at: https://doi.org/10.1093/infdis/jiw093.

Leitner, T. (2018) ‘The Puzzle of HIV Neutral and Selective Evolution’, *Molecular Biology and Evolution*. Edited by S. Kumar, 35(6), pp. 1355–1358. Available at: https://doi.org/10.1093/molbev/msy089.

Lemey, P. *et al.* (2003) ‘Tracing the origin and history of the HIV-2 epidemic’, *Proceedings of the National Academy of Sciences of the United States of America*, 100(11), pp. 6588–6592. Available at: https://doi.org/10.1073/pnas.0936469100.

Lemey, P. *et al.* (2007) ‘Synonymous Substitution Rates Predict HIV Disease Progression as a Result of Underlying Replication Dynamics’, *PLOS Computational Biology*, 3(2): e29.

Lemey, P., Rambaut, A. and Pybus, O.G. (2006) ‘HIV Evolutionary Dynamics Within and Among Hosts’, *AIDS Reviews*, 8:125-40.

Lemey, P., Vandamme, A.-M. and Salemi, M. (2009) *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. 2nd edn. Cambridge University Press.

Leslie, A. *et al.* (2005) ‘Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA’, *The Journal of Experimental Medicine*, 201(6), pp. 891–902. Available at: https://doi.org/10.1084/jem.20041455.

Leslie, A.J. *et al.* (2004) ‘HIV evolution: CTL escape mutation and reversion after transmission’, *Nature Medicine*, 10(3), pp. 282–289. Available at: https://doi.org/10.1038/nm992.

Levy, D.N. *et al.* (2004) ‘Dynamics of HIV-1 recombination in its natural target cells’, *Proceedings of the National Academy of Sciences*, 101(12), pp. 4204–4209. Available at: https://doi.org/10.1073/pnas.0306764101.

Li, B. *et al.* (2007) ‘Rapid Reversion of Sequence Polymorphisms Dominates Early Human Immunodeficiency Virus Type 1 Evolution’, *Journal of Virology*, 81(1), pp. 193–201. Available at: https://doi.org/10.1128/JVI.01231-06.

Li, X. *et al.* (2023) ‘Molecular basis of differential HLA class I-restricted T cell recognition of a highly networked HIV peptide’, *Nature Communications*, 14(1), p. 2929. Available at: https://doi.org/10.1038/s41467-023-38573-8.

Lingappa, J.R. *et al.* (2009) ‘Characteristics of HIV-1 Discordant Couples Enrolled in a Trial of HSV-2 Suppression to Reduce HIV-1 Transmission: The Partners Study’,

PLoS ONE. Edited by S. Emery, 4(4), p. e5272. Available at: https://doi.org/10.1371/journal.pone.0005272.

Lingappa, J.R. *et al.* (2011) 'Genomewide Association Study for Determinants of HIV-1 Acquisition and Viral Set Point in HIV-1 Serodiscordant Couples with Quantified Virus Exposure', *PLoS ONE*. Edited by R.F. Speck, 6(12), p. e28632. Available at: https://doi.org/10.1371/journal.pone.0028632.

Little, S.J. *et al.* (2008) 'Persistence of Transmitted Drug Resistance among Subjects with Primary Human Immunodeficiency Virus Infection', *Journal of Virology*, 82(11), pp. 5510–5518. Available at: https://doi.org/10.1128/JVI.02579-07.

Liu, D. *et al.* (2014) 'Preexisting compensatory amino acids compromise fitness costs of a HIV-1 T cell escape mutation', *Retrovirology*, 11(1), p. 101. Available at: https://doi.org/10.1186/s12977-014-0101-0.

Liu, R. *et al.* (1996) 'Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection', *Cell*, 86(3), pp. 367–377. Available at: https://doi.org/10.1016/S0092-8674(00)80110-5.

Liu, Y. *et al.* (2006) 'Selection on the Human Immunodeficiency Virus Type 1 Proteome following Primary Infection', *Journal of Virology*, 80(19), pp. 9519–9529. Available at: https://doi.org/10.1128/JVI.00575-06.

Liu, Y. *et al.* (2019) 'Onward Transmission of Multiple HIV-1 Unique Recombinant Forms Among Men Who Have Sex With Men in Beijing, China', *J Acquir Immune Defic Syndr*, 81(1).

Loh, L. *et al.* (2007) 'In Vivo Fitness Costs of Different Gag CD8 T-Cell Escape Mutant Simian-Human Immunodeficiency Viruses for Macaques', *Journal of Virology*, 81(10), pp. 5418–5422. Available at: https://doi.org/10.1128/JVI.02763-06.

Los Alamos National Laboratory (2024a) 'HIV Molecular Immunology Database'. Available at: https://www.hiv.lanl.gov/content/immunology/ (Accessed: 18 September 2024).

Los Alamos National Laboratory (2024b) 'HIV Sequence Database'. Available at: https://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html (Accessed: 18 September 2024).

Lyles, R.H. *et al.* (2000) 'Natural History of Human Immunodeficiency Virus Type 1 Viremia after Seroconversion and Proximal to AIDS in a Large Cohort of Homosexual Men', *The Journal of Infectious Diseases*, 181(3), pp. 872–880. Available at: https://doi.org/10.1086/315339.

Lynch, R.M. *et al.* (2009) 'Appreciating HIV Type 1 Diversity: Subtype Differences in Env', *AIDS Research and Human Retroviruses*, 25(3), pp. 237–248. Available at: https://doi.org/10.1089/aid.2008.0219.

Lythgoe, K.A. *et al.* (2017) 'Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections', *Trends in Microbiology*, 25(5), pp. 336–348. Available at: https://doi.org/10.1016/j.tim.2017.03.003.

Lythgoe, K.A. and Fraser, C. (2012) ‘New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels’, *Proceedings of the Royal Society B: Biological Sciences*, 279(1741), pp. 3367–3375. Available at: https://doi.org/10.1098/rspb.2012.0595.

Lythgoe, K.A., Pellis, L. and Fraser, C. (2013) ‘IS HIV SHORT-SIGHTED? INSIGHTS FROM A MULTISTRAIN NESTED MODEL’, *Evolution*, 67(10), pp. 2769–2782. Available at: https://doi.org/10.1111/evo.12166.

Magiorkinis, G. *et al.* (2003) ‘In vivo characteristics of human immunodeficiency virus type 1 intersubtype recombination: determination of hot spots and correlation with sequence similarity’, *Journal of General Virology*, 84(10), pp. 2715–2722. Available at: https://doi.org/10.1099/vir.0.19180-0.

Maljkovic Berry, I. *et al.* (2009) ‘The evolutionary rate dynamically tracks changes in HIV-1 epidemics: Application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data’, *Epidemics*, 1(4), pp. 230–239. Available at: https://doi.org/10.1016/j.epidem.2009.10.003.

Mansky, L.M. (1996a) ‘Forward Mutation Rate of Human Immunodeficiency Virus Type 1 in a T Lymphoid Cell Line*’, *AIDS Research and Human Retroviruses*, 12(4), pp. 307–314. Available at: https://doi.org/10.1089/aid.1996.12.307.

Mansky, L.M. (1996b) ‘The Mutation Rate of Human Immunodeficiency Virus Type 1 Is Influenced by thevprGene’, *Virology*, 222(2), pp. 391–400. Available at: https://doi.org/10.1006/viro.1996.0436.

Martin, M.P. (1998) ‘Genetic Acceleration of AIDS Progression by a Promoter Variant of CCR5’, *Science*, 282(5395), pp. 1907–1911. Available at: https://doi.org/10.1126/science.282.5395.1907.

Matthews, P.C. *et al.* (2009) ‘HLA Footprints on Human Immunodeficiency Virus Type 1 Are Associated with Interclade Polymorphisms and Intraclade Phylogenetic Clustering’, *Journal of Virology*, 83(9), pp. 4605–4615. Available at: https://doi.org/10.1128/JVI.02017-08.

McLaren, P.J. *et al.* (2015) ‘Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load’, *Proceedings of the National Academy of Sciences*, 112(47), pp. 14658–14663. Available at: https://doi.org/10.1073/pnas.1514867112.

McMichael, A.J. *et al.* (2010) ‘The immune response during acute HIV-1 infection: clues for vaccine development’, *Nature Reviews Immunology*, 10(1), pp. 11–23. Available at: https://doi.org/10.1038/nri2674.

McMichael, A.J. and Rowland-Jones, S.L. (2001) ‘Cellular immune responses to HIV’, *Nature*, 410(6831), pp. 980–987. Available at: https://doi.org/10.1038/35073658.

Mellors, J.W. *et al.* (1996) 'Prognosis in HIV-1 Infection Predicted by the Quantity of Virus in Plasma', *Science*, 272(5265), pp. 1167–1170. Available at: https://doi.org/10.1126/science.272.5265.1167.

Migueles, S.A. *et al.* (2000) 'HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors', *Proceedings of the National Academy of Sciences*, 97(6), pp. 2709–2714. Available at: https://doi.org/10.1073/pnas.050567397.

Mikhail, M. *et al.* (2005) 'Role of viral evolutionary rate in HIV-1 disease progression in a linked cohort', *Retrovirology*, 2(1), p. 41. Available at: https://doi.org/10.1186/1742-4690-2-41.

Minh, B.Q. *et al.* (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*. Edited by E. Teeling, 37(5), pp. 1530–1534. Available at: https://doi.org/10.1093/molbev/msaa015.

Mitov, V. and Stadler, T. (2018) 'A Practical Guide to Estimating the Heritability of Pathogen Traits', *Molecular Biology and Evolution*, 35(3), pp. 756–772. Available at: https://doi.org/10.1093/molbev/msx328.

Miura, T. *et al.* (2009) '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-Lymphotye Recognition', *Journal of Virology*, 83(6), pp. 2743–2755. Available at: https://doi.org/10.1128/JVI.02265-08.

Moore, C.B. *et al.* (2002) 'Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level', *Science*, 296(5572), pp. 1439–1443. Available at: https://doi.org/10.1126/science.1069660.

Moradigaravand, D. *et al.* (2014) 'Recombination Accelerates Adaptation on a Large-Scale Empirical Fitness Landscape in HIV-1', *PLoS Genetics*. Edited by M. Worobey, 10(6), p. e1004439. Available at: https://doi.org/10.1371/journal.pgen.1004439.

Moss, A.R. *et al.* (1988) 'Seropositivity for HIV and the development of AIDS or AIDS related condition: three year follow up of the San Francisco General Hospital cohort', *BMJ*, 296(6624), pp. 745–750. Available at: https://doi.org/10.1136/bmj.296.6624.745.

Mujugira, A. *et al.* (2011) 'Characteristics of HIV-1 Serodiscordant Couples Enrolled in a Clinical Trial of Antiretroviral Pre-Exposure Prophylaxis for HIV-1 Prevention', *PLoS ONE*. Edited by C. Thorne, 6(10), p. e25828. Available at: https://doi.org/10.1371/journal.pone.0025828.

Nagaraja, P., Alexander, H.K., Bonhoefferm, S, Dixit N.M., (2015), 'Influence of recombination on acquisition and reversion of immune escape and compensatory mutations in HIV-1', *Epidemics,* 14:11-25. doi: 10.1016/j.epidem.2015.09.001. Epub 2015 Oct 18. PMID: 26972510.

Nasir, A. *et al.* (2021) ‘Large Evolutionary Rate Heterogeneity among and within HIV-1 Subtypes and CRFs’, *Viruses*, 13(9), p. 1689. Available at: https://doi.org/10.3390/v13091689.

Navis, M. *et al.* (2007) ‘Viral Replication Capacity as a Correlate of HLA B57/B5801-Associated Nonprogressive HIV-1 Infection’, *The Journal of Immunology*, 179(5), pp. 3133–3143. Available at: https://doi.org/10.4049/jimmunol.179.5.3133.

Neher, R.A. and Leitner, T. (2010a) ‘Recombination Rate and Selection Strength in HIV Intra-patient Evolution’, *PLoS Computational Biology*. Edited by C. Fraser, 6(1), p. e1000660. Available at: https://doi.org/10.1371/journal.pcbi.1000660.

Neher, R.A. and Leitner, T. (2010b) ‘Recombination Rate and Selection Strength in HIV Intra-patient Evolution’, *PLoS Computational Biology*. Edited by C. Fraser, 6(1), p. e1000660. Available at: https://doi.org/10.1371/journal.pcbi.1000660.

Neogi, U. *et al.* (2017) ‘Recent increased identification and transmission of HIV-1 unique recombinant forms in Sweden’, *Scientific Reports*, 7(1), p. 6371. Available at: https://doi.org/10.1038/s41598-017-06860-2.

Nikolaitchik, O.A., Galli, A., Moore. M.D, Pathak, V.K., and Hu, W-S, (2011), ‘Multiple barriers to recombination between divergent HIV-1 variants revealed by a dual-marker recombination assay.’, *J Mol Biol*, 407(4):521–531, Apr 2011. doi: 10.1016/j.jmb.2011.01.052

Nikolaitchik, O.A., Dilley, K.A., Fu, W., et al. (2013) ‘Dimeric RNA Recognition Regulates HIV-1 Genome Packaging’, *PLOS Pathogens* 9(3): e1003249. https://doi.org/10.1371/journal.ppat.1003249

Nijhuis, M. *et al.* (1998) ‘Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy’, *Proceedings of the National Academy of Sciences*, 95(24), pp. 14441–14446. Available at: https://doi.org/10.1073/pnas.95.24.14441.

Novitsky, V. *et al.* (2013) ‘Intra-host evolutionary rates in HIV-1C env and gag during primary infection’, *Infection, Genetics and Evolution*, 19, pp. 361–368. Available at: https://doi.org/10.1016/j.meegid.2013.02.023.

O’Brien, S.J., Gao, X. and Carrington, M. (2001) ‘HLA and AIDS: a cautionary tale’, *Trends in Molecular Medicine*, 7(9), pp. 379–381. Available at: https://doi.org/10.1016/S1471-4914(01)02131-1.

Onafuwa, A. *et al.* (2003) ‘Human Immunodeficiency Virus Type 1 Genetic Recombination Is More Frequent Than That of Moloney Murine Leukemia Virus despite Similar Template Switching Rates’, *Journal of Virology*, 77(8), pp. 4577–4587. Available at: https://doi.org/10.1128/JVI.77.8.4577-4587.2003.

Oxenius, A. *et al.* (2004) ‘Loss of viral control in early HIV-1 infection is temporally associated with sequential escape from CD8+ T cell responses and decrease in HIV-1-specific CD4+ and CD8+ T cell frequencies’, *The Journal of Infectious Diseases*, 190(4), pp. 713–721. Available at: https://doi.org/10.1086/422760.

Payne, R. *et al.* (2014) ‘Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence’, *Proceedings of the National Academy of Sciences*, 111(50). Available at: https://doi.org/10.1073/pnas.1413339111.

Perelson, A.S. *et al.* (1996) ‘HIV-1 Dynamics in Vivo: Virion Clearance Rate, Infected Cell Life-Span, and Viral Generation Time’, *Science*, 271(5255), pp. 1582–1586. Available at: https://doi.org/10.1126/science.271.5255.1582.

Phillips, R.E. *et al.* (1991) ‘Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition’, *Nature*, 354(6353), pp. 453–459. Available at: https://doi.org/10.1038/354453a0.

Piantadosi, A. *et al.* (2009) ‘HIV-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response’, *AIDS (London, England)*, 23(5), pp. 579–587. Available at: https://doi.org/10.1097/QAD.0b013e328328f76e.

Plata, F. *et al.* (1987) ‘AIDS virus-specific cytotoxic T lymphocytes in lung disorders’, *Nature*, 328(6128), pp. 348–351. Available at: https://doi.org/10.1038/328348a0.

Poignard, P. *et al.* (1999) ‘Neutralizing Antibodies Have Limited Effects on the Control of Established HIV-1 Infection In Vivo’, *Immunity*, 10(4), pp. 431–438. Available at: https://doi.org/10.1016/S1074-7613(00)80043-6.

Prince, J.L. *et al.* (2012) ‘Role of Transmitted Gag CTL Polymorphisms in Defining Replicative Capacity and Early HIV-1 Pathogenesis’, *PLoS Pathogens*. Edited by J. Lifson, 8(11), p. e1003041. Available at: https://doi.org/10.1371/journal.ppat.1003041.

Pybus, O.G. and Rambaut, A. (2009) ‘Evolutionary analysis of the dynamics of viral infectious disease’, *Nature Reviews Genetics*, 10(8), pp. 540–550. Available at: https://doi.org/10.1038/nrg2583.

Quinn, T.C. *et al.* (2000) ‘Viral Load and Heterosexual Transmission of Human Immunodeficiency Virus Type 1’, *New England Journal of Medicine*, 342(13), pp. 921–929. Available at: https://doi.org/10.1056/NEJM200003303421303.

Raghwani, J. *et al.* (2016) ‘Exceptional Heterogeneity in Viral Evolutionary Dynamics Characterises Chronic Hepatitis C Virus Infection’, *PLoS pathogens*, 12(9), p. e1005894. Available at: https://doi.org/10.1371/journal.ppat.1005894.

Raghwani, J. *et al.* (2018) ‘Evolution of HIV-1 within untreated individuals and at the population scale in Uganda’, *PLOS Pathogens*. Edited by P. Lemey, 14(7), p. e1007167. Available at: https://doi.org/10.1371/journal.ppat.1007167.

Raghwani, J. *et al.* (2019) ‘High-Resolution Evolutionary Analysis of Within-Host Hepatitis C Virus Infection’, *The Journal of Infectious Diseases*, 219(11), pp. 1722–1729. Available at: https://doi.org/10.1093/infdis/jiy747.

Rambaut, A. *et al.* (2004) ‘The causes and consequences of HIV evolution’, *Nature Reviews Genetics*, 5(1), pp. 52–61. Available at: https://doi.org/10.1038/nrg1246.

Ramírez De Arellano, E. *et al.* (2019) 'Novel association of five HLA alleles with HIV-1 progression in Spanish long-term non progressor patients', *PLOS ONE*. Edited by S. Mummidi, 14(8), p. e0220459. Available at: https://doi.org/10.1371/journal.pone.0220459.

Rawson, J.M.O. *et al.* (2018) 'Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity', *Nucleic Acids Research* [Preprint]. Available at: https://doi.org/10.1093/nar/gky910.

Redd, A.D. *et al.* (2012) 'Previously Transmitted HIV-1 Strains Are Preferentially Selected During Subsequent Sexual Transmissions', *The Journal of Infectious Diseases*, 206(9), pp. 1433–1442. Available at: https://doi.org/10.1093/infdis/jis503.

Richardson, B.A. *et al.* (2003) 'Comparison of Human Immunodeficiency Virus Type 1 Viral Loads in Kenyan Women, Men, and Infants during Primary and Early Infection', *Journal of Virology*, 77(12), pp. 7120–7123. Available at: https://doi.org/10.1128/JVI.77.12.7120-7123.2003.

Roberts, H.E. *et al.* (2015) 'Structured Observations Reveal Slow HIV-1 CTL Escape', *PLOS Genetics*. Edited by T. Gojobori, 11(2), p. e1004914. Available at: https://doi.org/10.1371/journal.pgen.1004914.

Roberts, J.D., Bebenek, K. and Kunkel, T.A. (1988) 'The Accuracy of Reverse Transcriptase from HIV-1', *Science*, 242(4882), pp. 1171–1173. Available at: https://doi.org/10.1126/science.2460925.

Rocha, C. *et al.* (2013) 'Evolution of the human immunodeficiency virus type 2 envelope in the first years of infection is associated with the dynamics of the neutralizing antibody response', *Retrovirology*, 10, p. 110. Available at: https://doi.org/10.1186/1742-4690-10-110.

Romero, E.V. and Feder, A. (2024) 'Elevated HIV Viral Load is Associated with Higher Recombination Rate In Vivo'.

Ross, M.G., Russ, C., Costello, M. et al. (2013) 'Characterizing and measuring bias in sequence data', Genome Biol 14, R51. https://doi.org/10.1186/gb-2013-14-5-r51

Rouzine, I.M. and Coffin, J.M. (1999) 'Linkage disequilibrium test implies a large effective population number for HIV *in vivo*', *Proceedings of the National Academy of Sciences*, 96(19), pp. 10758–10763. Available at: https://doi.org/10.1073/pnas.96.19.10758.

Sanborn, K.B. *et al.* (2015) 'Recombination elevates the effective evolutionary rate and facilitates the establishment of HIV-1 infection in infants after mother-to-child transmission', *Retrovirology*, 12(1), p. 96. Available at: https://doi.org/10.1186/s12977-015-0222-0.

Sardanyés, J., Perales, C., Domingo, E. et al. 'Quasispecies theory and emerging viruses: challenges and applications', *npj Viruses* 2(54), https://doi.org/10.1038/s44298-024-00066-w

Sather, D.N. *et al.* (2009) 'Factors Associated with the Development of Cross-Reactive Neutralizing Antibodies during Human Immunodeficiency Virus Type 1

Infection', *Journal of Virology*, 83(2), pp. 757–769. Available at: https://doi.org/10.1128/JVI.02036-08.

Schierup, M.H. and Hein, J. (2000) 'Consequences of Recombination on Traditional Phylogenetic Analysis', *Genetics*, 156(2), pp. 879–891. Available at: https://doi.org/10.1093/genetics/156.2.879.

Schlub, T.E. *et al.* (2010) 'Accurately Measuring Recombination between Closely Related HIV-1 Genomes', *PLoS Computational Biology*. Edited by C. Fraser, 6(4), p. e1000766. Available at: https://doi.org/10.1371/journal.pcbi.1000766.

Schneidewind, A. *et al.* (2009) 'Transmission and Long-Term Stability of Compensated CD8 Escape Mutations', *Journal of Virology*, 83(8), pp. 3993–3997. Available at: https://doi.org/10.1128/JVI.01108-08.

Selhorst, P. *et al.* (2017) 'Replication Capacity of Viruses from Acute Infection Drives HIV-1 Disease Progression', *Journal of Virology*. Edited by F. Kirchhoff, 91(8), pp. e01806-16. Available at: https://doi.org/10.1128/JVI.01806-16.

Sengupta, S. and Siliciano, R.F. (2018) 'Targeting the Latent Reservoir for HIV-1', *Immunity*, 48(5), pp. 872–895. Available at: https://doi.org/10.1016/j.immuni.2018.04.030.

Shankarappa, R. *et al.* (1999) 'Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection', *Journal of Virology*, 73(12), pp. 10489–10502. Available at: https://doi.org/10.1128/JVI.73.12.10489-10502.1999.

Shriner, D. *et al.* (2004) 'Pervasive Genomic Recombination of HIV-1 in Vivo', *Genetics*, 167(4), pp. 1573–1583. Available at: https://doi.org/10.1534/genetics.103.023382.

Simek, M.D. *et al.* (2009) 'Human Immunodeficiency Virus Type 1 Elite Neutralizers: Individuals with Broad and Potent Neutralizing Activity Identified by Using a High-Throughput Neutralization Assay together with an Analytical Selection Algorithm', *Journal of Virology*, 83(14), pp. 7337–7348. Available at: https://doi.org/10.1128/JVI.00110-09.

Simon-Loriere, E. *et al.* (2009) 'Molecular Mechanisms of Recombination Restriction in the Envelope Gene of the Human Immunodeficiency Virus', *PLoS Pathogens*. Edited by E.C. Holmes, 5(5), p. e1000418. Available at: https://doi.org/10.1371/journal.ppat.1000418.

Simon-Loriere, E. *et al.* (2010) 'RNA Structures Facilitate Recombination-Mediated Gene Swapping in HIV-1', *Journal of Virology*, 84(24), pp. 12675–12682. Available at: https://doi.org/10.1128/JVI.01302-10.

Smith, S.J. *et al.* (2021) 'Integrase Strand Transfer Inhibitors Are Effective Anti-HIV Drugs', *Viruses*, 13(2), p. 205. Available at: https://doi.org/10.3390/v13020205.

Smyth, R.P. *et al.* (2014) ‘Identifying Recombination Hot Spots in the HIV-1 Genome’, *Journal of Virology*. Edited by B.H. Hahn, 88(5), pp. 2891–2902. Available at: https://doi.org/10.1128/JVI.03014-13.

Song, H. *et al.* (2018) ‘Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection’, *Nature Communications*, 9(1), p. 1928. Available at: https://doi.org/10.1038/s41467-018-04217-5.

Suchard, M.A. *et al.* (2018) ‘Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10’, *Virus Evolution*, 4(1). Available at: https://doi.org/10.1093/ve/vey016.

Tang, J. *et al.* (2002) ‘Favorable and Unfavorable HLA Class I Alleles and Haplotypes in Zambians Predominantly Infected with Clade C Human Immunodeficiency Virus Type 1’, *Journal of Virology*, 76(16), pp. 8276–8284. Available at: https://doi.org/10.1128/JVI.76.16.8276-8284.2002.

Tay, J.H., Kocher, A. and Duchene, S. (2024) ‘Assessing the effect of model specification and prior sensitivity on Bayesian tests of temporal signal’, *PLOS Computational Biology*. Edited by Y. Lu, 20(11), p. e1012371. Available at: https://doi.org/10.1371/journal.pcbi.1012371.

Theys, K. *et al.* (2018) ‘The impact of HIV-1 within-host evolution on transmission dynamics’, *Current Opinion in Virology*, 28, pp. 92–101. Available at: https://doi.org/10.1016/j.coviro.2017.12.001.

Timm, J. *et al.* (2004) ‘CD8 Epitope Escape and Reversion in Acute HCV Infection’, *The Journal of Experimental Medicine*, 200(12), pp. 1593–1604. Available at: https://doi.org/10.1084/jem.20041006.

Tomaras, G.D. *et al.* (2008) ‘Initial B-Cell Responses to Transmitted Human Immunodeficiency Virus Type 1: Virion-Binding Immunoglobulin M (IgM) and IgG Antibodies Followed by Plasma Anti-gp41 Antibodies with Ineffective Control of Initial Viremia’, *Journal of Virology*, 82(24), pp. 12449–12463. Available at: https://doi.org/10.1128/JVI.01708-08.

Tongo, M., De Oliveira, T. and Martin, D.P. (2018) ‘Patterns of genomic site inheritance in HIV-1M inter-subtype recombinants delineate the most likely genomic sites of subtype-specific adaptation’, *Virus Evolution*, 4(1). Available at: https://doi.org/10.1093/ve/vey015.

Townsend, A. *et al.* (1989) ‘Association of class I major histocompatibility heavy and light chains induced by viral peptides’, *Nature*, 340(6233), pp. 443–448. Available at: https://doi.org/10.1038/340443a0.

Trautmann, L. *et al.* (2006) ‘Upregulation of PD-1 expression on HIV-specific CD8+ T cells leads to reversible immune dysfunction’, *Nature Medicine*, 12(10), pp. 1198–1202. Available at: https://doi.org/10.1038/nm1482.

Troyer, R.M. *et al.* (2005) ‘Changes in Human Immunodeficiency Virus Type 1 Fitness and Genetic Diversity during Disease Progression’, *Journal of Virology*,

79(14), pp. 9006–9018. Available at: https://doi.org/10.1128/JVI.79.14.9006-9018.2005.

Troyer, R.M. *et al.* (2009) 'Variable Fitness Impact of HIV-1 Escape Mutations to Cytotoxic T Lymphocyte (CTL) Response', *PLoS Pathogens*, 5(4).

Turner, B.G. and Summers, M.F. (1999) 'Structural biology of HIV 1 1Edited by P. E. Wright', *Journal of Molecular Biology*, 285(1), pp. 1–32. Available at: https://doi.org/10.1006/jmbi.1998.2354.

Van Dorp, C.H., Van Boven, M. and De Boer, R.J. (2014) 'Immuno-epidemiological Modeling of HIV-1 Predicts High Heritability of the Set-Point Virus Load, while Selection for CTL Escape Dominates Virulence Evolution', *PLoS Computational Biology*. Edited by R.R. Regoes, 10(12), p. e1003899. Available at: https://doi.org/10.1371/journal.pcbi.1003899.

Vrancken, B. *et al.* (2014) 'The Genealogical Population Dynamics of HIV-1 in a Large Transmission Chain: Bridging within and among Host Evolutionary Rates', *PLoS Computational Biology*. Edited by C. Fraser, 10(4), p. e1003505. Available at: https://doi.org/10.1371/journal.pcbi.1003505.

Vrancken, B. *et al.* (2015) 'Disentangling the impact of within-host evolution and transmission dynamics on the tempo of HIV-1 evolution', *AIDS*, 29(12), pp. 1549–1556. Available at: https://doi.org/10.1097/QAD.0000000000000731.

Vrancken, B., Suchard, M.A. and Lemey, P. (2017) 'Accurate quantification of within- and between-host HBV evolutionary rates requires explicit transmission chain modelling', *Virus Evolution*, 3(2), p. vex028. Available at: https://doi.org/10.1093/ve/vex028.

Walker, B.D. *et al.* (1987) 'HIV-specific cytotoxic T lymphocytes in seropositive individuals', *Nature*, 328(6128), pp. 345–348. Available at: https://doi.org/10.1038/328345a0.

Wang, M. and Claesson, M.H. (2014) 'Classification of Human Leukocyte Antigen (HLA) Supertypes', in R.K. De and N. Tomar (eds) *Immunoinformatics*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 309–317. Available at: https://doi.org/10.1007/978-1-4939-1115-8_17.

Wange, R.L. and Samelson, L.E. (1996) 'Complex Complexes: Signaling at the TCR', *Immunity*, 5(3), pp. 197–205. Available at: https://doi.org/10.1016/S1074-7613(00)80315-5.

Wei, X. *et al.* (2003) 'Antibody neutralization and escape by HIV-1', *Nature*, 422(6929), pp. 307–312. Available at: https://doi.org/10.1038/nature01470.

West, A.P. *et al.* (2014) 'Structural Insights on the Role of Antibodies in HIV-1 Vaccine and Therapy', *Cell*, 156(4), pp. 633–648. Available at: https://doi.org/10.1016/j.cell.2014.01.052.

Whitney, J.B. *et al.* (2014) ‘Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys’, *Nature*, 512(7512), pp. 74–77. Available at: https://doi.org/10.1038/nature13594.

Wilke, C.O. (2005) ‘Quasispecies theory in the context of population genetics’, *BMC Evolutionary Biology*, 5(1), p. 44. Available at: https://doi.org/10.1186/1471-2148-5-44.

Williams, A. *et al.* (2023) ‘Geographic and Population Distributions of Human Immunodeficiency Virus (HIV)–1 and HIV-2 Circulating Subtypes: A Systematic Literature Review and Meta-analysis (2010–2021)’, *The Journal of Infectious Diseases*, 228(11), pp. 1583–1591. Available at: https://doi.org/10.1093/infdis/jiad327.

Woo, J., Robertson, D.L. and Lovell, S.C. (2014) ‘Constraints from protein structure and intra-molecular coevolution influence the fitness of HIV-1 recombinants’, *Virology*, 454–455, pp. 34–39. Available at: https://doi.org/10.1016/j.virol.2014.01.029.

Wood, N. *et al.* (2009) ‘HIV Evolution in Early Infection: Selection Pressures, Patterns of Insertion and Deletion, and the Impact of APOBEC’, *PLoS Pathogens*. Edited by D.C. Douek, 5(5), p. e1000414. Available at: https://doi.org/10.1371/journal.ppat.1000414.

Wright, S. (1920) ‘The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs’, *Proceedings of the National Academy of Sciences*, 6(6), pp. 320–332. Available at: https://doi.org/10.1073/pnas.6.6.320.

Wymant, C., Blanquart, F., *et al.* (2018) ‘Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver’, *Virus Evolution*, 4(1). Available at: https://doi.org/10.1093/ve/vey007.

Wymant, C., Hall, M., *et al.* (2018) ‘PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity’, *Molecular Biology and Evolution*, 35(3), pp. 719–733. Available at: https://doi.org/10.1093/molbev/msx304.

Yewdell, J.W. and Bennink, J.R. (1992) ‘Cell Biology of Antigen Processing and Presentation to Major Histocompatibility Complex Class I Molecule-Restricted T Lymphocytes’, in *Advances in Immunology*. Elsevier, pp. 1–123. Available at: https://doi.org/10.1016/S0065-2776(08)60875-5.

Yokomaku, Y. *et al.* (2004) ‘Impaired Processing and Presentation of Cytotoxic-T-Lymphocyte (CTL) Epitopes Are Major Escape Mechanisms from CTL Immune Pressure in Human Immunodeficiency Virus Type 1 Infection’, *Journal of Virology*, 78(3), pp. 1324–1332. Available at: https://doi.org/10.1128/JVI.78.3.1324-1332.2004.

Zanini, F. *et al.* (2015) ‘Population genomics of intrapatient HIV-1 evolution’, *eLife*, 4, p. e11282. Available at: https://doi.org/10.7554/eLife.11282.

Zanini, F. *et al.* (2017) '*In vivo* mutation rates and the landscape of fitness costs of HIV-1', *Virus Evolution*, 3(1), p. vex003. Available at: https://doi.org/10.1093/ve/vex003.

Zanini, F. and Neher, R.A. (2013) 'Quantifying Selection against Synonymous Mutations in HIV-1 *env* Evolution', *Journal of Virology*, 87(21), pp. 11843–11850. Available at: https://doi.org/10.1128/JVI.01529-13.

Zhang, X. *et al.* (2013) 'HLA-B*44 is associated with a lower viral set point and slow CD4 decline in a cohort of Chinese homosexual men acutely infected with HIV-1', *Clinical and vaccine immunology: CVI*, 20(7), pp. 1048–1054. Available at: https://doi.org/10.1128/CVI.00015-13.

Zhuang, J. *et al.* (2002) 'Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots', *Journal of Virology*, 76(22), pp. 11273–11282. Available at: https://doi.org/10.1128/JVI.76.22.11273-11282.2002.

9 Appendix

9.1 Chapter 3 Supplementary material

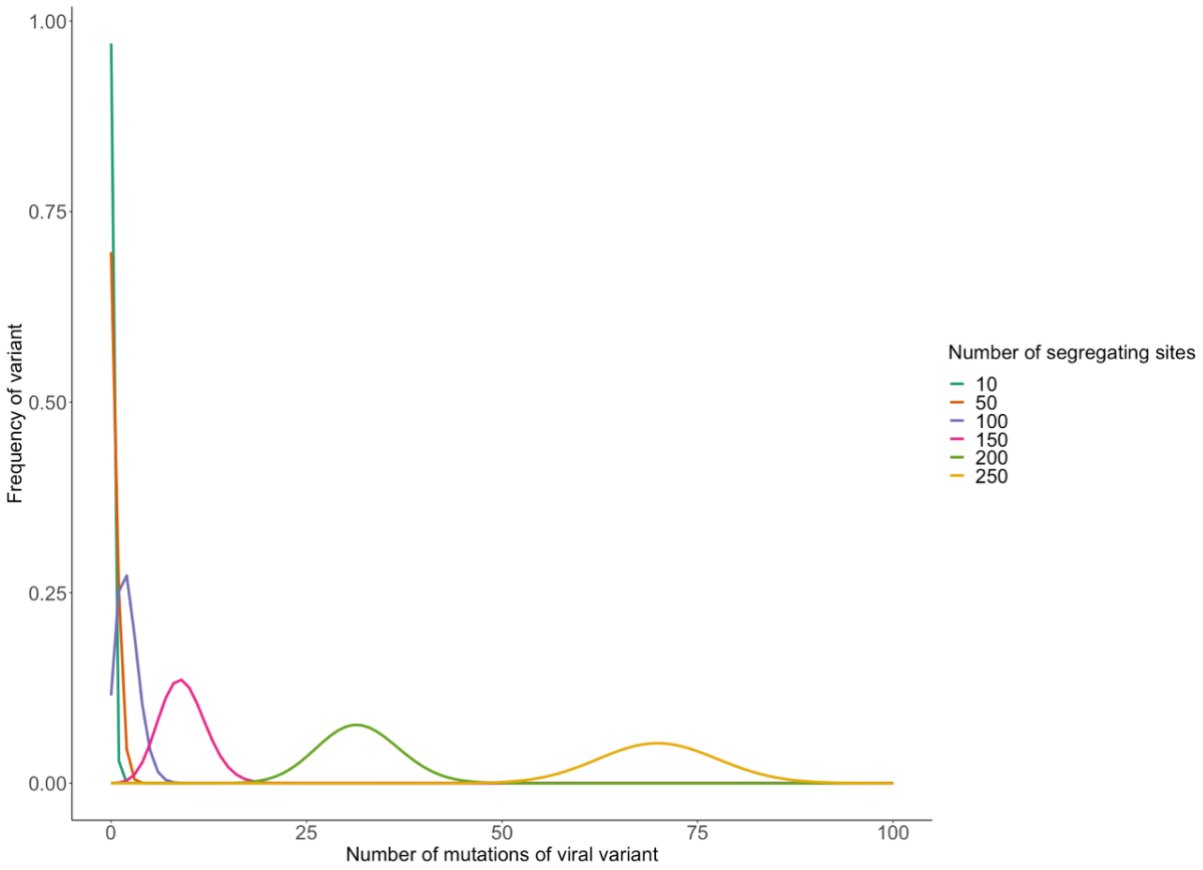

Figure 9.1 The frequency distribution of the viral population at equilibria. A specific viral type is defined by its number of mutations. When we consider few mutations of large effect, the population is dominated by a single virus type of high relative fitness. As we increase the number of segregating sites and lower the associated fitness cost, the population becomes increasingly diverse, which has implications for the viral variants that are transmitted and between-host evolution. This diversity in the within-host viral population creates a broader pool of viral variants available for transmission, potentially influencing both transmission dynamics and the trajectory of between-host evolution

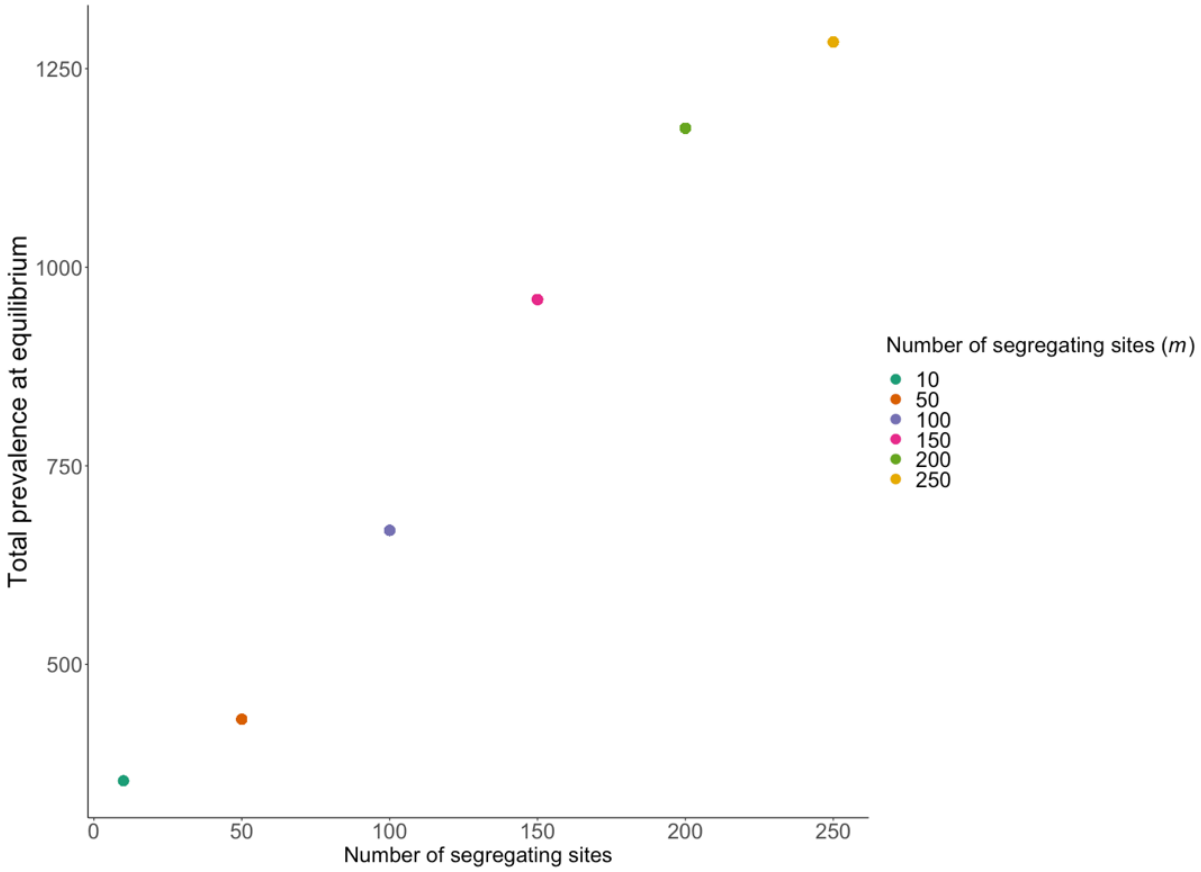

Figure 9.2 The total prevalence at the endemic steady state. As the number of segregating sites increases, the within-host dynamics slow down and the viral population is more diverse. As a result, between-host selection is able to select the virus types with greatest transmission potential, ultimately increasing the endemic prevalence.

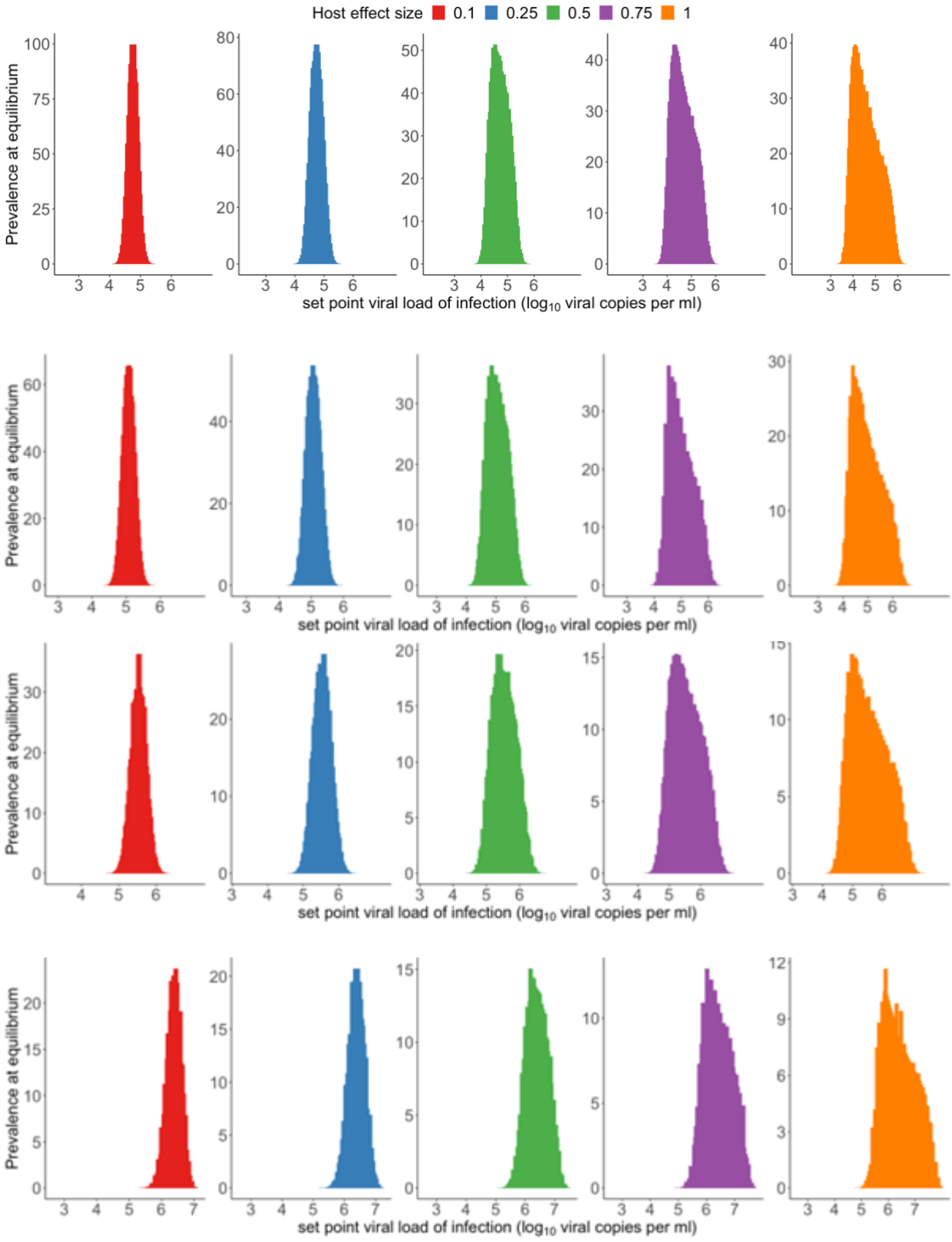

Figure 9.3 Between-host outcomes at equilibrium for an increasingly heterogenous host population for all models. Histograms of spVLs in heterogenous infected population for different maximum host effect size, *e,* for A)200 B)150, C)100 and D) 50 segregating sites. To account for the effect that host genetics has on viral load, we introduced a host specific additive effect to viral load. The size of the host effect is discretely uniformly distributed between -*e* and *e* and there are 50 host types. Increasing the host-effect broadens the distribution of spVLs.

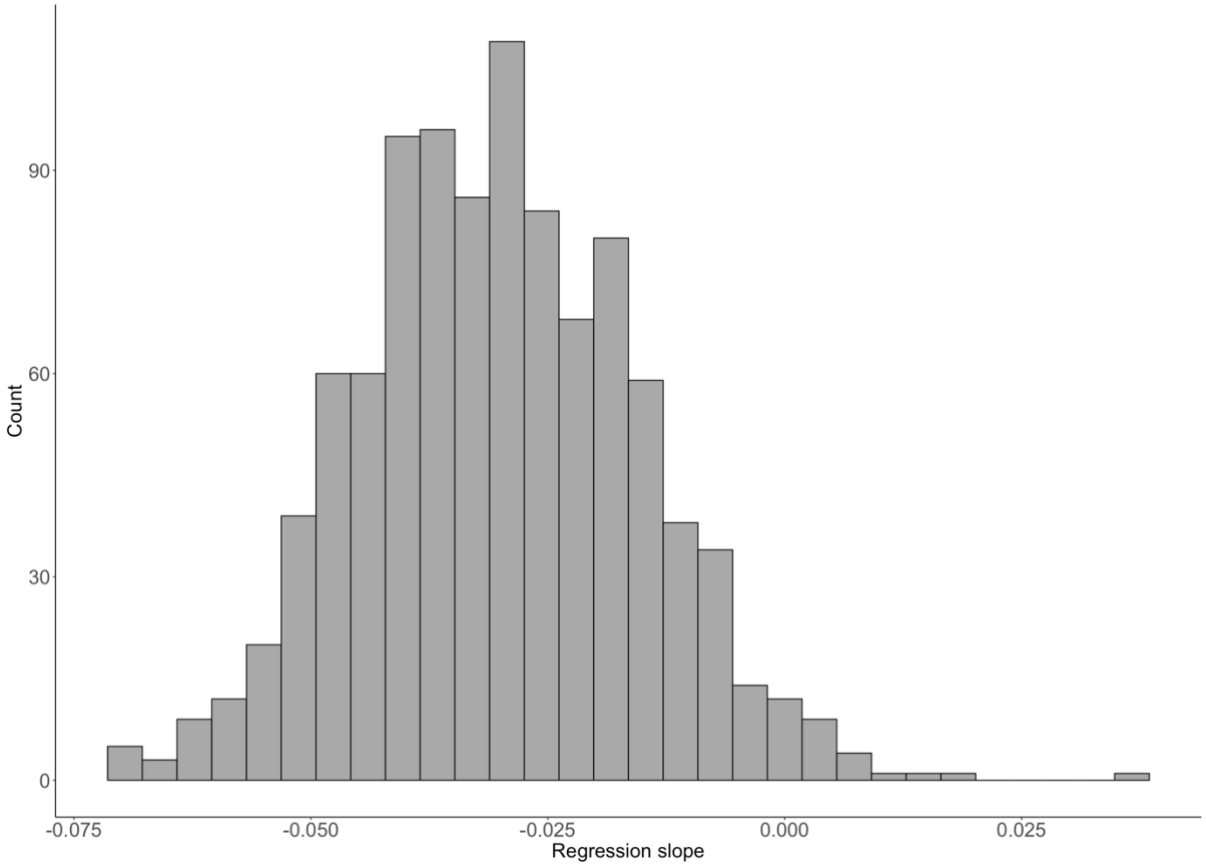

Figure 9.4 Regression slopes of simulated viral load trajectories with noise. Sets of 100 viral load trajectories with 3 measurements over time, with variation by noise generated by noise alone. The regression slope is the relationship between the within-host temporal trend in change in viral load with the initial viral load measure. The process was repeated 1000 times, and in every simulation the association was weaker than the value inferred from the real dataset of viral loads (Figure 3.7, main text).

9.2 Chapter 4 Supplementary material

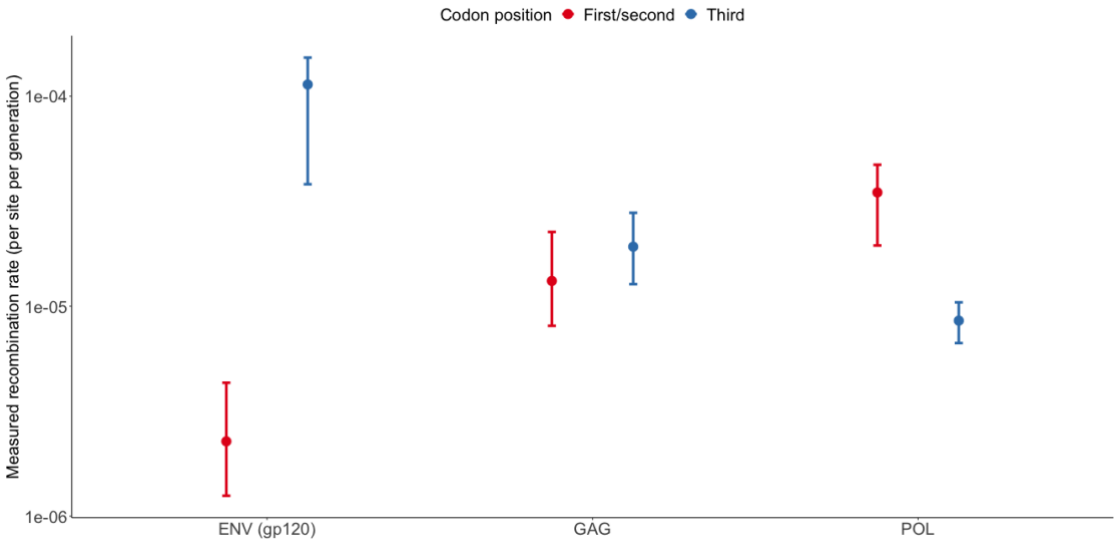

Figure 9.5 Substantial differences in recombination rate in ENV by codon position for sites at which allele frequency is stable. Pairs of sites have been filtered to only those at which the change in frequency across two time points is lower than 10% for both sites. The recombination rate is estimated separately by codon position (1 and 2 vs 3) for non-overlapping regions of ENV (GP120), GAG and POL. Error bars represent 95% confidence intervals generated by bootstrap replicates. Mutations at third codon positions are used as a proxy for synonymous changes and first/second for non-synonymous. A lower rate of measured recombination in pairs of sites at first/second positions signifies selection against recombinants.

9.3 Chapter 5 Supplementary material

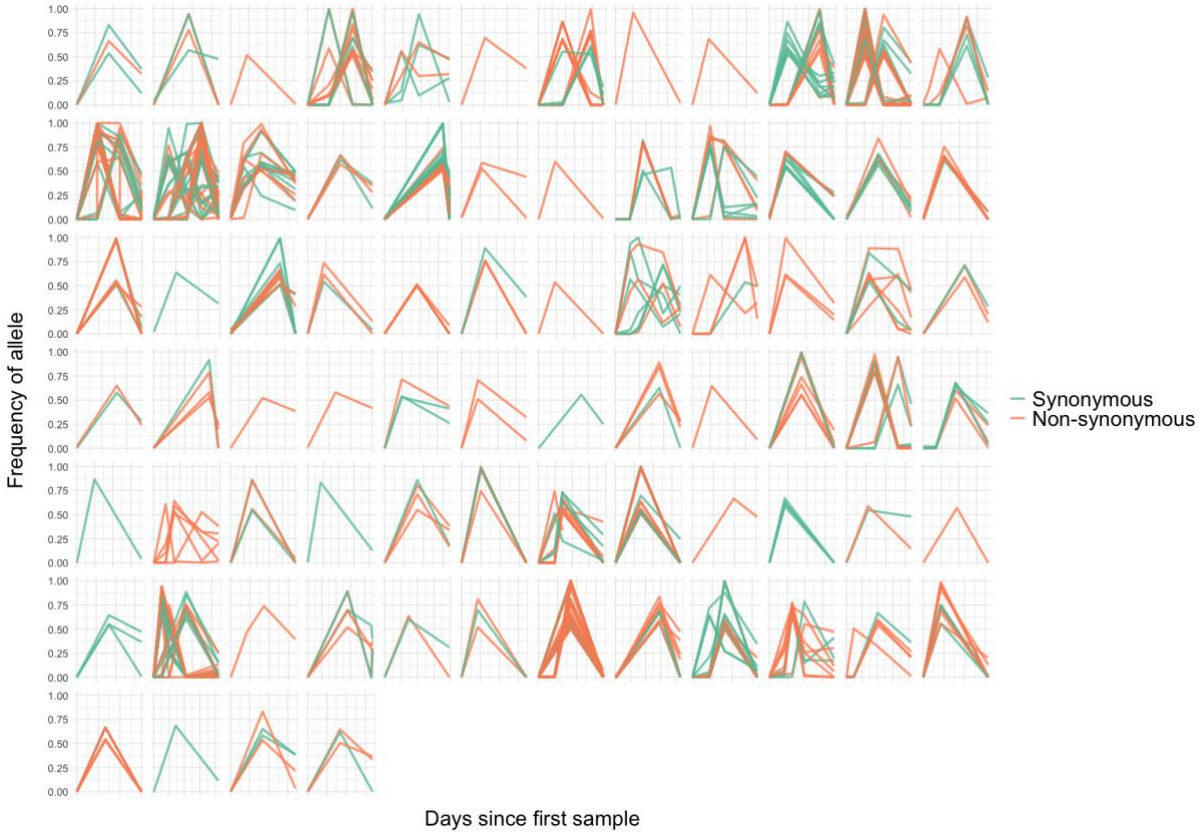

Figure 9.6 Trajectories of toggling mutations for all individuals. Toggling is defined by a change in consensus (50% frequency) which is later reversed (frequency less than 50%). Each grid visualises trajectories from a single individual. Across individuals there is substantial differences in the number of toggling mutations. For visualisation purposes, the x-axis is scaled differently for each individual.

9.4 Chapter 6 Supplementary material

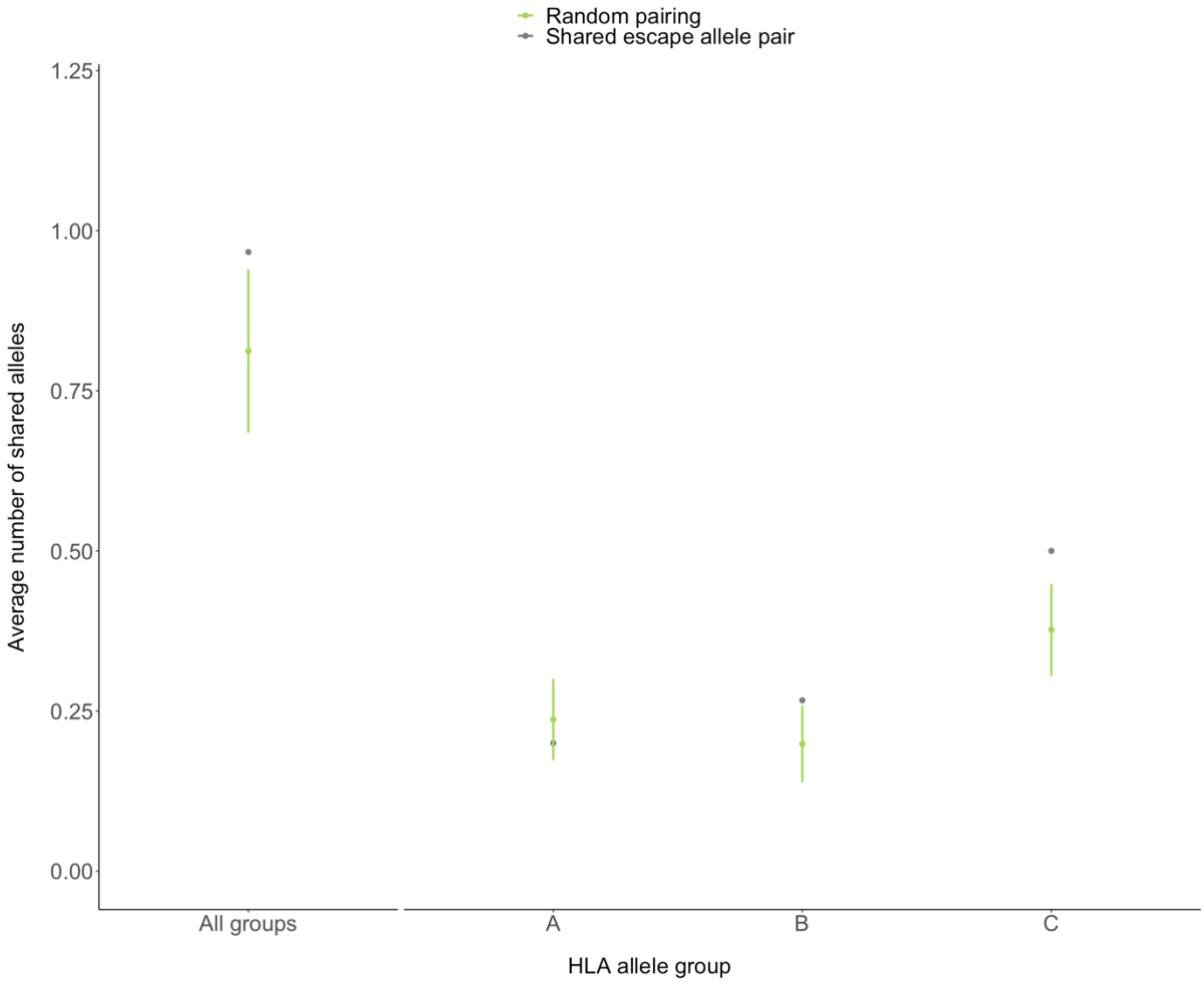

Figure 9.7 Individuals with similar HLA profiles drive the same escape mutation more often than by chance The average number of shared alleles between a pair sharing an escape allele for all HLA groups and for HLA-A, B and C specifically, where the escapes are filtered to only those identified without searching known epitope restrictions of the HLA. Green bars indicate confidence intervals of the mean number of shared alleles in 1,000 sets of 100 randomly selected pairs of individuals from the study (NB: not equivalent to transmission pair), representing a null distribution. The average number of shared alleles in pairs that share an allele is significantly higher than for the null comparator data, with exception of HLA allele group A. The equivalent plot with all escapes is given in the main text.

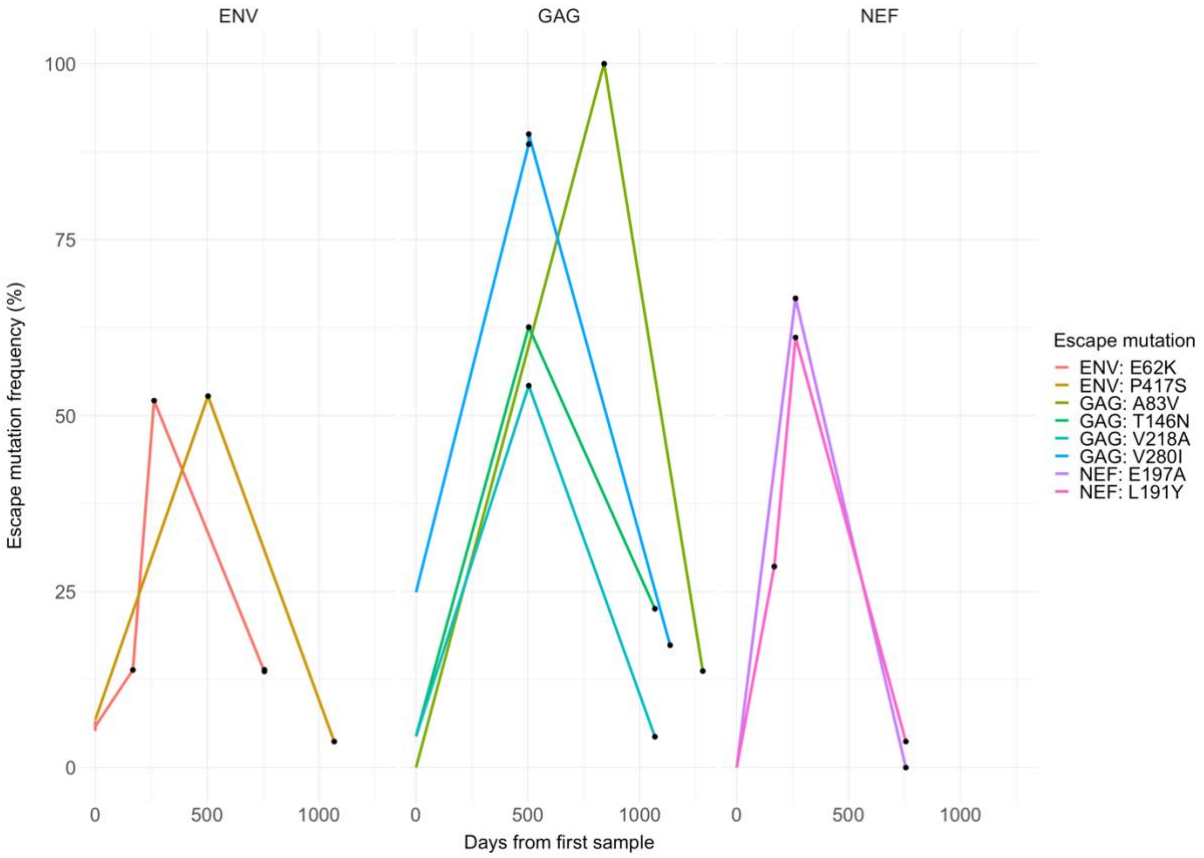

Figure 9.8 Within-host trajectories of toggling mutations in source individuals. The within-host trajectories of 8 escape mutations that are initially selected for but are later selected against, all within a single source individual, and often in favour of a mutation within the same epitope. Each trajectory is from a different individual infection. The escape mutation is defined in the legend.

Table 9.1 Positions of between-host toggling sites. A site was defined as toggling between hosts if the site was classified as a CTL escape site observed to be evolving under selection in the source, the site was transmitted, and there was then evidence of negative selection in the recipient. Sites are described by the name of the gene followed by the protein position. The variant amino acid (AA) is the variant under selection and transmitted, and the base AA is the subtype-level consensus AA under selection in the host.

Site (Gene, HXB2 pos)	Variant AA	Base AA
ENV_192	R	I
ENV_330	H	Y
ENV_344	K	E
ENV_347	K	E
ENV_347	T	I
ENV_350	R	G
ENV_415	T	I
ENV_85	H	P
GAG_124	N	S
GAG_127	K	T
GAG_31	L	I
NEF_187	S	C
NEF_191	R	Q
NEF_71	R	K
NEF_74	V	L
NEF_81	Y	F
NEF_81	Y	F
NEF_87	L	I
POL_68	T	I
POL_68	S	P
POL_70	K	R