ABSTRACT

HAMINS-PUÉRTOLAS, MARCO GABRIEL. Modeling and Quantifying Selection Pressures Across Multiple Evolutionary Scales. (Under the direction of David Rasmussen).

In host-pathogen systems, pathogen strains can be influenced by varying levels of fitness at the within host and between host scales. This can arise through changes in the ability to compete at the within host scale or through variation in the host’s ability to transmit. Here we develop two mathematical models that model both the host and the population level. The first is a set of nested Wright-Fisher models where we analyze how key parameters including bottleneck size, time between transmission events, and selection impact probability of fixation and time to fixation. We find that both host population and bottleneck size have a role in determining probability of fixation. As the time between transmission events increases, generations to fixation is almost entirely dependent on host population size. This model provides a quantitative framework to study how population dynamics alter evolutionary dynamics in viruses across scales. The second model we develop is a phylodynamic birth-death model that allows selection to act on a continuous-valued trait both at the within and between hosts scales. We use this method to infer the magnitude and direction of selection pressures within and between hosts along with other relevant parameters like mutation rates and effective population size from phylogenetic trees. Each pathogen lineage is assigned a continuous “trait” value that is an element of some closed and bounded set. In the context of a host-pathogen system, these continuous traits could represent a pathogen's viral load, allele (strain) frequency or a measure of genetic distance. We demonstrate using simulation based approaches that given a phylogeny with continuous-valued tip states, we can perform Bayesian inference utilizing our continuous type birth-death branching model. We apply this approach to answer how two continuous traits, viral load and distance to the consensus, impact transmission and evolutionary dynamics of HIV-1. We demonstrate that an intermediate viral load is associated with optimal
transmission potential, a pattern that is hypothesized to be present under the virulence trade-off hypothesis. For our application to the distance to consensus we were motivated by the fact that HIV-1 is characterized by rapid diversification within-host during the acute phase of infection. However, not all of these amino acid substitutions will be beneficial at the between host scale. In fact, we find correspondence across three distinct HIV-1 data sets that confirms the genome undergoes rapid diversification post-transmission, but is under strong purifying selection at the between-host scale.
Modeling and Quantifying Selection Pressures Across Multiple Evolutionary Scales

by
Marco Gabriel Hamins-Puértolas

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Biomathematics

Raleigh, North Carolina
2021

APPROVED BY:

Alun Lloyd                                  Cristina Lanzas

Kevin Gross                                  Jeff Thorne

David Rasmussen
Chair of Advisory Committee
DEDICATION

To David for your constant support. I am honored to be your first student.

To Clotilde, Anthony, and Adrian. For your love and support over the years.
BIOGRAPHY

Marco Hamins-Puértolas was born in Washington D.C. in June of 1994. After graduating from Winston Churchill High School he attended St. Mary's College of Maryland, a public university in St. Mary's County, Maryland near the confluence of the Potomac River and Chesapeake Bay where he double majored in Biology and Mathematics. It was during his time at St. Mary's that he attended a Research Experience for Undergraduates at the Mathematical and Theoretical Biology Institute hosted by Arizona State University. This program provided his first introduction to the world of mathematical modeling and is where he fell in love with the intersection of biology and mathematics. A few months later he applied to the Biomathematics program at North Carolina State University and accepted an offer to begin in the fall of 2016. He joined David Rasmussen's lab in the summer of 2018 and began working on multiscale phylodynamic models.
ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. David Rasmussen, for his incalculable support. His thoughtful feedback and contributions to this work have made me a stronger writer and better scientist. Dr. Ruian Ke for the countless moments of excitement that we shared when developing the first chapter of my thesis. I also want to thank the members of my committee for the insightful comments and questions raised throughout my time as a graduate student. In particular Dr. Alun Lloyd for serving as thoughtful mentor and my office soccer conversationalist. To my previous mentors, Dr. Emek Köse and Dr. Samantha Elliot, who helped me develop my first independent research topic while at St. Mary’s College of Maryland. Additionally, the past and present members of the Rasmussen Lab who have provided helpful feedback and exciting scientific conversations over the years.

To Hadley who has always been there for me throughout my time in graduate school. I truly believe I would not be where I am today without you. To the many incredible friendships I have developed during my time at NC State - from consuming cookout with Julian, cooking dinners with John, talking PDEs with Michael, and cutting Brandon’s hair. To the entire Biomathematics program for the research and non-research related conversations throughout the years. To my Craigslist roommates turned life-long friends, Anton and Eric, who provided many much needed escapes from graduate school. To my Covid “pod” who helped me maintain some semblance of sanity during difficult and stressful times.

Last but certainly not least, to my incredible family. I would not be where I am today without my parents to whom I am forever grateful. To my brother for always being there to share a laugh, a beer, or both. To my late grandparents who I know would be so proud. Lastly, to the Stryer, Evans, Puértolas, and Pitas, who I am so lucky to call my family.
# TABLE OF CONTENTS

List of Tables ................................................................. vii

List of Figures ............................................................... viii

Chapter 1 Introduction ..................................................... 1
  1.1 Application to HIV .................................................... 3
  1.2 Dissertation Outline ................................................. 6

Chapter 2 Modeling stochastic viral evolution: A multiscale Wright-Fisher model 7
  2.1 Introduction ........................................................... 8
  2.2 Methods ............................................................... 10
    2.2.1 Host Level Selection types ..................................... 14
    2.2.2 Deterministic Addition ........................................ 16
  2.3 Results ............................................................... 17
    2.3.1 Neutral Mutation .............................................. 17
    2.3.2 Host Level Selection ........................................... 21
    2.3.3 Within-Host Level Selection .................................. 24
    2.3.4 Selection at multiple scales ................................... 25
  2.4 Discussion ........................................................... 26
  2.5 Supplemental ......................................................... 28
    2.5.1 First Step Analysis ............................................ 28

Chapter 3 Inferring multiscale selection pressures from pathogen phylogenies using a Continuous Type Birth-Death Model 32
  3.1 Introduction ........................................................... 33
  3.2 Methods ............................................................... 35
    3.2.1 Multi-type Birth-Death Model ................................. 35
    3.2.2 Continuous-Type Birth-Death Model ......................... 38
    3.2.3 Strain Frequency Derivation .................................. 45
    3.2.4 Parameter estimation .......................................... 46
    3.2.5 Simulation: CTBD ............................................. 47
  3.3 Application to HIV-1 ................................................ 50
    3.3.1 Background .................................................... 50
    3.3.2 Transmission .................................................... 51
    3.3.3 Application Specific Methods ................................. 54
  3.4 Results ............................................................... 55
  3.5 Discussion ........................................................... 58
  3.6 Appendix ............................................................. 63
    3.6.1 Full Transmission Kernel ...................................... 63
    3.6.2 Simulation: CTBD ............................................. 65
    3.6.3 Viral Load Application ........................................ 71
    3.6.4 Extensions ...................................................... 72
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Descriptions for parameters and metrics utilized in the paper.</td>
<td>14</td>
</tr>
<tr>
<td>3.1</td>
<td>Likelihood and Bayes Factor calculation for various functional forms of the relationship between VL and fitness, ( f(x) ). The functional forms are exponential growth, Michaelis-Menten, and cubic splines with six knots.</td>
<td>71</td>
</tr>
<tr>
<td>A.1</td>
<td>A summary of acronyms used in alphabetical order.</td>
<td>127</td>
</tr>
<tr>
<td>A.1</td>
<td>A summary of acronyms used in alphabetical order.</td>
<td>128</td>
</tr>
<tr>
<td>B.1</td>
<td>A summary of parameters, variables, and metrics followed by their definitions in alphabetical order. The chapter in which they are utilized is also noted.</td>
<td>129</td>
</tr>
<tr>
<td>B.1</td>
<td>A summary of parameters, variables, and metrics followed by their definitions in alphabetical order. The chapter in which they are utilized is also noted.</td>
<td>130</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 Generalized Wright-Fisher model where $f_{i,j}$ represents the frequency of the $i$th host in the $j$th generation. Within each host, genetic drift is at play for $g_s$ generations. .............................................. 12

Figure 2.2 Host representation of multiscale Wright Fisher model. Darker shades of blue correspond to higher frequency of mutant allele. The red circle distinguishes the MRCA of the hosts from the fixation event (last row). (a) Resulting outcome of fixation event where MRCA is fully mutant. (b) Resulting outcome of fixation event where MRCA is semi mutant. 19

Figure 2.3 (a) Impact of varying levels of within ($N_w$) and between ($N_b$) host population size on probability of fixation. (b) Impact of varying levels of within-host population size ($N_w$) and stochastic generations ($g_s$) on the expected number of generations to fixation. Simulated assuming $N_b = 100$, $g_d = s_b = s_w = 0$ and $g_s = 1$ unless otherwise specified. .............................................................. 21

Figure 2.4 Probability of fixation under varying host level selection coefficients ($s_b$). Strong selection (a) and weak selection (b) are depicted. Points represent simulation results, solid lines correspond to numerical results from the branching process model and the dotted line is an analytical approximation to fixation for two sequential Wright-Fisher models. Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified. .............................................. 23

Figure 2.5 Effect of host level selection on generations to fixation for weak (left) and strong selection (right). Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified. ................................. 23

Figure 2.6 Probability of fixation under varying within-host level selection coefficients ($s_w$). This is paired with (a) varying levels of within-host stochastic generations, $g_s$, and (b) varying levels of within-host deterministic generations. The dotted lines correspond to analytical results from branching process model while points represent simulation results. Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_b = 0$ and $g_s = 1$ unless otherwise specified. ............................................... 25

Figure 2.7 Three exemplary host level selection paradigms. Presence, proportion, and fully mutant dependent. ............................................. 28

Figure 2.8 Percent difference between numerical simulations and analytical solution for generations to fixation across the host and within-host population size parameter regimes under a neutral model. .......... 30
Figure 2.9  Average number of generations to fixation under varying within-host
level selection coefficients ($s_w$). This is paired with (a) varying levels of
within-host stochastic generations, $g_s$ and (b) varying levels of within-
host stochastic generations. The solid and dashed lines correspond
to analytical results from diffusion and branching process theory
respectively, while points represent simulation results. Simulated
assuming $N_b = 1000, N_w = 10, g_d = s_b = 0$ and $g_s = 1$ unless otherwise
specified. .......................................................... 31

Figure 2.10  Probability of fixation when both host and within-host selection co-
efficients are positive. Strong selection (a) and weak selection (b) for
both scales are depicted. Points represent simulation results, solid
lines correspond to numerical results from the branching process
model and the dotted line is an analytical approximation to fixa-
tion for two sequential Wright-Fisher models. Simulated assuming
$N_b = 1000, N_w = 10, g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified. 31

Figure 3.1  (a) Phylogeny produced through forward simulation. The true state
of a lineage is visualized and sampled tips are represented by open
points. (b) Solution of the PDE representing probability density $D_N(x, t)$
backwards through time (Eq. 3.9) for the two lineages of interest
(boxed). Red and green curves along with their color gradients corre-
spond to the probability distribution for the noted lineage and time.
...................................................................................... 43

Figure 3.2  Illustration of how phylogenetic structure and sampled tip states
can impact phylodynamic parameter estimates. Starting from point
masses at the tips ($t = 0$) we can follow the probability distribu-
tion backwards in time to the root ($t = t_0$). In one tree, phylogenetic
relatedness is correlated with sampled state (a) while it is more in-
dependent in (b). Red (blue) represents the probability distribution
solved backwards in time for a large (small) diffusion coefficient. (c)
An example of how selective forces can be informed. Red (blue) rep-
resents the probability distribution solved backwards in time for a
large (small) selective coefficient. ........................................ 45
Figure 3.3  Phylodynamic analysis using the CTBD for within and between host selection pressures. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host variability, backward ($\delta^b$) and forward ($\delta^f$) selection. Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across the continuous trait space (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the fitness value of all trait values and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions. 48

Figure 3.4  Phylodynamic analysis using the CTBD for within and between host selection pressures. Estimates of median parameter values from ten forward simulations under the same evolutionary scenario. Simulations were conducted with a negative relationship between the continuous trait and fitness at the between host scale while no within host selection bias is present. 50

Figure 3.5  Maximum likelihood reconstructed phylogeny for the 529 sequenced protease and reverse transcriptase coding regions of HIV-1 infected hosts in King County, WA. Tips and branches are colored according to observed or reconstructed $\log_{10}$ viral load of the host using PastML (Ishikawa et al. 2019). 54

Figure 3.6  Phylodynamic analysis of 529 HIV-1 paired samples of $\log_{10}$ viral load per ml and PR/RT sequences using the CTBD. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host evolution, rate of selection ($\log_{10}$ copies per ml to consensus per year) and diffusion ($\log_{10}$ copies per ml to consensus$^2$ per year). Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across log VL (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the functional form and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions. 57

Figure 3.7  Phylodynamic analysis of 529 HIV-1 paired samples of $\log_{10}$ viral load and PR/RT sequences using the CTBD. Posterior distribution for the fitness mapping function weighted by the KDE estimate for the density of host $\log_{10}$ viral loads per ml. The 95% credible intervals are shown in shaded intervals. 58
Figure 3.8 Impact of varying decay rates on the proportion of $\beta_0$ found on the transmission kernel $\lambda(x, y)$. Note that $|x - y|$ is presented here as the transmission kernel is symmetric. ............................... 64

Figure 3.9 Phylodynamic analysis using the CTBD for within and between host selection pressures. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host variability, backward ($\delta^b$) and forward ($\delta^f$) selection. Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across the continuous trait space (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the functional form and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions. ............................... 65

Figure 3.10 Estimates of median parameter values for ten forward simulations under the same evolutionary scenario. Simulations were conducted with neutral processes driving the evolution of a continuous trait at both the within and between host scale. Estimates for all simulations surround the ground truth for both within and between host parameters. ................................................. 66

Figure 3.11 Estimates of median parameter values for ten forward simulations under the same evolutionary scenario. Simulations were conducted with a positive selection pressure between continuous trait and fitness at the between host scale while no within host selection bias is present. Estimates for all simulations surround the ground truth for both within and between host parameters. ............................... 67

Figure 3.12 Estimates of median parameter values under the strain frequency parameterization found in Section 3.2.3. The true underlying parameters were $N_e = 100$, $s_w = -1$, and $s_h = .2$. ................................................. 67

Figure 3.13 Estimates for the between-host transmission rate under the strain frequency parameterization found in Section 3.2.3. The 50% and 90% credible intervals are found in light and dark blue while the mean and true fitness curves are the solid black and dashed red lines respectively. The true underlying parameters were $N_e = 100$, $s_w = -1$, and $s_h = .2$. ................................................. 68

Figure 3.14 Estimates of median parameter values under the strain frequency parameterization found in Section 3.2.3. The true underlying parameters were $N_e = 100$, $s_w = .1$, and $s_h = -1$. ................................................. 68
Figure 3.15 Estimates for the between-host transmission rate under the strain frequency parameterization found in Section 3.2.3. The 50% and 90% credible intervals are found in light and dark blue while the mean and true fitness curves are the solid black and dashed red lines respectively. The true underlying parameters were $N_e = 100$, $s_w = .1$, and $s_h = -1$. 69

Figure 3.16 Distribution of $\log_{10}$ viral loads for HIV-1 positive individuals sequenced in King County, WA. Each individual also had the region of their genome that codes for the protease and reverse transcriptase enzymes sequenced. A reconstructed phylogeny is presented in Figure 3.5 71

Figure 4.1 Left: Estimates of the distance to consensus for internal nodes. Blue points represent true tip state while red points represent the inferred median state using the CTBD. Bars represent the 95% HPI. Right: Maximum likelihood reconstructed phylogeny for the sequenced protease and reverse transcriptase coding regions of HIV-1 infected hosts in King County, WA. Tips and branches are colored according to observed or reconstructed distance to consensus of the host. Distance to consensus for ancestral nodes and branches were reconstructed using PastML. (Ishikawa et al. 2019) 89

Figure 4.2 Left: Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the LANL data set. Median estimates are surrounded by the 95% credible interval. Note that fitness here is relative to the average transmission rate. The median fitness for $env$ and $pol$ at the population consensus remains above one, only suggesting that these viral lineages transmit more than the average transmission rate expected under a simple birth-death process. Right: Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the partial or full $pol$ genes from the three data sets. Median estimates are surrounded by the 95% credible interval (Ishikawa et al. 2019) 91

Figure 4.3 Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the LANL data set. Median estimates are surrounded by the 95% credible interval. We first estimated the background birth rate prior to running the full CTBD. The median fitness for $env$ and $pol$ at the population consensus remains above one, only suggesting that these viral lineages transmit more than the average transmission rate expected under a simple birth-death process 95
Figure 4.4 Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the LANL data set. Minimum distance to consensus values were used across each region independently for the lower bound during analysis with the CTBD. Median estimates are surrounded by the 95% credible interval. 96

Figure 4.5 Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the LANL data set. Minimum distance to consensus values were used across each region independently for the lower bound during analysis with the CTBD. Black vertical lines represent the median along with the 95% credible interval. 97

Figure 4.6 Distance to in-sample consensus measured across time for the sub-sampled phylogenies with 500 tips derived from the larger Swiss HIV-1 MSM data set. Each color is representative of a different phylogeny. 98

Figure 4.7 Tanglegram for the four genomic regions of interest from the LANL data set. From left to right: gag, pol, accessory, and env. Colors are representative of relative location shift across phlogenies in reference to gag gene. 99

Figure 4.8 Distance to in-sample consensus measured across time for the protease and reverse transcriptase protein coding regions found in 529 HIV-1 positive individuals from King County, WA. 100

Figure 4.9 Distance to in-sample consensus measured across time for the four regions of interest and the full genome sequences from the LANL data set. 101

Figure 4.10 Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the Swiss data set. Median estimates are surrounded by the 95% credible interval. 102

Figure 4.11 Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the Swiss data set. Black vertical lines represent the median along with the 95% credible interval. 103

Figure 4.12 Scatterplot of distance to in-sample consensus and viral load (log10 viral copies per ml) with color corresponding to their sampled date for the LANL data set. We find there to be no correlation between these two traits with a $R^2 = 1.5e-5$. 103

Figure 4.13 Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the LANL data set. Black vertical lines represent the median along with the 95% credible interval. 104

Figure 4.14 Posterior estimates for the selection and diffusion parameters of the CTBD for the partial or full pol genes from the three data sets. Black vertical lines represent the median along with the 95% credible interval. 104
Figure 4.15 Comparison between true continuous state and estimate from CTBD at internal nodes. Blue points represent the true state of internal nodes, green points represent true tip state, while red points represent the inferred median state using the CTBD. Bars represent the 95% credible interval.
The field of multiscale modeling has grown rapidly in the past few years. This is a natural progression as we now sit on a wealth of knowledge about the rules that govern the processes on individual scales. The world is inherently multi-level giving rise to “Matryoska dolls” of scales, with each containing another. Cells are contained within tissues, which are found within an individual who find themselves part of a sub-population. This layering covers the entirety of biological systems and so it is not surprising to see that multiscale models are ubiquitous across the applied mathematical modeling world. Some are developed to just characterize interactions within humans (as reviewed in Talman et al. (2019)), others are developed to be ambiguous (Luo 2014), and many attempt to link within and between host dynamics in host-pathogen systems (Schreiber et al. 2021; Pepin et al. 2017; Park et al. 2013; Hernandez-Vargas et al. 2019; Bansept et al. 2019). Deepening our knowledge about these
within and between host dynamics could provide valuable insight into how pathogens evolve; potentially providing an avenue to probabilistically predict evolutionary trajectories beyond a single scale.

The majority of work surrounding pathogen evolution in population genetics has remained at a single scale. This comes from the assumption that the timescales on which evolution reigns is rapid in comparison to how the population changes. Recent work (Luo 2014) has demonstrated the importance of relaxing this assumption in understanding how adaptation occurs. There have been other models successful in linking multi-scale processes in the past (Wright 1931; Maruyama 1970) and present (Traulsen and Nowak 2006). However, no population genetic models have been explicitly developed to be fully stochastic and to delve into the importance of relative strength of selection at these multiple scales. In Chapter 2 we develop a multiscale Wright-Fisher model that describes how the frequency of a novel pathogen variant changes within and between infected hosts. This is particularly important since pathogen populations undergo dramatic reductions in population sizes during bottleneck events, where genetic drift is known to have a large impact on the trajectory of evolution. In addition, we model how selection acts on the within and between host scales.

On the other hand, models developed using phylodynamics, the study of how immunological, epidemiological, and evolutionary processes impact phylogenetic structure, have had difficulties in bridging the gap across scales for reasons like lack of data availability as well as intrinsic mathematical barriers (Volz et al. 2009; Metcalf et al. 2015; Frost et al. 2015; Gog et al. 2015). In Chapter 4 we describe how multiscale evolution shapes how a continuous trait of interest changes along a phylogeny using a phylodynamic birth-death model. We call this the continuous-type birth-death (CTBD) model which is parameterized such that a continuous trait can change throughout the infection process via selection and genetic drift while simultaneously impacting the transmission dynamics at the host scale.
1.1 Application to HIV

In Chapters 3 and 4 we apply our multiscale model to Human immunodeficiency virus (HIV). Here we will delve into some of the biological and evolutionary processes that motivated this choice.

HIV is the cause of human acquired immunodeficiency syndrome (AIDS). HIV is one of the most damaging human pathogens and is thought to infect approximately 1.7 million people across the globe in 2019 alone, adding to the over 38 million people living with HIV (UNAIDS 2020). Its origin can be traced back to multiple zoonotic spillover events from genetically distinct simian immunodeficiency viruses (SIV) (Hahn et al. 2000). Since then it has diversified into multiple subtypes including HIV-1 subtype B, which will be the focus of the work here since it is the most predominant subtype in Europe and North America.

HIV is a canonical example of a pathogen evolving under multiscale selection pressures. At the within-host scale, HIV undergoes rapid diversification after being transmitted to a susceptible host (Chen et al. 2004; Neher and Leitner 2010; Bazykin et al. 2006). However, mutations at the between host scale accumulate three to six times slower than at the within-host level (Abecasis et al. 2009; Lemey et al. 2006; Pybus and Rambaut 2009; Lemey et al. 2007). Three hypotheses have emerged for why contrasting levels of selection are found in HIV-1 (Lythgoe and Fraser 2012; Raghwani et al. 2018). They are as follows: ‘stage-specific selection’, ‘store and retrieve’, and ‘adapt and revert’ (Pybus and Rambaut 2009; Herbeck et al. 2006, 2011; Redd et al. 2012). ‘Stage-specific selection’ proposes that transmission events occur early in the infection before immune pressures begin to preferentially select for host-specific adaptations. Evidence that HIV diversifies rapidly in the first few months of infection has led to this particular hypothesis falling out of favor (Herbeck et al. 2011; Fischer et al. 2010; Henn et al. 2012). ‘Store and retrieve’ is the concept that founder-like strains remain somewhere in the host and are preferentially transmitted. Finally, ‘adapt and revert’ hypothesizes that the initial rapid diversification of HIV-1 in the infected host reduces population scale transmissibility through the accumulation of host-specific adaptations.
The latter two hypotheses suggest that HIV strains most similar to the population consensus are preferentially transmitted. Some work has suggested that this is indicative of purifying selection at the between host scale in the sense that most variants generated within-host will be deleterious upon transmission to the new host and therefore selectively removed from the viral population (Carlson et al. 2014; Raghwani et al. 2018). We attempt to quantify the strength of purifying selection at the between host level as a "fitness cost to divergence" using the multiscale phylodynamic model we will outline in Chapter 3.

Another trait of HIV that is thought to be under multiscale selection pressures is set point viral load (SPVL). HIV is characterized by a rapid increase in viral particles after initial exposure during the acute or primary infection stage. This stage lasts approximately one to four weeks and hosts may exhibit flu like symptoms for portions of this period. It is during this time that seroconversion begins and the immune system begins producing HIV antibodies and T-cells responses to fight the infection. The acute stage is followed by what is called the asymptomatic, chronic, or latent stage of infection and can last up to ten years. During this stage, viral counts remain relatively stationary at the SPVL measured in units of HIV particles per milliliter (mL) of blood. If left untreated during this latent period, CD4+ T cell population densities will slowly diminish. When they fall below 200 cells per mm$^3$ that individual is classified as having progressed to AIDS.

SPVL is known to be one of if not the largest determinant of transmissibility for HIV-1 with higher viral loads leading to a higher probability of transmission (Operskalski et al. 1997; Pedraza et al. 1999; Quinn et al. 2000; Fideli et al. 2001; Loutfy et al. 2013). This positive relationship between SPVL and transmissibility suggests that higher viral loads provide a between-host viral fitness advantage. This is particularly relevant as there is evidence that suggests SPVL is heritable, measured by the amount of host level SPVL variability that can be explained by pathogen genotype (Alizon et al. 2009; Fraser et al. 2014). Some work has attempted to directly quantify the functional relationship between SPVL and transmissibility in HIV (Fraser et al. 2007; Wilson et al. 2008; Lingappa et al. 2010; Hughes
et al. 2012; Wertheim et al. 2019; Fraser et al. 2014). This can be a difficult relationship to quantify and is often dependent on long term studies of serodiscordant couples, which can be prohibitively expensive. However, our work demonstrates that phylodynamic models can be built to answer similar questions.

When thinking about viral load heritability and impacts on transmission rates one can also think about how viral load, or virulence, will evolve on longer time scales for HIV. For decades now an intense debate has persisted on the evolution of virulence. This debate can be distilled into the following question: when a novel parasitic relationship begins, does evolutionary theory predict virulence to decrease (the avirulence hypothesis) or approach an intermediate viral load (the trade-off hypothesis). Explicitly, the avirulence hypothesis advances the claim that well adapted parasites have evolved to be harmless to their hosts (i.e. low virulence). Alternatively, the trade-off hypothesis suggests that there are costs and benefits to changing virulence (Anderson et al. 1982; Ewald 1983). Increased virulence leads to higher transmission rates while simultaneously decreasing host life expectancy thereby leading to an intermediate value of virulence that is optimized for these factors. More recently, the increased availability of ART and PREP may add additional pressures to viral load evolution making it unclear whether the virulence-transmission trade-off holds (Herbeck et al. 2016). There is recent evidence that viral load may in fact be under strict directional selection to increase since drug treatment, rather than AIDS or death, typically limits the infectious period, however questions remain whether or not this is due to the onset of treatment as prevention programs (Wertheim et al. 2019). All of these hypotheses rest on the relationship between SPVL and transmission fitness which we attempt to quantify using a multiscale phylodynamic model.
1.2 Dissertation Outline

In this dissertation, I develop two mathematical models to understand the evolutionary dynamics of host-pathogen systems. Chapter 2 explores a theoretical population genetics approach to examine how selection at the within and between host scale can shape the evolutionary trajectory of the pathogen. In Chapter 3 we develop an extension of the multi-type birth-death model that quantifies how a continuous-valued trait evolves along phylogeny. We apply this model to data where HIV-1 infected hosts were sequenced and sampled to determine viral load. This allows us to directly estimate the relationship between viral load and between-host transmission fitness for HIV-1. Chapter 4 applies the CTBD to understand how divergence from the population consensus impacts the transmissibility of HIV-1. We analyze this relationship along with the strength of selection within infected hosts across three different data sets from different populations. Finally, Chapter 5 will tie together this work along with presenting some potential future directions of study.
CHAPTER

2

MODELING STOCHASTIC VIRAL EVOLUTION: A MULTISCALE WRIGHT-FISHER MODEL

Abstract

Viral evolution is influenced by both demographic changes during infection at the in-host level and transmission at the between-host level. Here we implement a Wright-Fisher model at both the host and the population level. We analyze how key parameters including bottleneck size, time between transmission events, and selection impact probability of fixation and time to fixation. Both host population and bottleneck size have a role in
determining probability of fixation. As the time between transmission events increases, generations to fixation is practically entirely dependent on host population size. This model provides a quantitative framework to study how population dynamics alter evolutionary dynamics in viruses across scales.

2.1 Introduction

Highly pathogenic infectious diseases have large socio-economic impacts and in turn the study of emerging diseases can lead to substantive change (Molinari et al. 2007). Spillover events, antibiotic resistance, and adaptive shifts are all potential routes for the rise and persistence of infectious diseases. Throughout any pathogen’s history, these routes present a variety of environments with varying levels of selection, population bottlenecks, along with a diversity of population structure at multiple scales (e.g. species, host population, tissue, cellular, etc.). These three mechanisms all interact to alter how evolution proceeds. The process of adaptation itself is driven by novel mutations becoming fixed in the population. This is true for all organisms, however few organisms impact the globe like infectious diseases. For example, although public debate still rages on the matter, SARS-CoV-2’s story certainly began in a reservoir host, thought to be a bat species, where adaptations accumulated through time (Boni et al. 2020; MacLean et al. 2020). Then either directly or through an intermediate host, the pathogen was able to enter and replicate in a human host. Across this timeline, multiple fixation events likely occurred at the host population scale. Here, we utilize a mathematical model that is motivated by such fixation events and we strive to quantify the probability and speed a novel pathogen variant becomes fixed in a host population.

Effectively neutral mutations are found across all evolving viral pathogens (Frost et al. 2018). For example, at the population scale HIV-1 is driven by neutral processes (Grenfell et al. 2004). Influenza A virus is characterized by long intervals of neutral evolution fol-
owed by rapid selection (Wolf et al. 2006). Evidently, strictly employing neutral theory to understand evolutionary dynamics paints an incomplete picture. In particular, although advantageous mutations are rare, their high fitness remain an important feature of pathogen evolution (Fyre-Walker and Keightley 2007; Sanjuán et al. 2004). Single amino acid substitutions can confer large structural changes and in turn can lead to phenotypic variation. A non-synonymous mutation that increases the pathogen’s ability to replicate within a host, or transmit between hosts, is of public health concern and examples have been found across multiple pathogens (Rasmussen and Stadler 2019). In SARS-CoV-2 multiple non-synonymous mutations including D614G (Korber et al. 2020), and more recently a group of changes found in the B.1.1.7 (Rambaut et al. 2020), B.1.351 (Tegally et al. 2020), and B.1.617.2 variants (deemed the Alpha, Beta, and Delta variants respectively) have gained global interest due to their rise in frequency (Hodcroft 2021). In addition to these novel variants, theoretical experiments using deep mutational scanning methods have found a wide array of potential amino acid substitutions that can increase binding affinity to cell receptors and others that may reduce the neutralizing effects of naturally acquired antibodies (Starr et al. 2021, 2020; Greaney et al. 2020). Clearly understanding how advantageous mutations spread in populations is vital information.

The spread of novel variants can also be driven by genetic drift. During transmission events, intense bottlenecks drastically reduce the founding population size by orders of magnitude. Small populations provide a landscape on which stochastic forces drive evolutionary dynamics. These bottleneck events are ubiquitous across emerging pathogens. For example, HIV-1 is known to have an extremely small bottleneck size with 80% of infections established with a single virion (Joseph et al. 2015). The bottleneck size of influenza A has had a wider distribution of bottleneck estimates with some finding similarly stringent bottlenecks of 1-2 variants (McCrone et al. 2018) and others finding looser bottlenecks with approximately 200 virions being transmitted (Sobel Leonard et al. 2017).

Another important contributor to evolutionary dynamics is the structure of the host
population (Gordo and Campos 2007; Tkadlec et al. 2019; Maruyama 1970). Previous work in theoretical population genetics has explored this factor. For example, extensive population structure is known to increase the probability that deleterious mutations become fixed in the population (Whitlock 2003). Later, Gordo and Campos (2007) show how a compartmental model overlayed on a network can present varying levels of diversity depending on host immune system clearance rates and within-host effective population. Most recently, Tkadlec et al. (2018) explore how contact network structure can dramatically change probability of fixation and generations to fixation. None of these however have explicitly modeled multiscale evolution in a population of infected hosts.

Historically, theoretical work on adaptation has used derivations for the probability of fixation along with the time to fixation of a novel allele to quantify the likelihood and speed of adaptation. The two most analyzed models are the Wright-Fisher and Moran model and the analytical approaches developed fall under either diffusion or branching process theory. A review of these techniques and their application to quantifying the probability of fixation for beneficial mutations can be found in (Patwa and Wahl 2008). Here we simulate pathogen evolution using a multiscale Wright-Fisher model and employ both diffusion and branching process theory. We expand upon previous theoretical work on adaptation by quantifying how a novel mutation in a viral pathogen population can spread through a host structured population while undergoing transmission bottlenecks as well as selection acting at multiple scales. Although previous work has looked at selection pressures and structured populations, to our knowledge none have incorporated both multiple scales of selection on a structured population and transmission bottlenecks.

2.2 Methods

We implement a Wright-Fisher model (a discrete time Markov Model) at both scales to quantify how the frequency of a novel mutation changes throughout the infection and
transmission processes. At the between host scale, we implement a Wright-Fisher model with a constant population of infected hosts \( (N_b) \) that can be interpreted in one of two ways. Either the infected host population is in endemic equilibrium or that \( N_b \) is the effective population size of infected hosts in a fluctuating population. This assumptions dramatically simplifies our analytical approximations, but could potentially be relaxed in future work (Ewens 1967; Kimura and Ohta 1974; Otto and Whitlock 1997).

**Within Host Models**

To reflect the exponential growth of a viral pathogen within a novel host post-introduction, we consider two discrete stages or epochs. In the epoch driven model, the first stage is characterized by a Wright-Fisher model which is used to reflect the initial stochasticity from a transmission bottleneck of size \( N_w \). Since we are considering a host-pathogen system, we assume that the host population is orders of magnitude larger than the pathogen bottleneck size (i.e. \( N_b >> N_w \)). The second stage begins once the viral load has reached its carrying capacity. It is during this stage that mutant allele dynamics are deterministic. By the Hardy-Weinberg principle, if this mutation is neutral its frequency will remain constant throughout this phase. This remains true as long as there are no mutations or gene flow in the population, assumptions we make throughout this paper. We use a logistic growth model to capture the dynamics of the mutant frequency. The stochastic and deterministic phases will occur for a specific number of non-overlapping within-host generations, deemed \( g_s \) and \( g_d \) respectively.
Figure 2.1: Generalized Wright-Fisher model where $f_{i,j}$ represents the frequency of the $i$th host in the $j$th generation. Within each host, genetic drift is at play for $g_s$ generations.

Here we analyze two metrics for evolution: probability of fixation and the expected number of generations to fixation for a novel mutant viral pathogen. These metrics provide an understanding of the probability and speed at which evolution occur. Previous work has developed these metrics for the Wright-Fisher model under various assumptions. Quickly summarized, in a constant population of haploid individuals of size $n$ under which a single neutral mutant is present at frequency $p$, the probability of fixation and generations to
fixation have been approximated to be Equations 2.1 and 2.2 respectively.

\[ \pi = \frac{1}{n} \]  
\[ \bar{t}_1 = \left[ -\frac{1}{p(2n(1 - p) \log(1 - p))} \right] \]

For novel mutations this frequency \( p \) is assumed to be \( p = \frac{1}{n} \). Both of these are derived from the diffusion approximation to the Wright-Fisher model, however additional types of analytical approximations have been derived for this model. One of particular interest is the branching process approximation. This method has been shown to be most accurate under strong selection regimes and we will consider it as we relax the assumption that the mutation is neutral. It is important to note that this system can be reduced to the original Wright-Fisher model if either population size (\( N_b \) or \( N_w \)) is reduced to one.

There are a few parameters of interest that drive dynamics in the system: host population size (\( N_b \)), within-host population size (\( N_w \)), stochastic and deterministic within-host generations (\( g_s \) and \( g_d \)), and selection coefficients acting on each scale (\( s_b \) and \( s_w \)). These along with the metrics mentioned above are summarized in Table 2.1. Our goal is to understand how these parameters qualitatively and quantitatively alter the dynamics of the system.
Table 2.1: Descriptions for parameters and metrics utilized in the paper.

<table>
<thead>
<tr>
<th>Parameter/Metric</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_b$</td>
<td>Host population size</td>
</tr>
<tr>
<td>$N_w$</td>
<td>Pathogen bottleneck size</td>
</tr>
<tr>
<td>$g_s$</td>
<td>Stochastic pathogen generations</td>
</tr>
<tr>
<td>$g_d$</td>
<td>Deterministic pathogen generations</td>
</tr>
<tr>
<td>$s_w$</td>
<td>Within-host selection coefficient</td>
</tr>
<tr>
<td>$s_b$</td>
<td>Max host level selection coefficient</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Probability of fixation at host/population level</td>
</tr>
<tr>
<td>$\bar{t}_1$</td>
<td>Generations to fixation at host/population level</td>
</tr>
</tbody>
</table>

2.2.1 Host Level Selection types

Understanding how within-host dynamics impact transmission at the host level is an essential part of multi-scale models. A novel pathogen variant may lead to its host having a higher probability of transmitting compared to a host infected entirely with the wild-type pathogen. It is quite possible that this relationship is complicated, with the increase in transmission being dependent on the frequency of the novel variant in the within-host population. In particular, we hope to model how between host selection is a function of coinfection. Previous work has demonstrated how the overall transmission rate of a co-infected host can be modeled as a reflection of the dominance of the co-infecting strains (Bushman and Antia 2019). Here we take a similar approach in laying out three potential scenarios for the relationship between co-infection and host level transmission, all reflective of the dominance of the co-infecting strains. These three scenarios will be referred to as presence dependent, proportional dependent, and fully mutant selection benefits. Before we precisely define how selection benefits occur at the host level recall that $s_b \in \mathbb{R}$ is the maximum fitness advantage of the novel mutation at the host level. Let $s_b$...
be the realized selection benefit of the mutant virus at the host level and $p \in [0, 1]$ be the proportion of a particular host’s pathogen population that is mutant. Presence dependent selection benefits will ensure any host containing a nonzero frequency of mutant ($p > 0$) receive the full fitness advantage $\hat{s}_b = s_b$. This scenario likens itself to one where the mutant strain is dominant over the wild-type strain. Frequency dependent benefits will be linearly dependent on the frequency of mutant present in the host (i.e. $\hat{s}_b = ps_b$). This is an analogue of incomplete dominance. The last scenario is one where the wild-type strain is dominant over the mutant strain and in turn hosts will only receive the benefits of the mutant allele when it is fixed within the host ($p = 1$). These three scenarios (presented in Figure 2.7) are not meant to encapsulate every possible relationship, but instead demonstrate three caricatures of this relationship that may provide insight into the multiscale dynamics.

Utilizing multi-type branching process theory we can obtain the probability of emergence as an approximation to the probability of fixation (Antia et al. 2003). This is done by defining the dynamics of the host population with a series of probability generating functions for each type of host, characterized by the number of mutant pathogens present at the time of transmission. They can be written as follows:

$$f_i(r_1, ..., r_{N_w}) = \exp[-B_{i}w_i(1-r_1) - \cdots - B_{N_w}w_N(1-r_{N_w})], \quad i = 0, ..., N_w - 1$$

$$f_{N_w}(r_1, ..., r_{N_w}) = \exp[-w_{N_w}(1-r_{N_w})]$$

where $B_{ij} = \binom{N_w}{i} \binom{N_w-j}{N_w-i} w^i$ and $w^i = 1 + ps_b$ is the Wrightian fitness when $p = i/N_w$. Recall that Wrightian fitness is the average number of offspring. From these series of probability generating functions we can numerically solve the system $f_i(r_1, ..., r_{N_w}) = r_i$ simultaneously and take the probability of fixation to be $\pi = 1 - r_1$, or equivalently the probability an extinction event does not occur.

We will compare these results to what two sequential Wright-Fisher models would produce, one at the within-host level and then one at the host level. This will assume that
the within-host and host level Wright-Fisher models are independent of one another. These assumptions lead to two possible scenarios; selection benefits are present at both levels, or selection only being present at the host level. For the sake of this paper we do not explore deleterious mutations.

For some population size $n$ and selection coefficient $s$, the generations to fixation in the original Wright-Fisher model can be quantified using the diffusion approximation (Kimura and Ohta 1969). This is defined as,

$$ t_{fix}(n, s) = J_1 + \frac{1 - \pi}{\pi} J_2, \quad (2.3) $$

$$ J_1 = \frac{2}{s(1 - \exp(-2ns))} \int_{1/n}^{1} \frac{(\exp[2ns\zeta] - 1)(\exp[2ns\zeta] - \exp[2ns])}{\zeta(1 - \zeta)} d\zeta $$

$$ J_2 = \frac{2}{s(1 - \exp(-2ns))} \int_{0}^{1/n} \frac{(\exp[2ns\zeta] - 1)(1 - \exp[2ns\zeta])}{\zeta(1 - \zeta)} d\zeta $$

Note that $\lim_{s \to 0} t_{fix}(n, s) = 2n$.

### 2.2.2 Deterministic Addition

Although we are now capable of analytically approximating the stochastic setting, we are continuing to assume that the population size at the within-host level remains small. This is clearly not telling the entire story and ignores any population growth that occurs post infection. When population sizes are large, the effects of genetic drift become outweighed by selection. Therefore, we can see that at the within-host level there are times where either drift or selection are stronger. For this reason we have separated the within-host generation time into a stochastic phase and a deterministic phase. During deterministic generations we model the frequency of the mutant virus using the logistic equation below,

$$ p(p_0, s_w, g_d) = p_0 e^{(s_w g_d)/(1 + p_0(e^{(s_w g_d)-1}))} $$
where $s_w$ is the selection coefficient within the hosts, $g_d$ is the number of deterministic generations and $p_0$ is the frequency of the mutant virus immediately after the stochastic phase.

### 2.3 Results

The result section below is organized by which scale a novel mutation confers a fitness advantage. In each subsection we present analytical expressions for the probability of fixation and expected number of generations to fixation. These expressions are some combination of branching process and diffusion theory depending on the strength of selection pressures. This is since diffusion and branching processes are known to better approximate weak and strong selection pressures respectively. We first outline our approximations for a novel neutral mutation (Section 2.3.1), then our approximations for selectively advantageous mutations at the host and within-host levels follow (Sections 2.3.2 and 2.3.3), and this is concluded by mutations that are advantageous at both scales (Section 2.3.4).

#### 2.3.1 Neutral Mutation

Here we develop an argument for how the probability of fixation and the expected number of generations to fixation will be approximated for the case where a mutation is neutral at both the between-host and within-host levels. The logic presented in this section will be extrapolated to later sections where the novel mutation is advantageous at one or both scales. In host-pathogen systems of interest, infections are characterized by stringent bottlenecks across large numbers of infected hosts. This leads to the host population size being orders of magnitude larger than the initial viral population size in a newly infected host. Under these assumptions, there are two potential avenues under which fixation can occur depicted in Figure 2.2 that is driven by whether the most recent common ancestor (MRCA) is fully or semi-mutant. Under the latter scenario multiple hosts must become fixed
independently of one another while the former scenario requires that single host becomes fixed after which all hosts infected with the mutant pathogen will be fully mutant. This process can lead to drastic shifts in the frequency of mutant allele in the host population.

We propose that the proliferation of multiple semi-mutant hosts occurs, but fixation events with the MRCA being fully mutant are more likely to occur. In turn we can consider the process as two sequential events, fixation of the mutant pathogen in a single host followed by fixation at the population level. This scenario is presented in Figure 2.2A. Once this host is fully mutant, we should consider the probability that it will lead to a population level fixation event. Both of these fixation events can be approximated by a classic Wright-Fisher model under their respective population parameters. In particular, these probabilities are $\frac{1}{N_w}$ and $\frac{1}{N_b}$ respectively under the assumption that the mutation is neutral. Assuming that these two events are independent, the probability of fixation for a novel mutation is then $\frac{1}{N_b N_w}$. The accuracy of this analytical approximation has been supported by simulation work (Figure 2.5.1) as well as first step analysis (2.5.1). These results are independent of the number of stochastic ($g_s$) or deterministic ($g_d$) within-host generations (Figure S1).
Figure 2.2: Host representation of multiscale Wright Fisher model. Darker shades of blue correspond to higher frequency of mutant allele. The red circle distinguishes the MRCA of the hosts from the fixation event (last row). (a) Resulting outcome of fixation event where MRCA is fully mutant. (b) Resulting outcome of fixation event where MRCA is semi mutant.

In a well-mixed population of virions of equivalent size \( (N_b, N_w) \), the probability of fixation would be precisely what was derived for our multiscale system, \( \frac{1}{N_b N_w} \). However, the expected number of generations to fixation for this same well-mixed population would be \( \bar{t}_f \approx 2N_e = 2N_b N_w \). However, this is not the case; instead we find that the generations to fixation is dramatically lower. In particular, both host population size and transmission bottleneck size have approximately linear relationships with the number of generations to fixation. This can be understood using the same logic used above to approximate the probability of fixation for a novel mutation. In particular, if we assume that the majority of
fixation events occur via a fully mutant MRCA then the expected number of generations to fixation can be split up into a series of sequential fixation events, one at each scale, i.e. fixation of the mutant allele within a single host followed by fixation at the host population level. For a given population of size $n$, the expected number of generations to fixation for a neutral mutation is $\bar{t}_1 = \left[-1/p(2n(1-p)\log(1-p))\right]$ when $p = 1/n$. As $n \gg 0$ this becomes approximately $2n$ (Kimura and Ohta 1969). Although this latter approximation assumes large population size that is inherently unfulfilled by the model during pathogen bottlenecks, we will use it for simplicity.

Viral populations grow on a shorter time scales which permits the assumption that number of within-host and between host generations are not added to the expected time to fixation $\bar{t}_1$. Instead, time to fixation is a measure of host level generations. Since the mutant frequency remains constant during deterministic generations, $g_d$ has no impact on the expected time to fixation. However, the addition of within-host stochastic events will only speed up this process as the mutant virus has the chance to drift towards a higher mutant frequency. Therefore, the expected generations to fixation within one host is $2N_w/g_s$ generations. Once a single host is fully mutant, the expected number of generations to fixation at the host level is approximately $2N_b$. It follows that an approximation for generations to fixation under the assumption that fixation events occur sequentially takes the form $\bar{t}_1 \approx 2N_b + 2N_w/g_s$. This approximation overestimates results from simulation not fare as well as the probability of fixation and in fact is consistently an overestimate (Figure 2.3). This approximation is known to be an overestimate under a single Wright-Fisher model, suggesting the multilevel framework will only magnify this (Kimura and Ohta 1969).

This is due to the fact that not all fixation events have a fully mutant MRCA where fixation events at each scale occur sequentially. Instead, some portion of fixation events will occur where the MRCA is semi-mutant (Figure 2.2a). In these events, the MRCA occurs before $t = 2N_w$ which in turn leads to faster than expected fixation events. However, there is no good approximation of the time to fixation for this type of fixation event. Therefore,
we present our approximation to the expected number of generations to fixation as an upper bound, \( \bar{t}_1 < 2N_b + 2N_w/g \). This approximation will increase in accuracy in parameter regimes where a higher proportion of MRCA are fully mutant. Examples of these parameter regimes are when either \( N_b \) or \( N_w \) is close to one and when \( g_s >> 1 \) as seen in Figure 2.3b. The former is due to the fact that when either \( N_b \) or \( N_w \) is one the model returns to a single Wright-Fisher model. This pattern is presented in Figure 2.8 where approximations are best when either \( N_b \) or \( N_w \) is small.

Figure 2.3: (a) Impact of varying levels of within \((N_w)\) and between \((N_b)\) host population size on probability of fixation. (b) Impact of varying levels of within-host population size \((N_w)\) and stochastic generations \((g_s)\) on the expected number of generations to fixation. Simulated assuming \( N_b = 100, g_d = s_b = s_w = 0 \) and \( g_s = 1 \) unless otherwise specified.

2.3.2 Host Level Selection

In this section we analyze the impact of a mutation that only confers a positive fitness advantage at the host level \((s_b > 0)\) while maintaining neutral at the within-host level \( s_w = 0 \). Under these assumptions, co-infected hosts may transmit more frequently than their fully wild type counterparts. It follows that as \( s_b \) increases, the majority of fixation events will occur with a semi-mutant MRCA (visualized in Figure 2.2). We analyze how the three host
level selection scenarios outlined above in Section 2.2.1 change the probability of fixation and the expected number of generations to fixation.

Under nearly neutral positive selection (Figure 2.4b), the diffusion approximation best explains the probability of fixation for all three scenarios. This diffusion approximation is:

\[ \pi = \left( \frac{1 - e^{-2s_b}}{1 - e^{-2N_b s_b}} \right) \left( \frac{1}{N_w} \right). \]  

(2.4)

Under strong selection (Figure 2.4a), the probability of fixation is heavily influenced by the host level selection scenario at play and is approximated well by the multi-type branching process. Note that the presence dependent selection scenario approaches unity while the fully mutant dependent scenario asymptotes at the probability that one host becomes fully mutant in the host population.

The expected number of generations to fixation for all three mechanisms is approximated by the sum of two diffusion approximations (Figure 2.5), one at each scale, as found above in Equation 2.4. Host population size \( N_b \) is expected to be orders of magnitude larger than the pathogen bottleneck size \( N_w \) and thus selection at the host scale leads to a sharp decrease in the expected number of generations to fixation (Figure 2.5a).
Figure 2.4: Probability of fixation under varying host level selection coefficients ($s_b$). Strong selection (a) and weak selection (b) are depicted. Points represent simulation results, solid lines correspond to numerical results from the branching process model and the dotted line is an analytical approximation to fixation for two sequential Wright-Fisher models. Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified.

Figure 2.5: Effect of host level selection on generations to fixation for weak (left) and strong selection (right). Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified.
2.3.3 Within-Host Level Selection

Here we analyze when the selection advantage of the mutant is present at the within-host level. For these events, we implement two strategies to approximate the probability of fixation: one which takes the product of the two diffusion approximations found in the literature and the other which multiplies the diffusion approximation for the level where the mutation is neutral and a branching process approximation for the level where the mutation is advantageous (Figure 2.6). Branching processes approximate the probability of fixation for a novel mutation in the Wright-Fisher framework when supercritical \( s_i > 0 \) for \( i \in w, h \). When a mutation is advantageous strictly at the within-host level, the probability of fixation does not depend on the number of stochastic generations \( g_s \). It is however, sensitive to the number of deterministic generations \( g_d \). Figure 2.6 demonstrates this is particularly true when the mutation is weak. As mentioned in Section 2.3.2, the two approximation techniques hold better under different scenarios. The diffusion approximation is supported by simulations under weak beneficial mutations whereas the branching process is supported under strong beneficial mutations. Note that both simulations and approximations approach \( 1/N_b \) as \( N_{sw} \) grows large (i.e. the probability of fixation at the host level for a neutral mutation).

The expected number of generations to fixation under varying deterministic and stochastic within host generations is approximated by the sum of two diffusion approximations (Figure 2.9), one at each scale, as found above in Equation 2.4. These simulations assume that \( N_w = 10 \) while \( N_b = 1000 \). It is for this reason that we do not see a large decrease in the number of generations to fixation across these varying levels of within-host selection. If the within-host population size was significantly larger than the host population size \( (N_b >> N_w) \) we would see similar qualitative patterns found in Figure 2.5a.
2.3.4 Selection at multiple scales

When selection is weak at both scales ($0 < N_{s_w}, N_{b_s} < 1$) we approximate the probability of fixation with the product of two diffusion processes, one at each scale:

$$\pi \approx \left( \frac{1 - e^{-2s_w}}{1 - e^{-2N_{w}s_w}} \right) \left( \frac{1 - e^{-2s_b}}{1 - e^{-2N_{b}s_b}} \right).$$  (2.5)

This assumes the independence of fixation at the two scales, an incorrect assumption. However, Figure 2.10b demonstrates that this assumption holds for varying levels of weak within-host selection pressures in conjunction with weak between-host selection.

The probability of fixation can be approximated by the product of two branching processes when the mutation is strong at both scales ($N_{s_w}, N_{b_s} > 1$). This approximation is accurate for all types of host selection as demonstrated in Figure 2.10a.
2.4 Discussion

Quantifying the probability that a mutation becomes prevalent and the speed at which it does so are two important metrics to understand viral evolution. We analyze the probability and expected number of generations to fixation of a novel mutation under a multi-scale Wright-Fisher model.

The expected number of generations to fixation is driven by the host population size and in turn selection pressures acting at this scale have the largest impact on this metric. This result is dependent on the fact that the host population size is orders of magnitude larger than the effective population size of the pathogen population. The expected number of generations to fixation for a novel mutation in this host-pathogen framework is significantly lower than one found in a panmictic population of equivalent size. This was demonstrated in Tkadlec et al. (2019) and is due to the compartmentalization of the viral population into discrete hosts. Initial small population sizes after bottleneck events provides time for the mutant frequency to randomly increase and in turn have a higher likelihood of being transmitted.

Here we assume that the host population size is constant, modeling either an effective epidemic population size, or a population where the pathogen is endemic. This is a strong assumption, but gives the general rule that the expected number of generations to fixation for a neutral mutation will be linearly dependent on the host population size. As the number of stochastic generations after a bottleneck event increase, the expected time to fixation decreases. This pattern becomes more apparent as the effective population size of the pathogen increases.

Under the assumption of a neutral mutation (i.e. genetic drift), we find that the probability of fixation is dependent on the product of the host population size and the viral bottleneck size. This product is equivalent to the probability of fixation for a novel allele in a panmictic population of equivalent size, a result found previously by Maruyama (1970). This stresses the importance of quantifying viral bottleneck size and in turn the effective
population size of a pathogen across the infection process.

The probability of fixation varies greatly when a mutation confers selection advantages at either or both the within and between-host levels. Under strong host level selection, the probability of fixation becomes a function of the selection scenario. The manner in which it is altered varies with respect to the three previously outlined selection scenarios; presence, proportional, and full mutant dependent. These scenarios present scenarios of dominance between the wild-type and mutant strain. Unsurprisingly, a selectively advantageous mutant strain that is dominant has the highest probability of fixation analyzed here. Additional dominance scenarios like under and over dominance could be analyzed in future work. Quantifying how selection benefits cross between scales in particular pathogens could be an insightful direction for future research.

At the within-host level we find that the number of deterministic generations has a large impact on the probability of fixation and leads to an equivalently high probability of fixation as their strongly selected for counterparts. Pathogen strains that are best adapted for their hosts will out-compete others in the environment and as the number of generations increases this pattern only magnifies. This pattern could be potentially reversed under antagonistic selection (where the mutation is deleterious at the host-level). This question has been explored in theoretical HIV-1 models and quantifying precisely how probability of fixation and expected number of generations to fixation vary could be insightful into evolutionary dynamics of pathogens (Lythgoe et al. 2013).
2.5 Supplemental

Figure 2.7: Three exemplary host level selection paradigms. Presence, proportion, and fully mutant dependent.

2.5.1 First Step Analysis

The above logic is consistent with a first step analysis approach to this same question. First step analysis is a technique used to quantify various attributes of a Markov Chain like average number of times being in a particular state, mean absorption time, etc. Here we use it to quantify the probability of hitting fixation at the population level. What follows here is a short explanation of this result. Consider a population that consists of one host, co-infected with both mutant and wild-type viruses. Let $P_k$ be the probability that the
mutant viruses will become fixed in the population at some point if $k$ is the initial number of mutant viruses present in the host where $0 \leq k \leq N_w$. It is clear that $P_0 = 0$ since this defines the entire host population as containing solely wild-type viruses. When $k = N_w$ all hosts will either consist entirely of the mutant or the wild-type as we do not consider the possibility of mutation. Therefore, this means the system would be equivalent to the classic Wright-Fisher model where past work (Kimura 1968) has utilized diffusion approximations to show that $P_N = 1/N_b$. In order to determine the probability of fixation for any other type of initially co-infected host type we use first step analysis. Let $y_{ij}$ be the rate a host with $j$ mutants passes on $i$ mutants to their offspring. Therefore, $P_i y_{ij}$ is equivalent to the probability that the $i$ mutants passed on from a host with $j$ mutants will end up becoming fixed in the entire population. If we consider every possible type of offspring a host with $j$ mutants can have and sum up all these probabilities then we can find the total probability that they will become fixed in the population. Therefore,

$$P_k = \sum_{i=1}^{N_w} P_i y_{ik}$$

When we consider this system of equations under the assumption that each $y_{ij}$ is governed by the binomial distribution then we can solve for each $P_k$. The binomial distribution makes analytical solutions very elusive. However, we solved this system numerically to find that

$$P_k = \frac{k}{N_b N_w}$$

As mentioned in Section 2.3.1, Figure 2.3a verifies the validity of the first step analysis approach.
Figure 2.8: Percent difference between numerical simulations and analytical solution for generations to fixation across the host and within-host population size parameter regimes under a neutral model.
Figure 2.9: Average number of generations to fixation under varying within-host level selection coefficients ($s_w$). This is paired with (a) varying levels of within-host stochastic generations, $g_s$, and (b) varying levels of within-host stochastic generations. The solid and dashed lines correspond to analytical results from diffusion and branching process theory respectively, while points represent simulation results. Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_b = 0$ and $g_s = 1$ unless otherwise specified.

Figure 2.10: Probability of fixation when both host and within-host selection coefficients are positive. Strong selection (a) and weak selection (b) for both scales are depicted. Points represent simulation results, solid lines correspond to numerical results from the branching process model and the dotted line is an analytical approximation to fixation for two sequential Wright-Fisher models. Simulated assuming $N_b = 1000$, $N_w = 10$, $g_d = s_w = 0$ and $g_s = 1$ unless otherwise specified.
In host-pathogen systems, pathogen strains can have varying levels of fitness at the within host and between host scales. This can arise through changes in the ability to compete at the within host scale or through variation in the host’s ability to transmit. Here we develop a phylodynamic birth-death model that allows selection to act on a continuous-valued trait both at the within and between hosts scales. We use this method to infer the
magnitude and direction of selection pressures within and between hosts along with other relevant parameters like mutation rates and effective population size from phylogenetic trees. Each pathogen lineage is assigned a continuous “trait” value that is an element of some closed and bounded set. In the context of a host-pathogen system, these continuous traits could represent a pathogen’s viral load, allele (strain) frequency or a measure of genetic diversity of divergence. Given a phylogeny with continuous-valued tip states, we can perform Bayesian inference utilizing our continuous type birth-death branching model. We apply this approach to quantify the relationship between viral load and between-host transmission rates for HIV-1.

### 3.1 Introduction

Viral phylogenies are known to contain considerable information about underlying evolutionary and epidemiological processes. The field of phylodynamics attempts to understand the forces that have led to the epidemiological and phylogenetic patterns found in the data (Grenfell et al. 2004; Volz et al. 2013). With increasing computational power and the onset of next-gen sequencing, these models have provided vital insights into viral population dynamics.

One recent advance has been the development and implementation of multi-type birth-death (MTBD) models (Maddison et al. 2007; FitzJohn 2012; Kühnert et al. 2016; Stadler 2013; Barido-Sottani et al. 2020). These models assume sampled lineages can be categorized into discrete types which can give rise to novel infections, be removed from the infectious class, transition amongst these discrete types, and be sampled at type-specific rates. A discrete type could be a specific host trait, like being a part of a known risk group, or a trait of the pathogen itself, like the presence or absence of a particular mutation of interest. The parameterization of the MTBD allows for the relative fitness, or transmission potential, of the sampled types to be determined based off of the type-specific birth and death rates
(Kühnert et al. 2018). Types that are more fit will on average produce more offspring while less fit types are more likely to be removed before leading to any secondary infections. Individuals removed from the infectious class are sampled and the phylogenetic tree is reconstructed from their sequence data on which the phylodynamic model is computed. The type specific rates of birth, death, and transition can be estimated using a likelihood based approach.

While MTBD models are very useful for modeling the evolution of some pathogen traits, other traits (e.g. viral load, allele frequencies) are modeled more naturally as continuous-valued traits. Historically, comparative phylogenetic approaches have been used to study how both discrete and continuous valued traits can undergo evolutionary adaptation on phylogenies (Felsenstein 1985; Blomberg et al. 2003; Garland Jr et al. 2005). These models often assume Brownian motion governs the evolution of the trait with phylogenetic relatedness being used to inform the speed at which the trait evolves (Cornwell and Nakagawa 2017). The field has grown considerably in the past few decades, with more intricate models of trait evolution allowing for deeper understanding of evolutionary history. However, these methods do not allow the trait to shape the branching process generating the phylogeny. This is where the phylodynamic models, like the CTBD, can step in.

By extending the theory of MTBD models to continuous traits, we demonstrate that genetic data can be paired with a continuous trait of interest and we can quantify the fitness of lineages with any particular trait value in continuous space. We call this the Continuous Type Birth-Death (CTBD) model. The trait need not be continuous in nature, but instead could be an ordered trait that has a considerably large range, like the number of mutations found in a genome compared to a population consensus. We apply the CTBD model to viral load data in HIV-1 infected hosts from whom we also have genetic data from the Protease and Reverse Transcriptase genes (PR/RT). We assume that the sampled viral loads are good estimates for host set point viral load (SPVL), the stable viral load count during the latent phase of HIV-1 infection. SPVL is a heritable phenotype in that it is partially
under viral genetic control and therefore transmittable between hosts (Alizon et al. 2010). The heritability of this trait suggests that it can be driven by evolutionary processes and potentially selection at multiple scales. At the same time, SPVL is thought to be one of the major contributors to transmission rates in HIV-1 (Operskalski et al. 1997; Pedraza et al. 1999; Quinn et al. 2000; Fidel et al. 2001; Loutfy et al. 2013). At the between-host level it is understood that higher SPVL leads to increased transmission (Fraser et al. 2007; Wilson et al. 2008; Lingappa et al. 2010; Hughes et al. 2012; Wertheim et al. 2019), but the exact functional relationship is unknown (Blaser et al. 2014). However, high SPVL is also associated with faster transition to AIDS, potentially leading to intermediate viral loads maximizing transmission (Alizon et al. 2009; Fraser et al. 2007; Blanquart et al. 2016) Here we quantify the relationship and demonstrate the utility of the CTBD for questions regarding the evolutionary dynamics of continuous traits in viral populations.

3.2 Methods

In this section a continuous type birth-death model (CTBD) is developed which is an extension of the multi-type birth-death model (MTBD). We begin with an explanation of the MTBD (Section 3.2.1) in which the probability of some discrete trait (sex, allele, transmission group, sub-type, etc.) is assigned to each lineage, starting from the sampled tips. This is followed by a derivation of the CTBD model (Section 3.2.2), where a continuous valued trait (viral load, allele frequency, distance from consensus sequence, etc.) is assigned to each sampled tip.

3.2.1 Multi-type Birth-Death Model

The birth-death model was first conceived by Yule in 1924 assuming a constant death rate of zero and a constant birth rate (Yule 1925). This work has been further advanced through the years by various statisticians and mathematicians. Kendall derived the first
and second moments for the time varying birth-death model (Kendall 1948). Since then, there have been various developments made to accommodate varying types of sampling strategies like complete sampling (Rannala and Yang 1996), sampling of only extant individuals (Thompson et al. 1975; Gernhard 2008; Stadler 2009), serial sampling (Stadler 2010), and incomplete sampling (Yang and Rannala 1997; Stadler 2009). These, along with many previously derived birth-death model variants are special cases of a more general birth-death-sampling process (MacPherson et al. 2021). Stadler and Bonhoeffer (2013) developed a maximum-likelihood based framework for the MTBD branching process under incomplete sampling for lineages sampled serially through time. In this work, the ordinary differential equations for two-states under complete sampling derived in Maddison et al. (2007) were extended for an arbitrary number of states and incomplete sampling. This model is referred to as the MTBD-$m$ or MTBD model.

The MTBD-$m$ process, where $m$ is the number of states, has the following parameters:

$$\lambda = (\lambda_{1,1}, \ldots, \lambda_{1,m}, \lambda_{2,1}, \ldots, \lambda_{2,m}, \ldots, \lambda_{m,1}, \ldots, \lambda_{m,m}),$$

$$\gamma = (\gamma_{1,1}, \ldots, \gamma_{1,m}, \gamma_{2,1}, \ldots, \gamma_{2,m}, \ldots, \gamma_{m,1}, \ldots, \gamma_{m,m}),$$

$$d = (d_1, d_2, \ldots, d_m),$$

$$s = (s_1, s_2, \ldots, s_m),$$

$$t_0.$$

The first set of parameters $\lambda$, defines the birth or transmission rates among and between the types. Therefore, $\lambda_{i,j}$ is the rate that an individual of type $i$ gives birth to an individual of type $j$. For example, in Stadler and Bonhoeffer (2013), the authors apply the MTBD-2 to a population where infected hosts are classified into two types, superspreaders and normal spreaders and estimated that the rate of birth/transmission from superspreaders to normal spreaders to be approximately nine times higher than from normal spreaders. More generally, these models allow for precise quantification of transmission rates within and
between different types of hosts, where the type of a host can either reflect a property of the host (acute/chronic) or a property of the infecting pathogen (drug resistant/susceptible).

The rate of transitions amongst the types, or additional state changes not attributed to transmission events, is defined by the $\gamma$ matrix. Here the rate $\gamma_{i,j}$ is the rate an individual of type $i$ transitions into an individual of type $j$. An example of this could be an individual migrating from one country to another or even the pathogen itself mutating from one genotype to another. The rate of death $d$ is defined on a per type basis although these parameters are often assumed to be constant across types (i.e. $d_i = d_j$ for all $i, j \in \{1, 2, ..., m\}$). When an individual of type $i$ dies, they are then sampled at some type specific probability $s_i$. Finally, $t_0$ defines the total tree height at which the birth-death process is assumed to begin.

To perform likelihood-based inference under the MTBD-$m$ model, Stadler and Bonhoeffer (2013) derived a system of ordinary differential equations describing the change in the probability density $D_{N_i}(t)$ (Equation 3.1) that lineage $N$ evolved exactly as observed given that it is in state $i$ at time $t$ and $E_i(t)$ (Equation 3.2) as the change in probability that an individual in state $i$ is not sampled and has no sampled descendants after time $t$.

$$\frac{d}{dt} D_{N_i}(t) = -\left( \sum_{j=1}^{m} (\lambda_{i,j} + \gamma_{i,j}) + d_i \right) D_{N_i}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{N_i}(t) + \sum_{j=1}^{m} \gamma_{i,j} D_{N_j}(t)$$

(3.1)

$$\frac{d}{dt} E_i(t) = \left( 1 - s_i \right) d_i - \left( \sum_{j=1}^{m} (\lambda_{i,j} + \gamma_{i,j}) + d_i \right) E_i(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) E_i(t) + \sum_{j=1}^{m} \gamma_{i,j} E_j(t)$$

(3.2)

For some phylogenetic tree $\mathcal{T}$ of interest, the likelihood that it evolved under the MTBD-$m$ model for some combination of parameters $\theta = (\lambda, \gamma, d, s, t_0)$ is defined as $\mathcal{P}(\mathcal{T}|\theta)$. The likelihood of the tree can be computed as follows. First, $D_{N_i}(0)$ is initialized for each tip. If the state of a tip is known, initialize $D_{N_i}(0) = s_i d_i$ for the true state $i$ and $D_{N_j}(0) = 0$ for $j \neq i$. 
This follows from the assumption that sampling occurs with probability $s_i$ immediately after an individual dies or becomes non-infectious at rate $d_i$. If the true state of the tip is unknown or there is some uncertainty in our sampling process we can give all states the same initial condition, $D_{N_i}(0) = s_i d_i$. After initializing all of the tips, Equations 3.1 and 3.2 are solved backwards along the branches of the tree until a branching event is encountered.

At a bifurcation event, two edges coalesce to a singular ancestor. Without loss of generality, let these edges $M$ and $N$ be descendants of $A$. Since no prior information is known about the direction of infection the fact that either $M$ infected $N$ or $N$ infected $M$ is accounted for using the following equation,

$$ D_{Ai}(t) = \sum_{j=1}^{m} \left( \lambda_{i,j} D_{Mj}(t)D_{Nj}(t) + \lambda_{i,j} D_{Mj}(t)D_{Ni}(t) \right). $$

This process is repeated until the root is reached at which point we condition on the fact that at least one birth began this process (Stadler 2013) to obtain the likelihood of the tree

$$ \mathcal{P}(\mathcal{T}|\theta) = \sum_{j=1}^{m} \frac{D_{Nj}(t_0)}{1-E_i(t_0)}. $$

### 3.2.2 Continuous-Type Birth-Death Model

The extension of the MTBD-$m$ model to continuous-valued traits begins by taking the number of types, $m$, to infinity. All probability densities become explicit functions of both time ($t$) and the continuous type ($x$) such that the probability that an individual $N$ evolved from time 0 to $t$ and the probability that an individual in state $x$ is not sampled and has no sampled descendants become $D_N(x,t)$ and $E(x,t)$ respectively. For any continuous valued trait, we remap to the interval $[0,1]$ using the following transformation: $y = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}$. The trait space is therefore constrained between the maximum and minimum observed values such that we do not extrapolate outside of the sample space. If necessary, this assumption can be relaxed by choosing a user defined upper and lower bound ($u.b.$ and $l.b.$) such that
\[ u.b. \geq \max(x) \text{ and } l.b. \leq \min(x). \] This would lead to the transformation being defined as:
\[ y = \frac{x - l.b.}{u.b. - l.b.}. \]

**PDE Derivation**

Consider the \( \gamma \) matrix as a transition rate matrix strictly defining changes occurring along a lineage independent of birth/transmission events. A discrete diffusion approximation is applied to allow for these transitions to occur. Lineages can only move to the states in direct contact with itself (i.e. an individual of type \( i \) can only change to types \( i-1 \) and \( i+1 \) between transmission events). This sparsifies the \( \gamma \) matrix to a tri-diagonal matrix, dramatically reducing the dimensionality of the problem. Using the notation from Malaspinas et al. (2012), equations 3.1 and 3.2 are rewritten as the following:

\[
\frac{d}{dt} D_{Ni}(t) = - \left( \sum_{j=1}^{m} \gamma_{i,j} + d_i \right) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Ni}(t) - \left( \gamma_{i,i+1} + \gamma_{i,i} + \gamma_{i,i-1} \right) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Nj}(t) + \gamma_{i,i+1} D_{Ni+1}(t) + \gamma_{i,i} D_{Ni}(t) + \gamma_{i,i-1} D_{Ni-1}(t) \tag{3.3}
\]

\[
\frac{d}{dt} E_i(t) = (1 - s_i) d_i - \left( \sum_{j=1}^{m} \lambda_{i,j} + d_i \right) E_i(t) - \left( \gamma_{i,i+1} + \gamma_{i,i} + \gamma_{i,i-1} \right) E_i(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) E_i(t) + \gamma_{i,i+1} E_{i+1}(t) + \gamma_{i,i} E_i(t) + \gamma_{i,i-1} E_{i-1}(t) \tag{3.4}
\]

This can be re-written by combining like terms to be:

\[
\frac{d}{dt} D_{Ni}(t) = - \left( \sum_{j=1}^{m} \lambda_{i,j} + d_i \right) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Nj}(t) \tag{3.5}
\]

\[
\frac{d}{dt} E_i(t) = (1 - s_i) d_i - \left( \sum_{j=1}^{m} \lambda_{i,j} - d_i \right) E_i(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) E_i(t) \tag{3.6}
\]
Note that \( m \) is the number of classes present in the system. If \( \gamma_{i,i-1} = \gamma_{i,i+1} \) for all \( i \in \{2, ..., m - 1\} \), then there is no bias in the direction of transition. The transition rates within the \( \gamma \) matrix are representative of the combined action of multiple evolutionary processes. For this reason, let us redefine \( \gamma = \gamma^D + \gamma^A \) where both \( \gamma^D \) and \( \gamma^A \) are matrices whose sum is the transition rate for all possible state changes. Let \( \gamma^D \) be symmetric such that \( \gamma^D_{i,i-1} = \gamma^D_{i,i+1} \) for all \( i \in \{2, ..., m - 1\} \). Thus, \( \gamma^D \) purely represents the diffusive forces (i.e. genetic drift) in the transition matrix. Then, let \( \gamma^A = \gamma^A_{i,i+1} - \gamma^A_{i,i-1} \) such that \( \gamma^A \) purely represents the advective forces (i.e. directional selection) in the transition matrix. This implies that \( \gamma^A_i > 0 \) if the advective forces lead to a higher state \((i + 1)\), whereas \( \gamma^A_i < 0 \) if the opposite is true.

Therefore, the terms associated with \( \gamma \) in the equations above (3.5-3.6) can be split into the following:

\[
\frac{d}{dt} D_{Ni}(t) = -\left( \sum_{j=1}^{m} \lambda_{i,j} + d_i \right) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Ni}(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) D_{Nj}(t) \\
+ \gamma^A_{i,i-1} (D_{Ni-1}(t) - D_{Ni}(t)) + \gamma^A_{i,i+1} (D_{Ni+1}(t) - D_{Ni}(t)) \\
+ \gamma^D (D_{Ni+1}(t) - 2D_{Ni}(t) + D_{Ni-1}(t)) \quad (3.7)
\]

\[
\frac{d}{dt} E_i(t) = (1 - s_i) d_i - \left( \sum_{j=1}^{m} \lambda_{i,j} - d_i \right) E_i(t) + \sum_{j=1}^{m} \lambda_{i,j} E_j(t) E_j(t) \\
+ \gamma^A_{i,i+1} (E_{i+1}(t) - E_i(t)) + \gamma^A_{i,i-1} (E_{i-1}(t) - E_i(t)) \\
+ \gamma^D (E_{i-1}(t) - 2E_i(t) + E_{i+1}(t)) \quad (3.8)
\]

Let \( m \to \infty \) such that the probability densities of interest are explicitly functions of both continuous time and space (i.e. \( D_N(x,t) \) and \( E(x,t) \)). In particular, they are both defined from \( t = [0, t_0] \) where \( t_0 > 0 \) and on \( x = [0, 1] \).

Equations 3.9 and 3.10 then follow:
\[
\frac{\partial}{\partial t} D_N(x, t) = - \left( \int_0^1 \lambda(x, y) + d(x) \, dy \right) D_N(x, t) + \int_0^1 \lambda(x, y) E(y, t) D_N(x, t) \, dy \\
+ \int_0^1 \lambda(x, y) E(x, t) D_N(y, t) \, dy \\
+ \gamma^A(x) \frac{\partial}{\partial x} D(x, t) + \gamma^D(x) \frac{\partial^2 D(x, t)}{\partial x^2} \\
(3.9)
\]

\[
\frac{\partial}{\partial t} E(x, t) = (1 - s(x)) d(x) - \left( \int_0^1 \lambda(x, y) - d(x) \, dy \right) E(x, t) + \int_0^1 \lambda(x, y) E(x, t) E(y, t) \, dy \\
+ \gamma^A(x) \frac{\partial}{\partial x} E(x, t) + \gamma^D(x) \frac{\partial^2 E(x, t)}{\partial x^2} \\
(3.10)
\]

Each individual of state \(x\) gives birth to an individual of state \(y\) at rate \(\lambda(x, y)\). For simplicity, let \(\lambda(x, y) = 0\) if \(x \neq y\). This ensures that individuals can only give birth to individuals of the same type. An alternative transmission kernel that relaxes this assumption will be presented in Section 3.6.1. When \(x = y\), let \(\lambda(x, y) = f(x) \cdot \beta_0\) where \(f(x)\) is the relative population-level fitness advantage of a lineage in state \(x\) while \(\beta_0\) is the base transmission rate for the phylogeny.

The diffusion (\(\gamma^D(x)\)) and advection (\(\gamma^A(x)\)) parameters are general functions that can be defined in a variety of ways. For now, we will keep them in this form to maintain generalizability. We will do this by redefining the rate of moving to larger or smaller continuous traits as \(\delta^f\) and \(\delta^b\), respectively. Note that this reparameterization is done for computational purposes as it can be more stable in some examples. This accounts for the fact that if advection is positive and driving the continuous trait to larger values at the within host scale it should be found on the upper diagonal, whereas the opposite is true if advection is
negative and driving the continuous trait to smaller values. To be precise let

\[ \delta^f(x) = \begin{cases} 
\gamma^D(x) + \gamma^A(x) & \gamma^A(x) \geq 0 \\
\gamma^D(x) & \gamma^A(x) < 0 
\end{cases} \]

\[ \delta^b(x) = \begin{cases} 
\gamma^D(x) + |\gamma^A(x)| & \gamma^A(x) \geq 0 \\
\gamma^D(x) & \gamma^A(x) < 0 
\end{cases} \]

This parameterization leads to the following tridiagonal \( \gamma \) matrix being populated as follows. Recall that we discretize our continuous space \( x \in [0, 1] \) into \( n \) intervals:

\[
\gamma = \begin{pmatrix}
0 & \delta^f(0) & 0 \\
\delta^b(\frac{1}{n}) & 0 & \delta^f(\frac{1}{n}) \\
& \ddots & \ddots & \ddots \\
& & \delta^b(\frac{n-1}{n}) & 0 & \delta^f(\frac{n-1}{n}) \\
0 & & & \delta^b(1) & 0 
\end{pmatrix}.
\]

One thing of note is that the constant forward and backward transition along with the advection-diffusion parameterizations both use two parameters to allow for the \( \gamma \) matrix to be non-symmetric. This is a key component of this work as we endeavor to quantify within host processes.
Figure 3.1: (a) Phylogeny produced through forward simulation. The true state of a lineage is visualized and sampled tips are represented by open points. (b) Solution of the PDE representing probability density $D_N(x, t)$ backwards through time (Eq. 3.9) for the two lineages of interest (boxed). Red and green curves along with their color gradients correspond to the probability distribution for the noted lineage and time.

The parameter estimation process for this phylodynamic model is dependent on two sets of data, the phylogenetic structure and the sampled tip states. As presented in Figure 3.1, we can see how the probability distribution evolves as Equations 3.9 and 3.10 are solved backwards in time along the phylogeny. How these sets of data interact to create specific estimates of high or low diffusivity as well as directional selection can be slightly mystifying. Figure 3.2 provides a simplified illustration of how phylogeny and tip states interact. In particular, two identical phylogenies are presented with the same distribution of tip states. However, the relative pairwise phylogenetic and continuous state distance between the tips are distinct. In Figure 3.2a, the tip states that are closest on the continuous domain are also most phylogenetically related to one another. The opposite is true in Figure 3.2b. The most phylogenetically related tips are also sampled to be on opposite ends of the continuous domain. This ensures that the diffusion coefficient, or equivalently $\min(\delta^f, \delta^b)$, must be larger for the latter data set. Note that the red (blue) distributions represent the probability of being in a particular state when the diffusion coefficient is large (small).
addition, recall that the parameter combination that is most likely will be the one that leads
to the largest amount of probability density being conserved at the root. We can see in
this caricatured illustration that a large diffusion coefficient allows for sufficient density
overlap when phylogenetically related tips are sampled to be far from one another on the
continuous domain. However, a small rate of diffusion is more likely when we find that
phylogenetically related tips are sampled close to one another on the continuous domain.

The directionality (i.e. $\text{sign}[\delta^f - \delta^b] = \text{sign}[\gamma^D]$) and strength (i.e. $|\delta^f - \delta^b| = \gamma^D$) of
selection is informed in a similar manner. Figure 3.2c illustrates how the directionality and
strength of selection may be quantified. Note that the red (blue) distributions represent the
probability of being in a particular state when the strength of selection is strong (weak). In
the central monophyletic group we can see that all but one of the tips is sampled on the far
right of the continuous space. With the one sampled tip on the far left of the continuous
space the rapid branching events along lineages with high probability of being on the
far right of state space implies strong selection. We can see the theoretical probability
distribution across the continuous space is significantly reduced when a small selective
force is assumed (see branching event at $\tau^*$). This is due to the lack of overlap across during
this rapid diversification event. In practice a phylogeny of this size would lead to extremely
large confidence intervals on any parameter estimates. However, one can imagine how if this
pattern were repeated many times on a large phylogeny consisting of a monophyletic clade
sampled in a cluster on the continuous domain along with multiple sets of paraphyletic
samples found to be distant on the continuous domain, we would be able to reasonably
estimate the directionality as well as strength of selection.
Figure 3.2: Illustration of how phylogenetic structure and sampled tip states can impact phylodynamic parameter estimates. Starting from point masses at the tips \((t = 0)\) we can follow the probability distribution backwards in time to the root \((t = t_0)\). In one tree, phylogenetic relatedness is correlated with sampled state (a) while it is more independent in (b). Red (blue) represents the probability distribution solved backwards in time for a large (small) diffusion coefficient. (c) An example of how selective forces can be informed. Red (blue) represents the probability distribution solved backwards in time for a large (small) selective coefficient.

### 3.2.3 Strain Frequency Derivation

Up to this point the derivation of the PDEs for the CTBD model assume nothing about the scale at which evolutionary processes are occurring, just that the trait is evolving along lineages. However, we will now consider special cases of the CTBD model where the advection/diffusion parameters are assumed or can be interpreted to reflect within host processes. One potential continuous trait of interest could be the frequency of a particular variant or strain of interest, as has been widely studied throughout population genetics. This application naturally is constrained (i.e. \(x \in [0, 1]\)). In particular, we define the diffusion \((\gamma^D(x))\) and advection \((\gamma^A(x))\) parameters in terms of the classic population genetic values effective population size \((N_e)\) and within-host fitness advantage \((s_w)\). This parameterization follows directly from the Wright-Fisher model to be:

\[
\gamma^D(x) = \frac{x(1-x)}{2N_e} \quad \text{and} \quad \gamma^A(x) = \frac{x(1-x)}{2\bar{w}} \frac{d\bar{w}}{dx} = \frac{x(1-x)}{2(1+s_wx)}s_w
\]  

(3.11)
where the average fitness at the within host level, \( \bar{w} \), is defined to be:

\[
\bar{w} = (1 + s_w) \cdot x + 1 \cdot (1 - x) = 1 + s_w \cdot x.
\]

Note that both \( \gamma_D \) and \( \gamma_A \) are proportional to the variance \( x(1 - x) \) in allele frequencies between discrete generations. As noted above, the within-host transition matrix \( (\gamma) \) is populated with \( \gamma_A \) terms in the upper diagonal of a tridiagonal matrix when \( s_w \geq 0 \) and in the lower diagonal when \( s_w < 0 \). In addition, both diffusion and advection are both maximized at \( x = 0.5 \).

### 3.2.4 Parameter estimation

In this section we will outline the results of parameter estimation for data simulated under the CTBD. To perform inference we implemented the CTBD model in BEAST 2, a Bayesian phylogenetic computational resource (Bouckaert et al. 2019). BEAST 2 allows us to jointly infer model parameters together with the phylogenetic tree itself. Before any simplifications were made, every parameter is continuous leading to an infinitely large parameter space of interest. The transmission matrix, \( \lambda \), and the transition matrix \( \gamma \) are of particular interest as they define the rates of branching events and transition events on branches that occur across the multiple types in our model. The assumptions outlined in Section 3.2.2 are vital as they ensure this is a computationally tractable problem. The base transmission rate can be jointly estimated along with the fitness mapping function, but in practice estimating the base transmission rate first and then fixing it when estimating the fitness mapping function leads to considerable improvements in MCMC mixing. We then run the CTBD, ensuring that all fitness estimates are relative to the base transmission rate for the phylogeny. The function that maps the continuous trait to between host fitness, \( f(x) \), is estimated using a nonparametric natural cubic splines approach (Burden and Faires 1989). Using a small number of knots (\( \approx 5 \)) allows for the estimation of a smooth functional form that is flexible.
across the continuous space of interest while maintaining a relatively small parameter space to explore.

### 3.2.5 Simulation: CTBD

We first tested the CTBD model by forward simulating multiple epidemics across a variety of parameter combinations. In particular, we wanted to determine our ability to reconstruct within and between host evolutionary processes under a variety of sampling fractions, fitness mapping functions, and rates of diffusion and advection. This is done using stochastic forward simulations to generate phylogenies under which a continuous trait is evolving and impacting host level transmission rates. Provided a single phylogenetic tree along with the continuous trait each host is sampled in, we use the CTBD to estimate the parameters of interest.
Figure 3.3: Phylodynamic analysis using the CTBD for within and between host selection pressures. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host variability, backward ($\delta^b$) and forward ($\delta^f$) selection. Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across the continuous trait space (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the fitness value of all trait values and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions.
Accurate estimates of model parameters using the continuous type birth-death model on simulated phylogenies is demonstrated under two different parameterizations of the gamma matrix. For each parameterization we used the CTBD to estimate two different combinations of the within and between host associated rates. The first set of results are presented in Figures 3.3 and 3.9. Here we simulated a phylogeny with the gamma matrix parameterized such that selection is driving lineages towards larger continuous values while all trait values have the same between-host fitness (neutral evolution). The 95% credible intervals contain the true values of the within-host parameters. A similar result is found for the cubic splines approach estimating transmission fitness effects across continuous state space. The number of samples found in each particular region of the continuous trait is correlated with posterior uncertainty of the fitness effects. We also show that the model estimates are accurate for simulation results for when there is no directional selection at the within-host scale while lineages that have a larger continuous value are more fit at the between host scale. The ability to estimate parameter combinations is consistent across multiple simulations as demonstrated in Figures 3.4, 3.10 and 3.11.

In the second set of simulations, the gamma matrix is parameterized to model allele frequencies under the Wright Fisher models as outlined in Section 3.2.3 and transmission fitness increases at higher within host allele frequencies. We are able to make accurate estimates regarding the within-host selection pressures, effective population size of the pathogen, and between-host transmission function across various parameter combinations (Figures 3.12-3.14).
3.3 Application to HIV-1

In this section we apply the CTBD model to a set of viral genomic data from a population of hosts infected with human immunodeficiency virus (HIV) type 1. We will provide background insight for the questions we hope to answer for this application. In Chapter 4 we will take a deeper dive into HIV-1 and provide more background on the intricacies of its history and evolutionary background.

3.3.1 Background

HIV is one of the most damaging human pathogens and is thought to infect approximately 1.7 million people across the globe in 2019 alone, adding to the over 38 million people living with HIV (UNAIDS 2020). Mathematical models have provided considerable insight into the
evolutionary, epidemiological, and infection dynamics of HIV. Here we will demonstrate that our model can be used to estimate the impact of an important continuous trait, viral load, on transmission rates among HIV-1 infected individuals from King County, Washington. Viral load is the number of virions present during the asymptomatic period of HIV, prior to the onset of AIDS.

HIV is characterized by a rapid increase in viral particles after initial exposure called the acute or primary infection stage. This stage lasts approximately one to four weeks and hosts may exhibit flu like symptoms for portions of this period. It is during this time that seroconversion begins. Seroconversion is when the immune system begins producing HIV antibodies to fight the infection. This stage is followed by what is called the asymptomatic, chronic, or latent stage of infection and can last up to ten years. Here, viral counts remain relatively stationary at what is called the set point viral load (SPVL) measured in units of HIV particles per milliliter (mL) of blood. If left untreated during this latent period CD4+ T cell population densities will slowly decrease and when they fall below 200 cells per mm$^3$ that individual is classified as having progressed to AIDS.

3.3.2 Transmission

Viral load is known to be one of if not the largest determinant of transmissibility for HIV-1 with higher viral loads leading to a higher probability of transmission (Operskalski et al. 1997; Pedraza et al. 1999; Quinn et al. 2000; Fideli et al. 2001; Loutfy et al. 2013). This positive relationship between viral load and transmissibility suggests a between-host viral fitness advantage to higher viral loads. This is particularly relevant as there is evidence that suggests it is a heritable trait (Alizon et al. 2009; Fraser et al. 2014). Some work has attempted to directly quantify the functional relationship between viral load and transmissibility in HIV (Fraser et al. 2007; Wilson et al. 2008; Lingappa et al. 2010; Hughes et al. 2012; Wertheim et al. 2019; Fraser et al. 2014). Most of this work has interrogated this question in regards to log$_{10}$ viral load. A review of previous methods demonstrates how much of the heterogeneity
in transmission rates estimates is due to data set choice where multiple models find similar patterns when accounting for this (Blaser et al. 2014).

When thinking about viral load heritability and impacts on transmission rates one can also think about how viral load, or virulence, will evolve on longer time scales for HIV. For decades now an intense debate has persisted on the evolution of virulence. This debate can be distilled into the following question: when a novel parasitic relationship begins, does evolutionary theory predict virulence to decrease (the avirulence hypothesis) or approach an intermediate viral load (the trade-off hypothesis). Explicitly, the avirulence hypothesis advances the claim that well adapted parasites have evolved to be harmless to their hosts (i.e. low virulence). This hypothesis was so ingrained in the field that it was referred to as “conventional wisdom” by Anderson et al. (1982) when they and Ewald (1983) first proposed the trade-off hypothesis. This hypothesis suggests that there are costs and benefits to changing virulence. In particular, increased virulence leads to higher transmission rates while simultaneously decreasing host life expectancy thereby leading to an intermediate value of virulence that is optimized for these factors. More recently, the increased availability of ART and PREP may add additional pressures to viral load evolution making it unclear whether the virulence-transmission trade-off holds (Herbeck et al. 2016). There is recent evidence that viral load may in fact be under directional selection to increase transmissibility, however questions remain whether or not this is due to the onset of treatment as prevention programs (Wertheim et al. 2019).

All of these hypotheses rest on the relationship between SPVL and transmission fitness. Here we use our CTBD model to directly estimate the functional relationship between SPVL and transmissibility using data from King County, Washington, which includes the Seattle Metropolitan area. Since 2015, King County accounts for approximately half of all confirmed cases of HIV in the state of Washington (Seattle and King County Public Health 2020). Because of this prominent HIV problem, state and local officials have developed initiatives to help combat its spread. Part of this initiative has been to provide sequence
data to identify and combat the main sources of ongoing HIV transmission. The data set we analyze here consists of demographic information, sequence data, viral load counts, and results of drug resistance assays for individual HIV-1 infected hosts. Having paired sequence and viral load data allows us to ask questions about transmission dynamics that are directly linked to epidemiological and evolutionary questions of interest. The CTBD allows us to directly define the functional relationship between $\log_{10}$ viral load and transmissibility. In particular we utilize a non-parametric approach to avoid potential model mis specification.
Figure 3.5: Maximum likelihood reconstructed phylogeny for the 529 sequenced protease and reverse transcriptase coding regions of HIV-1 infected hosts in King County, WA. Tips and branches are colored according to observed or reconstructed log_{10} viral load of the host using PastML (Ishikawa et al. 2019).

3.3.3 Application Specific Methods

We obtained 18,460 sequences from either the Integrase (IN) or Protease/Reverse Transcriptase (PR/RT) coding regions of HIV-1 infected patients. These sequences are host
specific consensus sequences meaning any minority variants are not present in the data. The majority of sequences, 73.3%, came from patients residing in King County, WA. Of these King County specific sequences 11.5% and 88.5% are of the IN and PR/RT coding regions respectively. Some patients were sampled multiple times and for these patients we chose to use their first sequence. We do this to increase the probability that the hosts were not on antivirals at the time of sampling. This left 139 and 880 sequences from the IN and PR/RT coding regions respectively. Additionally, we removed any sequences from individuals without viral load data or when the CD4 counts were less than 200 cells per mm$^3$, an indicator used for AIDS diagnosis. To ensure samples were all subtype B we used COMET, an HIV-1 subtype identification tool (Struck et al. 2014). After these final data processing steps we were left with 88 IN and 529 PR/RT sequences. We utilized MAFFT for multiple alignment of the sequences using the Progressive method (Katoh and Standley 2013). Subsequently we used FastTree to obtain maximum likelihood phylogenies assuming a GTR model of sequence evolution with Gamma-distributed rate variation among sites (Price et al. 2010). Finally, to date the maximum likelihood phylogenies we used LSD assuming a molecular clock rate of 2e-3 substitutions per site per year (To et al. 2016). We run the CTBD model in BEAST 2 on these fixed phylogenies. Attempts were made to fit the CTBD to the IN data, but MCMC chains were not mixing sufficiently to obtain reasonable estimates for multiple parameters. This is likely due to the small sample size.

### 3.4 Results

We use a cubic splines approach to estimate an arbitrary fitness mapping function between VL and transmission risk. Recall that we are using the measured VL from infected hosts as a proxy for SPVL. We find that transmission rates have a nonlinear relationship with increased VL. In particular, individuals with VLs of approximately 5 log$_{10}$ copies per ml were found to have the highest transmission rates while those with viral loads below 4 and
above 6 log_{10} copies per ml were found to have low transmission rates (Figure 3.6). We also present the probability that a given transmission event occurs from a host in each VL class (Figure 3.7). This is calculated by weighting the inferred fitness function by the population VL distribution founding using a kernel density estimate (KDE). This shifts the VL of highest priority slightly closer to lower values, but the peak transmission potential remains centered at intermediate VLs.

We compare the fit using the cubic splines approach to two other between-host fitness mapping functions that have been used previously to model transmissibility. This was done as it was not necessarily expected for intermediate VLs to have the largest fitness. We compare the likelihood of the data under our nonlinear model to two models that assume a monotonic increasing relationship between VL and fitness, Michaelis-Menten equation and an exponential rate function, to see if they explain the data equally well (Blaser et al. 2014). We find that the marginal likelihood of the data under the model is lower for these approaches in comparison to the cubic splines fit (Table 3.1).

In Figure 3.16 we can see how log_{10} viral load is distributed across time. We use the CTBD to estimate how much VL changes within individuals and across time. Figure 3.6 shows evidence for an increase in VL at a median rate of 0.064 log_{10} copies per ml per year with 95% credible intervals of (0.023, 0.11). Previous estimates for the average rate of VL increase across time fall between 0.009 and 0.016 log_{10} copies per ml per year depending on the stage of infection (Wertheim et al. 2019). Instead our results suggest that a mixture of both within and population level processes are what is pushing VL to larger values.
Figure 3.6: Phylodynamic analysis of 529 HIV-1 paired samples of log_{10} viral load per ml and PR/RT sequences using the CTBD. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host evolution, rate of selection (log_{10} copies per ml to consensus per year) and diffusion (log_{10} copies per ml to consensus^2 per year). Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across log VL (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the functional form and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions.
Figure 3.7: Phylodynamic analysis of 529 HIV-1 paired samples of log_{10} viral load and 
PR/RT sequences using the CTBD. Posterior distribution for the fitness mapping function 
weighted by the KDE estimate for the density of host log_{10} viral loads per ml. The 95% 
credible intervals are shown in shaded intervals.

3.5 Discussion

In this Chapter we develop a phylodynamic model capable of quantifying how multiscale 
selection pressures are acting on an evolving continuous trait by leveraging phylogenetic 
structure along with host information to quantify functional relationships of interest. We 
apply the CTBD to simulated data to demonstrate the ability to estimate all parameters of 
interest under a variety of scenarios.

In the last portion of this chapter, we apply the CTBD to hosts infected with HIV - 1
subtype B. Phylogenies are paired with viral load data to quantify the relationship between viral load and transmissibility, a relationship that is of interest to public health officials. We quantified this relationship using our CTBD model paired with a cubic splines approach to fit a non-parametric fitness function. A nonlinear relationship between transmissibility and $\log_{10} VL$ was found where intermediate VLs between 4 and 6 $\log_{10}$ copies per ml have the largest rates of transmission. Presently, the gold standard for transmissibility estimation comes from studies that follow individuals in serodiscordant relationships (Kiwanuka et al. 2009; Fraser et al. 2007; Hughes et al. 2012; Loutfy et al. 2013). These studies provide a tried and true strategy to answer questions about how transmissibility is dependent on a variety of demographic and virus specific characteristics. However, these studies can be costly to attain and with the explosion of high-throughput next generation sequencing we show that phylodynamics can provide an alternative route to these answers.

While we did not directly test for a fitness trade-off, our results indicate a potential trade-off between virulence and transmission. The decreased transmission rate we infer for individuals with higher VL may not necessarily reflect reduced transmission potential, but rather faster removal from the infectious class or the sexually active population as the virulence trade-off hypothesis suggests (Alizon et al. 2009). We are not the first to present evidence of virulence trade off’s existence in HIV-1 (Fraser et al. 2007; Blanquart et al. 2016). Our estimates for the peak transmission fitness are $0.5\ldots 1 \log_{10}$ viral load per ml higher than this previous work. One possible explanation could be due to variability in optimal viral load between subtypes and/or regions of the world. In particular, Blanquart et al. (2016) utilized data from serodiscordant partnerships from Uganda who were infected primarily with HIV-1 subtypes A and D which are known to have faster progression to AIDS than HIV-1 subtype B (Venner et al. 2016; Baeten et al. 2007; Kaleebu et al. 2002; Kiwanuka et al. 2008). This would lead to a steeper dropoff in population-level fitness at high viral loads. Therefore it is quite possible that HIV-1 subtype B holds the same qualitative pattern of highly fit intermediate viral loads, but transmission peaks at larger viral loads.
Estimates for the functional relationship between log_{10} viral load and transmissibility in HIV have been shown to be data set dependent (Blaser et al. 2014). It would be insightful if we could apply the CTBD to a previously analyzed data set to compare to previous estimates as done in Kiwanuka et al. (2009). However, as this is a phylodynamic method, the requirement for data sets to pair viral load and genomic information can make this difficult. It is also important to note that we assume that the sampling and removal rate is constant across all viral loads. We do not estimate these parameters across viral load. Future work could look into the feasibility of estimating the impact of continuous valued traits on birth, death, and sampling rates. This would potentially allow for a direct test of the virulence trade-off as it predicts a faster removal from the infectious population (i.e. a higher death rate in the CTBD model) as transmission increases with larger VL.

The functional relationship we infer between VL and transmission must be interpreted carefully as we were unable to control for confounding factors such as the stage of HIV infection. We attempted to account for stage of infection by removing any hosts who have transitioned to AIDS (CD4 count less than 200 cells per mm^{3}) and using the first sequence from hosts who had multiple samples in the data set. However, during the acute phase of infection, known for high transmission rates, viral loads are at their peak leading to correlation between VL and transmission that may not be causal. If we could obtain a good metric to classify hosts being in the primary or secondary stage of infection we could perform independent analyses on how viral load impacts transmission rates at these two stages. We also reiterate that the viral load counts sampled from infected hosts are not necessarily SPVLs. We attempt to account for this using some of the above techniques, however it is possible that we have just captured individuals at time points where there VL was exceptionally high or low in relation to their true SPVL. The average VL of our data set is $4.64 \pm 1.0 \ \text{log}_{10}$ copies per ml which generally corresponds with previous estimates of average SPVL (Fraser et al. 2007; Huang et al. 2012; Wertheim et al. 2019).

Lastly there are additional factors that can lead to increased transmission rates, poten-
tially confounding a direct interpretation of the relationship between VL and transmission. For example, risk behavior is known to have a large impact on transmission rate and is highly dependent on awareness of infection. It is estimated that approximately 25% of HIV positive individuals are unaware of their status, but that this group transmits at 3.5 times the rate than those who are aware of their status (Marks et al. 2006). This is a pattern seen across infectious disease modeling and is a classic paradigm in human behavior, defined as the Pareto principle and colloquially referred to as the 80/20 rule (Woolhouse et al. 1997; Lloyd-Smith et al. 2005). This disparity in transmission rates between aware and unaware HIV positive hosts demonstrates how vital testing and education are to reduce prevalence (Marks et al. 2006). Additionally, both ulcerative and non-ulcerative sexually transmitted diseases (STD) are known to increase transmissibility of HIV, contributing further complexities (Fleming and Wasserheit 1999). Likely the largest confounding variable when attempting to quantify the relationship between viral load and transmissibility is stage of infection. It is known that during the acute stage of infection, the probability of transmission is higher than during the asymptomatic phase, likely due to the impact of higher viral loads (Wawer et al. 2005; Brenner et al. 2007; Powers et al. 2008; Miller et al. 2010). However, the relative contribution of acute versus asymptomatic phases is still an open question as the extensive duration of the asymptomatic phase may in turn lead to it having a large contribution to transmission (Hollingsworth et al. 2008; Ratmann et al. 2016). The interaction between increased viral load and stage of infection is a complicated one and relative increases in viral load appear to have greater impacts after the acute phase of infection (Wertheim et al. 2019).

In addition to demonstrating the flexibility of the CTBD, we present results on the CTBD being applied to a highly relevant question in HIV research —the functional relationship between viral load and transmissibility. We show that transmissibility is highly dependent on \( \log_{10} \) viral load and evidence that both within and between host selection pressures may lead to higher VLs while transmission rates may decline above a certain threshold as predicted by
the virulence-transmission trade-off hypothesis. We also present two potential extensions of the CTBD model that allow for a wider range of potential applications. The first extension builds upon the CTBD framework to allow for multiple continuous traits of interest to evolve simultaneously on a phylogenetic tree. The second extension demonstrates how generalizable the CTBD can be in answering field specific questions. There we develop the framework for the application to an antimicrobial pathogen.
3.6 Appendix

3.6.1 Full Transmission Kernel

The original development of this model assumes that individuals of a particular type only give birth to individuals of the same type. This is found in the construction of the transmission kernel where $\lambda(x, y) = f(x) \cdot \beta$ when $x = y$ and $\lambda(x, y) = 0$ otherwise, where $f(x)$ is the continuous fitness function of interest. This can be both technically problematic and potentially a bad assumption for specific applications. The technical problem appears when a set of closely related samples are found to be in vastly divergent continuous states. If the branches are short enough such that the probability density of these two child nodes do not overlap at their branching event, then their parent will have no probability density across all continuous states. This zero probability density will propagate to the root and lead to the likelihood function equalling zero.

Additionally, there is an inherent claim that the trait is completely heritable (i.e. does not change at birth/transmission events). This assumption may be more or less appropriate for certain applications. To account for these possibilities we populate the off-diagonals of the transmission kernel, $\lambda(x, y) > 0$ when $x \neq y$. This ensures that a non-zero probability density is always passed to the parent at branching events and represents how heritable the trait is at transmission. The probability of transmission between two states, $x$ and $y$, decreases as the distance between $x$ and $y$ increases (i.e. $\lambda(x, x + \delta_1) \geq \lambda(x, x + \delta_2)$ for all $x \geq |\delta_2| > |\delta_1| \geq 0$). We do this by defining the transmission kernel to be the following:

$$
\lambda(x, y) = \begin{cases} 
\beta_0 \cdot f(x) & x = y \\
\beta_0 \cdot f(x) \cdot \frac{\sqrt{2}}{\sigma \pi} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right) \text{erf}\left(\frac{\sigma}{\sigma}\right) & x \neq y
\end{cases}
$$

Here, we use a truncated Gaussian distribution to model how transmission rate decreases as the distance between child and parent trait increases. We do this as we believe it is the most
general functional representation for heritability. Note that other functional representations of this decay can be utilized for any specific applications. We use a truncated Gaussian to ensure that the total fecundity of an individual in state $x$ is equal to the product of the base birth rate and the fitness mapping function (i.e. $\int_{0}^{1} \lambda(x, \xi) d\xi = \beta_{0} \cdot f(x)$). This ensures that the fitness mapping function still determines the overall transmission rate from a particular trait value in continuous space. The spread of the transmission kernel is driven by the standard deviation or $\sigma$, a parameter that we can jointly infer. Here we show how the truncated Gaussian distribution varies for various combinations of $x$ and $\sigma$ (Figure 3.8).

![Figure 3.8](image)

Figure 3.8: Impact of varying decay rates on the proportion of $\beta_{0}$ found on the transmission kernel $\lambda(x, y)$. Note that $|x - y|$ is presented here as the transmission kernel is symmetric.

This full transmission kernel will be particularly beneficial to data sets where individual hosts are sequenced multiple times as the amount of trait variability due to pathogen bottlenecks would potentially be informed.
3.6.2 Simulation: CTBD

Figure 3.9: Phylodynamic analysis using the CTBD for within and between host selection pressures. The top two rows show the posterior distributions and MCMC trace plot for the parameters associated with within host variability, backward ($\delta^b$) and forward ($\delta^f$) selection. Uniform priors were placed on these parameters. True parameter values are represented by dashed red lines whereas posterior means and 95% credible intervals are represented by solid blue vertical and horizontal lines respectively. The bottom most row contains the posterior distribution (blue) for the fitness mapping function across the continuous trait space (left) in addition to the histogram of sampled tip states (light gray). A log normal prior distribution was placed on the functional form and is visualized in light grey. Both 50% and 95% credible intervals are shown in shaded intervals for the prior and posterior distributions.
(a) Median estimates for the between host fitness function. The solid lines represent median estimates from each simulation while the dashed red line represents the ground truth.

(b) Median estimates for the backward and forward rates of selection. The solid black lines represent the ground truth while the connected colored points represent each simulation.

Figure 3.10: Estimates of median parameter values for ten forward simulations under the same evolutionary scenario. Simulations were conducted with neutral processes driving the evolution of a continuous trait at both the within and between host scale. Estimates for all simulations surround the ground truth for both within and between host parameters.
Figure 3.11: Estimates of median parameter values for ten forward simulations under the same evolutionary scenario. Simulations were conducted with a positive selection pressure between continuous trait and fitness at the between host scale while no within host selection bias is present. Estimates for all simulations surround the ground truth for both within and between host parameters.

(a) Median estimates for the the between host fitness function. The solid lines represent median estimates from each simulation while the dashed red line represents the ground truth.

(b) Median estimates for the backward and forward rates of selection. The solid black lines represent the ground truth while the connected colored points represent each simulation.

Figure 3.12: Estimates of median parameter values under the strain frequency parameterization found in Section 3.2.3. The true underlying parameters were $N_e = 100$, $s_w = -1$, and $s_h = .2$.

(a) Violin plot for the within host selection pressure ($s_w$). The interior box and whisker plot defines the quartiles while the red dashed line represents the ground truth.

(b) Violin plot for the effective population size ($N_e$). The interior box and whisker plot defines the quartiles while the red dashed line represents the ground truth.
Figure 3.13: Estimates for the between-host transmission rate under the strain frequency parameterization found in Section 3.2.3. The 50% and 90% credible intervals are found in light and dark blue while the mean and true fitness curves are the solid black and dashed red lines respectively. The true underlying parameters were $N_e = 100$, $s_w = -.1$, and $s_h = .2$.

Figure 3.14: Estimates of median parameter values under the strain frequency parameterization found in Section 3.2.3. The true underlying parameters were $N_e = 100$, $s_w = .1$, and $s_h = -.1$.

(a) Violin plot for the within host selection pressure ($s_w$). The interior box and whisker plot defines the quartiles while the red dashed line represents the ground truth.

(b) Violin plot for the effective population size ($N_e$). The interior box and whisker plot defines the quartiles while the red dashed line represents the ground truth.
Figure 3.15: Estimates for the between-host transmission rate under the strain frequency parameterization found in Section 3.2.3. The 50% and 90% credible intervals are found in light and dark blue while the mean and true fitness curves are the solid black and dashed red lines respectively. The true underlying parameters were $N_e = 100$, $s_w = .1$, and $s_h = -.1$.

### Changes to Lumiere Estimation

We have added a few additional computational changes. These are implemented to either allow for sufficient probability to travel backwards along the tree or as a conservative approach to the likelihood calculation. The first alteration we performed was for the possibility that two tips were sampled on dramatically different portions of the continuous trait of interest.

Previous versions of code implemented in BEAST 2 were unable to calculate the likelihood of any trees with more than 500 tips when the continuous trait is discretized into approximately 50 states. This was due to an underflow issue where re-scaling factors were utilized across all traits along with the conditional likelihood of the entire tree. In our case the probability of being in any particular state of our continuous trait diffuses and advects across all potential states. This leads to comparably low probabilities $D_{Ni}$ within a particular state position which only leads to smaller and smaller state probabilities after more and
more branching events occur. After some critical number of branching events, numerical issues are run into as the probability of being in a particular state is more than 2040 times smaller than the conditional likelihood of the entire tree. We ameliorate this numerical issue by allowing for each trait, as well as the conditional likelihood of the entire tree, to have a trait specific scaling factor that is accounted for in the final likelihood calculation. We liken this task to taking out a loan as we “borrow” probability density that we eventually pay back once we reach the root of the tree. The goal of this technique is to accurately compute the probability density and in turn should provide a good estimate of the true probability density. This metaphor is surprisingly apt as the “interest” we pay on our loan is computational.

The MTBD and CTBD are both leveraging information from the sampled tips to estimate the probability of being in each state as we look deeper and deeper into the past. Therefore there is an implicit assumption here that we believe that our model can accurately estimate ancestral states back to the root. These techniques are what allow phylogeographic techniques to estimate the originating location of a set of samples. We believe that there are some potential applications where this assumption may break. For this reason we incorporate an option for the user to choose some point in the past after which we do not solve the partial differential equations 3.9 and 3.10 backwards in time. This time should be deep enough in the past such that we can leverage all samples in our data set, but close enough to the present such that the assumption about ancestral state reconstruction is not entirely abused. The choice of this time point is data set and problem specific.
3.6.3 Viral Load Application

Figure 3.16: Distribution of log_{10} viral loads for HIV-1 positive individuals sequenced in King County, WA. Each individual also had the region of their genome that codes for the protease and reverse transcriptase enzymes sequenced. A reconstructed phylogeny is presented in Figure 3.5

Table 3.1: Likelihood and Bayes Factor calculation for various functional forms of the relationship between VL and fitness, f(x). The functional forms are exponential growth, Michaelis-Menten, and cubic splines with six knots.

<table>
<thead>
<tr>
<th>Model</th>
<th>Median Likelihood</th>
<th>Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential Growth</td>
<td>-4648.07</td>
<td>1.021</td>
</tr>
<tr>
<td>Michaelis-Menten</td>
<td>-4634.36</td>
<td>1.018</td>
</tr>
<tr>
<td>Cubic Splines</td>
<td>-4550.52</td>
<td>Ref</td>
</tr>
</tbody>
</table>
3.6.4 Extensions

In this section we will outline a few potential extensions of the CTBD from above. The first is the most straightforward extension, where multiple continuous valued traits are modeled simultaneously. The second extension will be directed towards applying our model to the evolution of antimicrobial resistance. This is a system where multi-level fitness tradeoffs are known to occur between environments with and without antimicrobial drugs, for which our model is well-suited.

Multi-Site CTBD Model: msCTBD

Developing a higher dimensional PDE system should be a straightforward extension of the PDE-based model. Although there are potentially more tractable ways of modeling multiple continuous traits evolving along a phylogeny, we outline precisely how this would be completed for \( m \) evolving continuous traits, which we will refer to as the multi-site CTBD model.

Let us define the probability densities to be \( D_N(x_1, \ldots, x_m, t) \) and \( E(x_1, \ldots, x_m, t) \). In particular, they are both defined from \( t = [0, t_0] \) where \( t_0 > 0 \) and on \( x_1, \ldots, x_m \in [0, 1] \). In Equations 3.9 and 3.10 the parameters of interest, \( \theta = [\lambda, \gamma, s, d] \), are defined to be continuous. Here they are representative of multiple continuous traits so we must re-define them as the transmission (\( \lambda \)) and transition (\( \gamma \)) kernels will likely vary for each combination of type \((x_1, \ldots, x_m)\). With \( \beta_0 \) being the base birth rate of the system, let \( f_\lambda(x_1, \ldots, x_m) \) be the fitness effect of being a particular type \((x_1, \ldots, x_m)\). We impose the same assumptions found above in the single trait CTBD model, where an individual of type \((x_1, \ldots, x_m)\) can only give birth to individuals of type \((x_1, \ldots, x_m)\). Let \( \lambda_i(x_i) \) define the fitness effects of being in state \( x_i \) at site \( i \) and in turn we can define \( \lambda \) to be the ordered list of these site specific fitness functions, or \( \lambda = [\lambda_1, \lambda_2, \ldots, \lambda_m] \). We can define the transmission rate from individuals of type \((x_1, \ldots, x_m)\) to individuals of type \((y_1, \ldots, y_m)\) to be \( \beta_0 \cdot f_\lambda(x_1, \ldots, x_m) \).

There are an infinite number of ways to choose \( f_\lambda(x_1, \ldots, x_m) \), however we will allow for
site specific fitness effects to either be additive or multiplicative. In other words,

\[ f_\lambda(x_1, ..., x_m) = \lambda_1(x_1) \cdot \lambda_2(x_2) \cdots \lambda_m(x_m) \text{ or} \]

\[ f_\lambda(x_1, ..., x_m) = \lambda_1(x_1) + \lambda_2(x_2) + ... + \lambda_m(x_m) \text{ iff } x_1 = y_1, x_2 = y_2, ... x_m = y_m \]

Note that both choices assume that the fitness effect of each site is independent of the
others, but that their interpretation of \( \lambda_i(x_i) \) differs between these two models.

Recall that in the CTBD model, \( \gamma \) combined diffusive and advective forces. Here we
allow each trait to be acted upon independently by drift and selective pressures. Note
that selection can still act on the combined fitness effects of all traits via \( f_\lambda \). Therefore, let
\[ \gamma = [\gamma_1, \gamma_2, ..., \gamma_m] \]. Note that \( \gamma_i = \gamma^D_i + \gamma^A_i \) where \( \gamma^D_i \) and \( \gamma^A_i \) are the diffusive and advective
forces acting on trait \( i \). Each \( \gamma_i \) is parameterized following the diffusion approximation
outlined above and in Malaspinas et al. (2012). In particular, \( \delta_{f,i} \) and \( \delta_{b,i} \) are defined as the
forward and backward transition rates for the \( i \)th site.

The death and sampling rates may vary for each combination of type \( (x_1, ..., x_m) \). There-
fore, \( d(x_1, ..., x_m) \) and \( s(x_1, ..., x_m) \) become the death and sampling kernels respectively. For
most analyses, the death and sampling rates here will be assumed to be constant across
trait and site space.

Therefore, we can rewrite Equations 3.9 and 3.10 to be the following when considering
Here we write them formally as the msCTBD model:

\[
\frac{\partial}{\partial t} D_N(x_1, \ldots, x_m, t) = -\left(\beta f + d(x_1, \ldots, x_m)\right) D_N(x_1, \ldots, x_m, t) + m \beta_0 f E(x_1, \ldots, x_m, t) D_N(x_1, \ldots, x_m, t)
\]

\[
+ \int_0^1 \gamma_C(x_1, y_1) \frac{\partial D_N(x_1, \ldots, x_m, t)}{\partial x_1} d y_1 + \int_0^1 \gamma_D(x_1, y_1) \frac{\partial^2 D_N(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
\]

\[
+ \cdots + \cdots
\]

\[
+ \int_0^1 \gamma_m(x_m, y_m) \frac{\partial D_N(x_1, \ldots, x_m, t)}{\partial x_m} d y_m + \int_0^1 \gamma_{m}^D(x_m, y_m) \frac{\partial^2 D_N(x_1, \ldots, x_m, t)}{\partial x_m^2} d y_m
\]

(3.12)

\[
\frac{\partial}{\partial t} E(x_1, \ldots, x_m, t) = (1 - s(x_1, \ldots, x_m)) d(x_1, \ldots, x_m) - \left(\beta f + d(x_1, \ldots, x_m)\right) E(x_1, \ldots, x_m, t)
\]

\[
+ \beta_0 f E(x_1, \ldots, x_m, t) E(x_1, \ldots, x_m, t)
\]

\[
+ \int_0^1 \gamma_C(x_1, y_1) \frac{\partial E(x_1, \ldots, x_m, t)}{\partial x_1} d y_1 + \int_0^1 \gamma_D(x_1, y_1) \frac{\partial^2 E(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
\]

\[
+ \cdots + \cdots
\]

\[
+ \int_0^1 \gamma_m(x_m, y_m) \frac{\partial E(x_1, \ldots, x_m, t)}{\partial x_m} d y_m + \int_0^1 \gamma_{m}^D(x_m, y_m) \frac{\partial^2 E(x_1, \ldots, x_m, t)}{\partial x_m^2} d y_m
\]

(3.13)

Note that Equations 3.12 and 3.13 look significantly simpler than their counterparts in Equations 3.9 and 3.10. This is because of the assumption that transmission events lead to births of the same type as the parent, allowing us to eliminate the integrals found in the first few terms. Although this assumption was also administered in the CTBD-1 model, we retained the most generalized model, as relaxation of the transmission assumption is still a viable option. However, the addition of new sites in the msCTBD model will only increase the number of parameters in the system, making the transmission assumption essential.
parameters we need to estimate scales linearly with \( m \) to be \( m(k + 2) \). Note that we do not include the inference of \( \beta_0 \), or the base birth rate, as this is done a priori assuming a constant birth and death rate for every state. This allows for the \( \lambda \) estimates to be relative fitness’ for a particular trait of interest. For larger numbers of sites this becomes increasingly intractable and it is likely that a parametric approach should be taken to estimate the transmission kernel as this would reduce parameter count.

**Within-Host Mixed msCTBD Model**

In Section 3.6.4 we assume that diffusion for the continuous traits of interest are independent. They can provide an understanding of the relationship among the traits of interest and their impact to within-host level dynamics. Something we have not incorporated are the mixed partial derivatives. Diffusivity is a site specific and to our knowledge can not be correlated with other diffusion processes. Advection on the other hand can have multiple moments that are being driven by multiple correlated sites. Ignoring this fact is likely less biologically realistic, but complicates the system significantly by incorporating additional parameters that may confound estimation procedures. For the sake of completion we demonstrate how this change would present itself in the msCTBD setting.

Let \( \gamma = [\gamma_{1,1}, \gamma_{1,2}, ..., \gamma_{1,m}, \gamma_{2,1}, \gamma_{2,2}, ..., \gamma_{m,m}] \). We assume that the correlation structure is symmetric (i.e. \( \gamma_{1,2} = \gamma_{2,1} \)). Note that this correlation structure only contains diffusive processes and in turn \( \gamma_{i,j} = \gamma_{i,j}^D \) where both the upper and lower parts of the tridiagonal matrix presented in the CTBD-1 will be populated by \( \delta_{i,j}^f = \delta_{i,j}^b \). Therefore, the number of additional parameters due to mixed partials is \( m(m - 1)/2 \). This could be a relatively minor addition if \( m \) is small. However, this gets intractable as \( m \) gets large. In those scenarios one could either revert back to the model presented in Section 3.6.4 or potentially use a LASSO type approach on all \( \delta_{i,j} \) values, ensuring that the most important relationships remain in the model. The latter may provide valuable insight into which continuous traits impact one another at the within-host level.
We deem this model the Within-Host Mixed msCTBD model and it is defined as follows:

\[
\frac{\partial}{\partial t} D_N(x_1, \ldots, x_m, t) = -\left[\beta_0 f_\lambda + d(x_1, \ldots, x_m)\right] D_N(x_1, \ldots, x_m, t)
+ m \beta_0 f_\lambda E(x_1, \ldots, x_m, t) D_N(x_1, \ldots, x_m, t)
+ \int_0^1 \gamma_1^C(x_1, y_1) \frac{\partial D_N(x_1, \ldots, x_m, t)}{\partial x_1} d y_1
+ \int_0^1 \gamma_1^D(x_1, y_1) \frac{\partial^2 D_N(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
+ \ldots + \int_0^1 \gamma_1^C(x_1, y_1) \frac{\partial D_N(x_1, \ldots, x_m, t)}{\partial x_1} d y_1
+ \int_0^1 \gamma_1^D(x_1, y_1) \frac{\partial^2 D_N(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
\]

\[
\frac{\partial}{\partial t} E(x_1, \ldots, x_m, t) = (1 - s(x_1, \ldots, x_m))d(x_1, \ldots, x_m)
- \left[\beta f_\lambda + d(x_1, \ldots, x_m)\right] E(x_1, \ldots, x_m, t)
+ \beta f_\lambda E(x_1, \ldots, x_m, t) E(x_1, \ldots, x_m, t)
+ \int_0^1 \gamma_1^C(x_1, y_1) \frac{\partial E(x_1, \ldots, x_m, t)}{\partial x_1} d y_1
+ \int_0^1 \gamma_1^D(x_1, y_1) \frac{\partial^2 E(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
+ \ldots + \int_0^1 \gamma_1^C(x_1, y_1) \frac{\partial E(x_1, \ldots, x_m, t)}{\partial x_1} d y_1
+ \int_0^1 \gamma_1^D(x_1, y_1) \frac{\partial^2 E(x_1, \ldots, x_m, t)}{\partial x_1^2} d y_1
\]

This method allows for directional processes and correlation structure amongst the sites be accounted for. This correlation structure is representative of potential mutational or selective forces acting concurrently on the sites.
Note that we do not delve into the possibility of a between-host correlated msCTBD model due to the severe limitations presented by the necessity to quantify correlation structure amongst all continuous variables across their entire domains. Even for the CTBD-2 case, this extension would become intractable quite quickly assuming that a cubic splines approach remains the mechanism of inference across the between-host fitness domain. As mentioned above, a parametric approach would have to be implemented and this has led to identifiability issues and less accurate estimates in the past.

Application to AMR in Bacteria

Here we lay out an extension to the CTBD model, summarized above in Section 3.2.1, adapted to model antimicrobial resistance in bacteria. This extension accounts for the status of antimicrobial therapy of the infected host and utilizes the following assumptions. We still track the probability densities across both continuous time \( t \) and space \( x \), where \( x \) represents a quantitative trait that reflects the antimicrobial resistance of the pathogen. This could be the frequency of an antimicrobial strain in a host or potentially the minimum inhibitory concentration (MIC) of the pathogen in the host. MIC provides a precise metric for the concentration of antibiotic required to inhibit pathogen growth. However, now individuals can be classified into three drug treatment groups: antimicrobial being administered at inhibitory concentrations (2), sub-inhibitory concentrations (1), or not being administered (0). Therefore, we will be explicitly solving six partial differential equations. They are \( D_{N,2}(x, t) \), \( D_{N,1}(x, t) \), \( D_{N,0}(x, t) \), \( E_2(x, t) \), \( E_1(x, t) \), and \( E_0(x, t) \). The model parameters, as described in Section 3.2.2, are \( \theta = [\lambda, \gamma, s, d] \). In this extension, let

\[
\lambda = [\lambda_{22}(x), \lambda_{21}(x), \lambda_{20}(x), \lambda_{12}(x), \lambda_{11}(x), \lambda_{10}(x), \lambda_{02}(x), \lambda_{01}(x), \lambda_{00}(x)]
\]

where \( \lambda_{i,j}(x) \) is the transmission rate from the \( i \)th to \( j \)th antimicrobial status group \( (i, j \in \{0, 1, 2\}) \) for type \( x \). We define this rate on the state of just the host’s type, \( x \), as we assume
that individuals only give birth to individuals of the same type. However, this allows indi-
viduals to give birth to individuals of a different antimicrobial status. This is made explicit
since individuals who are on antimicrobial treatment may still potentially transmit the
microbe to a different host who will not be on the antimicrobial treatment. Here a reason-
able assumption is made that regardless of treatment status, individuals infected with the
microbe will only give birth to individuals not on antimicrobial treatment (i.e. $\lambda_{ij}(x) = 0$
for all $x$ when $j \in \{1, 2\}$). Therefore, we can reduce $\lambda = [\lambda_{20}(x), \lambda_{10}(x), \lambda_{00}(x)]$.

For the transition rate within and between treatment status we use the following

$$\gamma = [\gamma_{22}(x, y), \gamma_{21}(x, y), \gamma_{20}(x, y), \gamma_{12}(x, y), \gamma_{11}(x, y), \gamma_{10}(x, y), \gamma_{02}(x, y), \gamma_{01}(x, y), \gamma_{00}(x, y)]$$

where $\gamma_{i,j}(x)$ is the transition kernel for the $i$th to $j$th antimicrobial treatment group
($i, j \in \{0, 1, 2\}$). Within treatment status groups ($\gamma_{22}(x, y), \gamma_{11}(x, y)$ and $\gamma_{00}(x, y)$) will be
defined in a similar manner to Equation 3.11. Both within treatment status transition rates
will be dependent on diffusive and advective forces parameterized using the Wright-Fisher
model (i.e. $\gamma_{00}(x, y) = \gamma_{00}^D(x, y) + \gamma_{00}^A(x, y)$). In particular, the within-host fitness advantage of
a particular antimicrobial pathogen will be defined by $s_{w,i}$ depending on the host’s antimi-
crobial treatment status while the effective population size of the within-host microbial
population ($N_e$) will drive diffusion. For the remainder of the section this rate of drift is
assumed to be constant across the three treatment status groups. However, this can be
relaxed if data suggests otherwise. In particular it is possible that individuals under an opti-
mal treatment regiment will have substantially lower viral population sizes leading to faster
diffusivity across allele frequency space. Between treatment status group transition rates
(i.e. $\gamma_{ij}(x, y)$ when $i \neq j$ and $i, j \in \{1, 2, 3\}$) will define the rate at which individuals are put
on antimicrobial treatments, the rate treatment is ended, or adherence otherwise changes
between optimal and sub-optimal concentrations. Since these parameters will define a
change in environment, they will not exact an immediate change in the frequency of AMR
pathogens in the host. Therefore, $\gamma_{ij}(x,y) = \gamma_{ji}(x,y) = 0$ if $x \neq y$, $i \neq j$ and $i, j \in \{1, 2, 3\}$. In addition we will assume this rate is constant for all $x$. This is another assumption that will make analysis considerably less complicated. However, it is very possible that increased frequency of AMR strains could impact transition between the treatment groups. Let $\gamma_{ij}(x,x) = \gamma_{ij}$ for all $x$ where $i \neq j$ and $i, j \in \{1, 2, 3\}$. The value for these parameters could be informed by knowledge on antimicrobial prescription and adherence amongst patients.

Lastly, we assume that the sampling and death rates for the three treatment status groups will differ by a set of constants, $\kappa^1_s, \kappa^2_s, \kappa^1_d,$ and $\kappa^2_d$ respectively, such that $s_1(x) = \kappa^1_s s_0(x)$, $d_1(x) = \kappa^2_d d_0(x)$, etc. This allows for the the rate of removal from the system (i.e. death) to vary between hosts on different treatments.
\[\frac{\partial}{\partial t} D_{N,2}(x, t) = -d_2(x)D_{N,2}(x, t) + \gamma_{02} D_{N,0}(x, t) + \gamma_{12} D_{N,1}(x, t) \]
\[ + \int_0^1 \gamma_{22}^C(x, y) \frac{\partial}{\partial x} D_{N,2}(x, t) dy + \int_0^1 \gamma_{22}^D(x, y) \frac{\partial^2 D_{N,2}(x, t)}{\partial x^2} dy \tag{3.16}\]

\[\frac{\partial}{\partial t} D_{N,1}(x, t) = -d_1(x)D_{N,1}(x, t) + \gamma_{21} D_{N,2}(x, t) + \gamma_{01} D_{N,0}(x, t) \]
\[ + \int_0^1 \gamma_{11}^C(x, y) \frac{\partial}{\partial x} D_{N,1}(x, t) dy + \int_0^1 \gamma_{11}^D(x, y) \frac{\partial^2 D_{N,1}(x, t)}{\partial x^2} dy \tag{3.17}\]

\[\frac{\partial}{\partial t} D_{N,0}(x, t) = -(\lambda_{0,0}(x) + d_0(x)) D_{N,0}(x, t) + 2\lambda_{00} E_0(y, t) D_{N,0}(x, t) \]
\[ + \lambda_{10}(x) E_1(y, t) D_{N,0}(x, t) + \lambda_{10}(x) E_0(y, t) D_{N,1}(x, t) \]
\[ + \lambda_{20}(x) E_2(y, t) D_{N,0}(x, t) + \lambda_{20}(x) E_0(y, t) D_{N,2}(x, t) \]
\[ + \int_0^1 \gamma_{00}^C(x, y) \frac{\partial}{\partial x} D_{N,0}(x, t) dy + \int_0^1 \gamma_{00}^D(x, y) \frac{\partial^2 D_{N,0}(x, t)}{\partial x^2} dy \tag{3.18}\]

\[\frac{\partial}{\partial t} E_2(x, t) = (1 - s_2(x))d_2(x) - d_2(x) E_2(x, t) + \gamma_{12} E_1(x, t) + \gamma_{02} E_0(x, t) \]
\[ + \int_0^1 \gamma_{22}^C(x, y) \frac{\partial}{\partial x} E_2(x, t) dy + \int_0^1 \gamma_{22}^D(x, y) \frac{\partial^2 E_2(x, t)}{\partial x^2} dy \tag{3.19}\]

\[\frac{\partial}{\partial t} E_1(x, t) = (1 - s_1(x))d_1(x) - d_1(x) E_1(x, t) + \gamma_{21} E_2(x, t) + \gamma_{01} E_0(x, t) \]
\[ + \int_0^1 \gamma_{11}^C(x, y) \frac{\partial}{\partial x} E_1(x, t) dy + \int_0^1 \gamma_{11}^D(x, y) \frac{\partial^2 E_1(x, t)}{\partial x^2} dy \tag{3.20}\]

\[\frac{\partial}{\partial t} E_0(x, t) = (1 - s_0(x))d_0(x) - (\lambda_{00}(x) + d_0(x)) E_0(x, t) + \lambda_{00}(x) E_0(x, t) E_0(x, t) \]
\[ + 2\lambda_{10}(x) E_1(x, t) E_0(x, t) + 2\lambda_{20}(x) E_2(x, t) E_0(x, t) \]
\[ + \int_0^1 \gamma_{00}^C(x, y) \frac{\partial}{\partial x} E_0(x, t) dy + \int_0^1 \gamma_{00}^D(x, y) \frac{\partial^2 E_0(x, t)}{\partial x^2} dy \tag{3.21}\]

This parameterization allows for the CTBD model to be adapted to understand AMR in bacteria. In particular, this could provide a tool to quantify the multi-level fitness tradeoffs that AMR is under. It would be quite interesting to apply this model to data and determine how AMR frequency and host environment interact to produce transmission rate.
CHAPTER 4

QUANTIFYING WITHIN AND BETWEEN HOST SELECTION PRESSURES ON HIV-1 USING A CONTINUOUS-TYPE BIRTH-DEATH MODEL

Pathogens can experience varying rates of evolution at multiple scales. Human Immunodeficiency Virus (HIV) is no exception and is characterized by rapid diversification within-host during the acute phase of infection while strains most similar to the population consensus are transmitted at the population scale. Using sequence data from HIV-1 infected hosts, we apply the continuous type birth-death (CTBD) model, developed in Chapter 3, to deter-
mine how the distance to consensus evolves along a phylogenetic tree. This allows us to model how within and between host selection pressures impact the pathogen's evolutionary trajectory. We find correspondence across three distinct HIV-1 data sets that the genome undergoes rapid diversification post-transmission paired with purifying selection at the between-host scale.

4.1 Introduction

Human immunodeficiency virus (HIV) is the cause of human acquired immunodeficiency syndrome (AIDS) and infects millions of individuals every year (UNAIDS 2020). Like any host-pathogen system, HIV is under strong evolutionary pressures. One such pressure is the strict bottlenecks at the time of transmission, where infected individuals are often colonized by a single genetic variant (Wolinsky et al. 1992; Haaland et al. 2009; Carlson et al. 2014). These bottlenecks are complex, with HIV virions being presented multiple barriers throughout a transmission event including transit through host and recipient genital tracts and the novel environment presented in the recipient host (Joseph et al. 2015). This is paired with the understanding that HIV-1 rapidly accumulates genetic diversity throughout the infection process (Coffin 1995). Rapid viral turnover (Ho et al. 1995), high viral loads (Campbell et al. 2003), selection to escape host immune responses (Chen et al. 2004; Neher and Leitner 2010; Bazykin et al. 2006), recombination (Robertson et al. 1995), and lack of proofreading for both Reverse Transcriptase (RT) and RNA polymerase II are all factors in this rapid diversification (Pathak and Temin 1990; Mansky and Temin 1994, 1995; Kim et al. 1996).

There is evidence that selection bias at the between host scale may impact evolutionary dynamics. Sequences that have diverged from the population consensus are transmitted less frequently (Carlson et al. 2014; Herbeck et al. 2006; Raghwani et al. 2018). This suggests that many of the diversifying mutations at the within host scale are deleterious at the between.
host scale. In addition, this body of work demonstrates that the diversity accumulated within a host can be used as an indicator for transmission success at the between host scale.

Regions of the HIV genome that recognize the host immune system are of particular interest as they are likely under strong diversifying selection pressures to escape recognition by the immune system. At the same time not all mutations accumulated by HIV-1 within one host will be advantageous in other hosts. It has been shown that one in three sequence changes revert throughout infection (Zanini et al. 2015). One potential reason is due to HLA polymorphisms in human populations, which modulate the recognition of HIV epitopes by T cells (Bhattacharya et al. 2007). Mismatched HLA types lead to more reversions post infection, a suggestion that escape mutations from previous hosts can be deleterious in a novel host (Li et al. 2007; Leslie et al. 2004). Other examples of regions that are likely under selection against diversification, or purifying selection, are the pol and gag genes that encode essential components to the viral infection process. Without reverse transcriptase and integrase, the viral genome could not be integrated into DNA and in turn transcribed by the host cell’s machinery. Similarly, the functionality of protease and the Gag proteins is of vital importance for viral envelope structure.

A rich field has developed countless techniques to determine which of the three major types of selection, neutral, purifying selection, and diversifying selection as well as the strength of selection that a particular gene or coding region is under. One of the most commonly implemented methods quantifies the ratio of nonsynonymous to synonymous substitutions. This ratio is often referred to as $K_a / K_s$, $dN/dS$, and $\omega$. When $dN/dS$ ratios are greater than 1, less than 1, and equal to 1 they suggest that diversifying selection, purifying selection, and neutral evolution respectively is present at the sites of interest. What these methods do not allow is for a direct quantification of the fitness advantage or disadvantage of diversification at the within and between host scales.

Phylodynamic models provide another potential avenue through which selection pres-
sures can be quantified. The Multi-Type Birth-Death (MTBD) model can allow for the fitness of a particular pathogen type (e.g. a variant with a particular amino acid change) to be estimated in a likelihood-based framework (Stadler and Bonhoeffer 2013). In Chapter 3, I outline an extension of the MTBD model, the Continuous Type Birth-Death model (CTBD), which allows continuous valued traits to evolve along each branch of a tree and shape a lineage's fitness through its birth or transmission rate. Using BEAST 2, we are able to infer both the fitness of a continuous trait at the host level and the parameters associated with within-host evolution from a phylogeny (Bouckaert et al. 2019). Here we apply these methods to quantify the selection pressures acting on HIV-1 at the within and between host scales.

4.2 Methods: CTBD Application

HIV-1 evolutionary dynamics are driven by multiscale processes. Evidence suggests that neutral processes drive between host scale evolution while undergoing rapid diversifying selection at the within host scale post-transmission. This rapid diversification is thought to contribute to a reduction in transmissibility. Therefore, we can use sequence divergence from the population-level consensus as a measure of diversification. Here we apply the CTBD, detailed in Chapter 3, to multiple HIV-1 data sources to elucidate the relationship between this sequence divergence and transmissibility. These data sets come from varying population structures and sizes allowing us to see whether patterns of HIV evolution remain constant or change across these factors.

For each data set we compute the distance to consensus sequence. This is done by first finding the amino acid at highest frequency at each site. Then for each sequence we calculate the Hamming distance (special case of Manhattan distance when all features are binary) between it and the consensus sequence. Gaps (-) and sites classified as an undetermined amino acid (X) do not add to the distance between the two sequences. Note
that for the application to distance to consensus we do not bound the continuous trait below by the minimum number of non-synonymous mutations we have found in a particular data set. Instead, we choose the minimum distance to consensus to be zero, allowing ancestral lineages to be inferred in this state. We do this as theoretical estimates demonstrate that we are considerably better at inferring ancestral state when this assumption is made (Figure 4.15). However, we can still fit the CTBD to the more constrained bounds and we present these results in the appendix (Figures 4.4 and 4.5). This does not impact the qualitative patterns of inference that there is purifying selection at the between-host scale while diversifying selection occurs simultaneously at the within host scale.

Note that by having the continuous trait of interest be distance to consensus sequence, selection for divergence is equivalent to the accumulation of nonsynonymous mutations, or diversifying selection. Selection against divergence at the within host scale however would be demonstrative of purifying selection within hosts. Equivalently, a negative relationship between distance to consensus and fitness at the between host scale would be suggestive of there being purifying selection. A positive relationship at this scale would mean that there is positive directional selection.

Prior to running the full CTBD model we first estimate the base transmission rate under a single type birth-death model. This leads to an average birth rate on which we estimate the relative fitness mapping function mentioned above. Since we are running a simple birth-death model we do not need to define any transition rates. This ensures that the diffusion and selection coefficients that we estimate are independent of this initial base transmission rate estimation. We then run the CTBD where the relative fitness of any particular trait value, $x$, at the between-host scale is captured using a fitness mapping function, $f(x)$. The continuous trait value is assumed to be inherited by those they transmit to such that individuals of type $x$ strictly give birth to individuals of type $x$. The fitness mapping function is parameterized using a cubic splines approach that is flexible enough to capture non-linear relationships (Burden and Faires 1989). Within host dynamics are
captured by the gamma matrix where the continuous valued trait is able to evolve under both diffusive and advective forces, defined as $\gamma^D$ and $\gamma^A$ respectively.

4.2.1 Data

We apply the CTBD model to three HIV-1 subtype B sequence data sets, all sampled from distinct host populations: 1) partial pol genes from King County, WA, 2) full genomes from the Los Alamos National Labs database (www.hiv.lanl.gov), and 3) partial pol genes from men who have sex with men (MSM) in Switzerland.

4.2.2 Phylogenetic Reconstruction

For King County, we obtained sequences of the Protease/Reverse Transcriptase (PR/RT) coding region of HIV-1 infected patients. These sequences are host specific consensus sequences meaning any minority variants are not present in the data. The majority of sequences, 73.3%, came from patients residing in King County, WA. Individuals are sampled from multiple risk groups including heterosexual, men who have sex with men (MSM), and intravenous drug users (IDU). Some patients were sampled multiple times and for these patients we chose to use their first sequence. We do this to increase the probability that the hosts were not on antivirals. This left 880 sequences from the PR/RT coding region. Additionally, we removed any sequences from individuals without viral load data or when the CD4 counts were less than 200 cells per mm$^3$, an indicator used for AIDS diagnosis. To ensure samples were all subtype B we used COMET, an HIV-1 subtype identification tool (Struck et al. 2014). After these final data processing steps we were left with 529 PR/RT sequences. We utilized MAFFT for multiple alignment of the sequences using the Progressive method (Katoh and Standley 2013). Subsequently we used FastTree to obtain maximum likelihood phylogenies assuming a GTR model of sequence evolution with Gamma-distributed rate variation among sites (Price et al. 2010). Finally, to date the maximum likelihood phylogenies we used LSD assuming a molecular clock rate of 2e-3 substitutions per site per year.
Full genome sequences of HIV-1 subtype B infected hosts sampled between 1990 and 2021 were downloaded from the LANL HIV database (www.hiv.lanl.gov) on 3/17/21. There were a total of 3,380 sequences that fit these criteria. These individuals are known to be sequenced in the United States at the time of sampling, a wide geographic range in comparison to the other data sets in this analysis. For individuals with multiple sequences in the database, we chose the earliest sequence sampled from each host to be the representative sequence for said host. Note that only the year of sampling is known. If multiple samples were taken for a single host in the first year of their sequencing, we randomly selected one sequence. This left 335 total sequences. Sequences were aligned using the Progressive method in MAFFT (Katoh and Standley 2013). We then scanned the sequences for significant signals of recombination using RDP4 (Martin et al. 2015). Any sequences with multiple recombination tests finding a significant signal of recombination were analyzed and removed if this paired with a phylogenetic signature of recombination. This left a total of 328 sequences. We then partitioned the remaining sequences into the four major genomic regions of HIV-1: \textit{gag}, \textit{pol}, \textit{env} and all accessory genes. Using RAxML we obtained the maximum likelihood phylogenies for these regions as well as the full genomes assuming a GTR model of sequence evolution with Gamma-distributed rate variation among sites (Stamatakis 2014). To date the maximum likelihood phylogenies we used LSD assuming a molecular clock rate of 3e-3 substitutions per site per year (To et al. 2016).

Here we show a tanglegram, a visual representation of the concordance between the phylogenies, for the four genomic regions of interest (Figure 4.7). There is significant variability in the phylogenetic relationship amongst the sequences when splitting by major regions. This demonstrated the importance that we perform analyses on each region separately.

The Swiss data set consists strictly of HIV-1 subtype B sequences isolated from men who have sex with men (MSM) (Kouyos et al. 2010). This is a particularly large data set with 4103 partial \textit{pol} sequences. To reduce computational time we subsampled the data into
four groups of 500 tips and fit the CTBD to each independently. The distance to in-sample consensus across time is presented in Figure 4.6. The Swiss sequence data is not publicly available, so we relied on previously reconstructed maximum likelihood trees (Rasmussen et al. 2017).

4.3 Results

Using the CTBD we estimate how the fitness of a viral lineage depends on the distance to consensus, a measure of how many non-synonymous amino acid mutations have accumulated in reference to the population consensus sequence. We do so by estimating how the strength of selection pressures at both the within and between host scales act upon this continuous trait. We find that across all data sets and genomic regions, the CTBD infers that ancestral lineages deeper in the past become more similar to the population consensus (Figure 4.1a). We performed an independent analysis for each data set using PASTML to reconstruct ancestral sequences along the phylogeny and in turn calculate the distance to consensus for ancestral lineages (Ishikawa et al. 2019). We find the same pattern using this approach, with ancestral estimates for the distance to consensus approaching zero as we reach the root (Figure 4.1b). This correspondence occurs across all three of our data sets even though they all have different signals of distance to consensus across time (Figures 4.8, 4.6, 4.9).
Figure 4.1: Left: Estimates of the distance to consensus for internal nodes. Blue points represent true tip state while red points represent the inferred median state using the CTBD. Bars represent the 95% HPI. Right: Maximum likelihood reconstructed phylogeny for the sequenced protease and reverse transcriptase coding regions of HIV-1 infected hosts in King County, WA. Tips and branches are colored according to observed or reconstructed distance to consensus of the host. Distance to consensus for ancestral nodes and branches were reconstructed using PastML. (Ishikawa et al. 2019).

Distance to consensus is estimated to be under diversifying selection at the within-host scale in all three data sets. Viral lineages diverge from the consensus at a median rate of 2.62, 0.16, and 1.37 amino acid substitutions per year in the LANL, King County, and Swiss data sets respectively (Figures 4.5a-4.14a). All of these estimates have 95% credible distributions only contain positive values for the rate of divergence. When comparing the results for the pol gene across all three data sets we find that King County data set has the smallest estimated within host selection pressure while the Swiss and LANL pol regions have comparably high within-host selection coefficients (Figure 4.14a).

We estimate there to be a negative relationship between distance to consensus and population level fitness, what we call the transmission cost to divergence, for all three data sets. Individuals who are infected with a strain of HIV-1 more similar to the consensus transmit at a higher rate. This general relationship remains consistent across various parts
of the genome as well as across our geographically disparate data sets. After partitioning the LANL data set into four major genomic regions (gag, env, pol, and all accessory genes) we find that all four regions have their peak fitness somewhere in the range of 10-25 amino acid substitutions from the consensus, after which fitness declines with larger distances. The confidence bounds for all regions are quite wide when near the lower bound, where the distance to consensus is zero. This is likely due to no sequences being sampled in this range.

Next, we estimated the transmission cost to divergence for the full genome and find that the peak is further away from the consensus than the four genomic regions of interest (Figure 4.3). It is very likely that at the full genome scale purifying selection is relaxed in non-coding and other less conserved regions, allowing for more divergence. We can compare how the fitness cost to divergence varies across the other genomic regions. In order of largest to smallest drop we find that gag and pol have the most precipitous drops followed by the accessory genes, and finally env (Figure 4.2a). When comparing the estimates for the fitness cost to divergence across the different host populations where we have the pol gene we find that the Swiss data set has the most exponential like decay in transmissibility while both LANL and King County have their respective max transmission rates at approximately 10-20 amino acid substitutions from the consensus sequence (Figure 4.2b).
4.4 Discussion

In Chapter 3 we presented a theoretical framework for the quantification of how a continuous valued trait of interest evolves along a phylogeny. We developed a multiscale evolutionary model where inference can be conducted on how a continuous trait evolves at the within and between host scale. Within host processes are allowed to change without directional bias through diffusion while we simultaneously estimate directional processes driving positive or negative selection. In this Chapter we apply this model to understand the selection pressures at multiple scales. We do so by inferring parameters associated with multiscale evolution for three data sets of HIV-1 subtype B infected hosts. In addition, we partition whole genome sequences from one of these data sets into multiple regions of interest that may be under variable selection pressures. The CTBD allows us to test and
compare how the distance to consensus impacts evolution across the genome.

Multiscale evolutionary pressures drive the phylogenetic structure and epidemiological dynamics of pathogens. This work explores this evolutionary mismatch at the within and between host scales by leveraging the phylogenetic history of HIV-1 subtype B infected hosts. We find there to be positive selection, or rapid diversification occurring at the within-host scale while purifying selection acts simultaneously at the between-host scale. This agrees with previous work showing that HIV-1 undergoes rapid diversification during the acute phase of infection due to evolutionary processes associated with immune escape (Henn et al. 2012; Liu et al. 2012). The direction of selection at the between host scale is also consistent with previous work which has shown that founder strains are more similar to the population level consensus than expected by chance (Herbeck et al. 2006; Carlson et al. 2014; Raghwani et al. 2018). While previous work has focused extensively on the biological mechanisms favoring transmission of more conserved variants, our work allows us to quantify the transmission cost to divergence at the host-population level.

We were able to directly compare the direction and strength of selection at multiple scales across the genome. We find that purifying selection is ubiquitous at the between-host scale, but is weaker in \textit{env} when compared to \textit{gag} and \textit{pol}. This coincides with previous work that has shown purifying selection is acting on \textit{gag} and \textit{pol} while the \textit{gp120} and \textit{gp41} genes that encode for \textit{env} are under diversifying and purifying selection respectively (Seibert et al. 1995b; Novitsky et al. 2009; Seibert et al. 1995a). The variability in selection on \textit{env} is likely due to how the \textit{gp120} glycoproteins sit exposed on the envelope spike while the \textit{gp41} glycoproteins remain tucked away closer to the infected cell's membrane. Thus, one reason we estimate \textit{env} to be under weaker purifying selection due to this mixture of selective processes on the region. In the future we hope to explore this pattern more in depth.

When comparing our analyses across data sets we find a strong signal of purifying selection at the between-host scale, although the precise functional form does vary. It is possible that some of the variability we find is due to differences in demographics and
risk behavior across host populations. The Swiss data set for example consists entirely of MSM individuals, a group known to be at higher risk for HIV-1 infection. The LANL and King County data sets on the other hand have an unknown mixture of different risk groups. This could explain some of the variability in selection pressures observed between these populations. Another pattern of interest is that we see the largest fitness cost to divergence in the HIV-1 data set from the MSM infected individuals in Switzerland. It has been noted previously that the population consensus has a higher likelihood of being transmitted to male recipients, however this study was done in heterosexual couples (Carlson et al. 2014). Another potential explanation for the inferred transmission cost to divergence is the potential correlation between stage of infection and divergence. If individuals are sampled in the acute phase of infection then the sequenced viruses will have less time for within-host evolution to drive diversification. If acutely infected individuals transmit at a higher rate then there could be a pattern of transmission cost to divergence due to this correlation. Since viral loads reach their maximum during the acute phase then it is possible that this pattern could be perceptible by jointly plotting an individual's distance to consensus and sampled viral load. We find there to be no correlation between these two traits with a $R^2 = 1.5e - 5$ (Figure 4.12).

Of note, the three pol phylogenies were reconstructed using different portions of the pol gene. The LANL data set contains full pol sequences while the King County, WA and Swiss data sets only consist of the PR/RT protein coding regions. This means that the LANL data set also holds the Integrase protein coding region of pol, a region that contains a cluster of CD4 T cell epitope positions on its 3’ end which have a particularly high amino acid diversity at the between-host scale when compared to the rest of the pol gene (Li et al. 2015). This may contribute to the less dramatic transmission cost to divergence that we infer for the LANL data set.

One potential factor that could shape the estimates for transmission/fitness costs of divergence is the natural branching structure of HIV phylogenies where long external
branches lead to most branching events occurring in the deep past. This can be particularly difficult for the CTBD as no novel information is presented on long branches. Pathogens with this type of branching pattern could benefit from additional sequence data from the same infected host across time. This would provide more data on how within host processes are driving the continuous trait’s evolution which would increase model accuracy. However, the work presented here provides an exciting technique to quantify multiscale evolution when this type of sequence data is unavailable or prohibitively expensive. Here we demonstrate that using sequences from HIV-1 infected hosts provides substantial information about within and between-host evolutionary processes.

In the future we also hope to apply this same technique to other host-pathogen systems where the distance to consensus is thought to be a phenotype on which selection acts. Influenza is known to be under strong selective pressures to escape host immunity at epitope sites while the majority of mutations outside of epitope sites are thought to be deleterious (Pybus et al. 2007; Koelle and Rasmussen 2015; Łuksza and Lässig 2014). The CTBD could be used to quantify the discrepancy in transmission costs at epitope versus non-epitope sites.
4.5 Appendix

Figure 4.3: Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the LANL data set. Median estimates are surrounded by the 95% credible interval. We first estimated the background birth rate prior to running the full CTBD. The median fitness for \textit{env} and \textit{pol} at the population consensus remains above one, only suggesting that these viral lineages transmit more than the average transmission rate expected under a simple birth-death process.
Figure 4.4: Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the LANL data set. Minimum distance to consensus values were used across each region independently for the lower bound during analysis with the CTBD. Median estimates are surrounded by the 95% credible interval.
Figure 4.5: Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the LANL data set. Minimum distance to consensus values were used across each region independently for the lower bound during analysis with the CTBD. Black vertical lines represent the median along with the 95% credible interval.
Figure 4.6: Distance to in-sample consensus measured across time for the subsampled phylogenies with 500 tips derived from the larger Swiss HIV-1 MSM data set. Each color is representative of a different phylogeny.
Figure 4.7: Tanglegram for the four genomic regions of interest from the LANL data set. From left to right: *gag, pol, accessory, and env*. Colors are representative of relative location shift across phlogenies in reference to *gag* gene.
Figure 4.8: Distance to in-sample consensus measured across time for the protease and reverse transcriptase protein coding regions found in 529 HIV-1 positive individuals from King County, WA.
Figure 4.9: Distance to in-sample consensus measured across time for the four regions of interest and the full genome sequences from the LANL data set.
Figure 4.10: Posterior estimate of the between host distance to consensus to fitness mapping function inferred using the CTBD for the regions of interest from the Swiss data set. Median estimates are surrounded by the 95% credible interval.
Figure 4.11: Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the Swiss data set. Black vertical lines represent the median along with the 95% credible interval.

4.5.1 LANL

Figure 4.12: Scatterplot of distance to in-sample consensus and viral load (log10 viral copies per ml) with color corresponding to their sampled date for the LANL data set. We find there to be no correlation between these two traits with a $R^2 = 1.5e-5$. 

(a) Posterior estimate for the within-host selection coefficient.

(b) Posterior estimate for the within-host diffusion coefficient.
Figure 4.13: Posterior estimates for the selection and diffusion parameters of the CTBD for the regions of interest from the LANL data set. Black vertical lines represent the median along with the 95% credible interval.

4.5.2 Pol

Figure 4.14: Posterior estimates for the selection and diffusion parameters of the CTBD for the partial or full pol genes from the three data sets. Black vertical lines represent the median along with the 95% credible interval.
4.5.3 Additional Figures

(a) Comparison between inferred and truth when we bound the CTBD below by the minimum threshold found in the data.

(b) Comparison between inferred and truth when we bound the CTBD below at a location significantly below the minimum sampled location.

Figure 4.15: Comparison between true continuous state and estimate from CTBD at internal nodes. Blue points represent the true state of internal nodes, green points represent true tip state, while red points represent the inferred median state using the CTBD. Bars represent the 95% credible interval.
In the previous three chapters I have presented two mathematical frameworks that provide insight into the evolutionary dynamics of host pathogen systems. There are clear and obvious implications for the study of these systems using mathematical models and here we hope to contribute to this flourishing and vitally important field. Chapter 2 is based off of theoretical population genetic models of evolving populations, developed almost 100 years ago. We extend this work to model multi-scale evolution that allows us to approximate the probability of and expected time it will take a novel strain to spread and rise to high frequencies in a population of infected hosts. In Chapter 3 we develop the machinery for a novel phylodynamic method to allow for inference of how a continuous trait of interest evolves along a phylogenetic tree. Finally, in Chapter 4 we apply this model to data from Human Immunodeficiency Virus (HIV) infected hosts and estimate important evolutionary
and epidemiological parameters.

5.1 Summary of Results

The spread of an emerging pathogen is a multifaceted process where epidemiological, immunological, and evolutionary factors are simultaneously at play. One such example is during the emergence of a novel mutation in a population of infected hosts where both demographic changes throughout infection at the within-host level and transmission at the between-host level can impact the evolutionary outcome. For example, two of the most costly infectious diseases, HIV-1 and influenza A, are thought to have strict bottlenecks (Joseph et al. 2015; McCrone et al. 2018). Small bottlenecks are known to impact the probability that beneficial and deleterious mutations spread and become fixed in a population (Wahl and Gerrish 2001; Wahl et al. 2002; Bergstrom et al. 1999).

In Chapter 2 we implement a Wright-Fisher model at both the host and the population level to allow for stochasticity to drive the evolutionary dynamics. This is particularly important immediately after transmission when viral populations are initially small and genetic drift can lead to rapid changes in allele frequencies. We analyze how key parameters including this bottleneck size, time between transmission events, and selection at multiple scales impact probability of fixation and time to fixation. These metrics are a good measure of how quickly and how likely a pathogen can adapt.

We found that both host population and bottleneck size have a large role in determining probability of fixation. This further demonstrates the importance of quantitative studies that aim to measure the precise number of viral particles that establish infection. Assuming a mutant strain is neutral, then smaller bottlenecks leads to a higher likelihood that the mutant fixes in the population. This remains the case for beneficial mutations, however other factors begin to impact this relationship. In particular we find that the functional relationship that maps the frequency of the mutant strain within a host to the between
host rate of transmissibility can change the probability of fixation by an order of magnitude when host level selection pressures are strong.

The main driver of the expected time to fixation for a novel strain is the number of infected hosts in the population. This is due to the fact that this usually outnumbers the viral bottleneck size by several orders of magnitude. Therefore, the more transmissible a novel variant is the faster it can become fixed in the population. This is something we have seen multiple times throughout the spread of SARS-CoV-2. The first of these non-synonymous mutations that rose to high frequencies was D614G, which has become fixed across the globe (Korber et al. 2020). Since then other variants like B.1.1.7, B.1.351, and now B.1.617.2 (also referred to as (Alpha, Beta, and Delta) have emerged into larger infected population sizes which has led to a slower rise in frequency (Rambaut et al. 2020; Tegally et al. 2020; Hodcroft 2021). It is likely that all of these variants had some positive impact on transmissibility, or equivalently were under strong positive selection at the host scale.

In Chapter 3 we continue to explore some of the same overarching themes of multi-scale evolution and host-pathogen dynamics as found in Chapter 2. Pathogen strains can have varying levels of fitness at the within and between host scales. Often times these variants are under antagonistic selection pressures, with benefits being incurred at one scale paired with detrimental effects at another. These changes in fitness can impact the ability to compete at the within host scale or by altering the host's ability to transmit.

In Chapter 3 we developed a phylodynamic birth-death model that extends the multi-type birth-death models (Maddison et al. 2007; FitzJohn 2012; Kühnert et al. 2016; Stadler 2013; Barido-Sottani et al. 2020). We aimed to develop an extension that could model the evolution of a continuous-valued trait like allele frequency or viral load. Using a system of partial differential equations, we can compute the likelihood of the phylogeny evolving as observed given a function that maps the value of a continuous trait to a lineage's fitness in terms of its birth or death rate. This likelihood based approach allows us to infer the magnitude and direction of selection pressures within and between hosts along with other
relevant parameters from phylogenetic trees. The CTBD was then implemented in BEAST 2, a Bayesian phylogenetic program. Using this approach we apply the model to both simulated and real data sets.

We then applied the CTBD model to a population of HIV-1 infected hosts from whom we have paired partial pol sequences and viral load counts. This allows us to quantify the relationship between viral load and between-host transmission fitness for HIV-1. We estimate that transmission rates are optimized at intermediate set point viral load (SPVL) values of approximately $10^5$ virions. We hypothesize that the inferred nonlinear relationship between SPVL and fitness arises due to fitness trade-offs between virulence and transmission although we are unable to directly demonstrate this. In particular, increasing SPVLs will increase transmissibility while simultaneously increase the speed of host removal from the infectious class (Alizon et al. 2009). This removal from the infectious class could be due to experiencing more severe symptoms which reduce contact rates or by inducing death. This is the first phylodynamic method used to estimate the impact of SPVL on transmission rates. However, we are not the first method to infer that intermediate SPVLs have a higher likelihood of transmitting (Blanquart et al. 2016).

Pathogens can experience varying selection pressures at multiple scales. Human Immunodeficiency Virus (HIV) is no exception and is characterized by rapid diversification within-host during the acute phase of infection while strains most similar to the population consensus are transmitted at the population scale. In Chapter 4 we apply the CTBD, developed in Chapter 3, to determine how the distance to consensus evolves along a phylogenetic tree. This allows us to model how within and between host selection pressures impact the pathogen's evolutionary trajectory. We run this model on multiple genomic data sets of individuals infected with HIV-1. Across three distinct HIV-1 data sets we consistently infer that the genome undergoes rapid diversification post-transmission paired with purifying selection at the between-host scale. Viral lineages with sequences closer to the population consensus are considerably more transmissible than more divergent sequences and this
pattern is consistent across the envelope proteins as well as structural, and accessory genes. Notably we find that *gag* and *pol* are under the strongest purifying selection followed by the accessory genes and *env*.

### 5.2 Future Work

#### 5.2.1 Chapter 2

We develop a quantitative framework to study how population dynamics alter evolutionary dynamics in viruses across scales. One potential future direction would be to incorporate a non-stationary infected host population. We know that in a classic Wright-Fisher model, exponential growth increases the probability that a positively selected for mutation becomes fixed (Otto and Whitlock 1997). However, quantifying how this relationship changes in a multi-scale model could provide insight into how more realistic infection dynamics changes adaptation in host pathogen systems.

Another prospect is to pair the model with a likelihood-based inference from Tataru et al. (2017) to estimate parameters from allele frequency data (Tataru et al. 2017). One immediate roadblock for this approach is in the rarity of necessary data. In particular, allele frequencies would have to be sampled at least twice throughout infection in multiple hosts in order to inform parameters associated with within host selection. Assuming one could obtain a data set capable of providing sufficient information, it is possible that an analytical expression for the likelihood function under the model may be intractable. Instead I believe that pairing sufficient statistics from simulation results with machine learning approaches could lead to parameter estimates. Previously, Approximate Bayesian Computation (ABC) has shown some promise in this field (Beaumont et al. 2002; Joyce and Marjoram 2008).
5.2.2 Chapter 3

There are many potentially fruitful directions for future work. For one, we hope to directly test for the virulence transmission trade-off that we allude to above. This can be done by allowing the death rate to be dependent on SPVL and perform simultaneous inference on both the birth and death rate functional forms. It is likely that we will have to simplify the functional relationships as strong correlations will be present amongst these parameters. Applying this model to other data sets where SPVL is fully characterized and paired with sequence data would provide additional support for this relationship. Although this data is rare, I believe that phylodynamic methods like these have a bright future in answering a plethora of interesting evolutionary and epidemiological questions. Lastly, incorporating multiple evolving traits that can be any combination of continuous and discrete could provide valuable insight. For example, we could potentially begin to untangle how risk factors along with SPVL shape dynamics.

In addition, we have developed multiple extensions for the CTBD in Section 3.6.4 of Chapter 3 that have not been developed and incorporated into BEAST 2. This will allow us to perform inference on simulated data, giving us the ability to diagnose whether these additional methods could be used in real world applications.

5.2.3 Chapter 4

Future directions delve into additional applications of the CTBD to quantifying the relationship between viral sequence divergence and fitness. For example, Carlson et al. (2014) show that additional factors like donor SPVL and recipient gender impact the selection bias at transmission. Using extensions of the CTBD could allow for these factors to impact transmission fitness. In addition, the application of the CTBD to other host pathogen systems could provide more general insights into what factors primarily determine pathogen fitness. Influenza for example also presents a mismatch of selection biases at the within
and between host scales, with evolution appearing essentially neutral within hosts, but diversifying selection driving evolution at the between-host scale. This is further confounded by the fact that influenza has a few epitope sites where mutations can lead to antigenic drift while the majority of nonsynonymous mutations in non-epitope sites are deleterious in nature (Pybus et al. 2007; Koelle and Rasmussen 2015; Łuksza and Lässig 2014). The CTBD phylodynamic model could quantify the fitness cost to accumulating deleterious mutations in terms of transmission potential as we did for HIV.
REFERENCES


115


Stadler, T. (2013). How can we improve accuracy of macroevolutionary rate estimates?


A summary of all acronyms is documented in Table A.1.

Table A.1: A summary of acronyms used in alphabetical order.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acquired Immunodeficiency Syndrome</td>
<td>AIDS</td>
</tr>
<tr>
<td>Antimicrobial Resistance</td>
<td>AMR</td>
</tr>
<tr>
<td>Anti-Retroviral Therapy</td>
<td>ART</td>
</tr>
<tr>
<td>Cluster of Differentiation 4</td>
<td>CD4</td>
</tr>
<tr>
<td>Continuous-Type Birth-Death</td>
<td>CTBD</td>
</tr>
<tr>
<td>Kernel Density Estimate</td>
<td>KDE</td>
</tr>
<tr>
<td>Human Immunodeficiency Virus</td>
<td>HIV</td>
</tr>
<tr>
<td>Integrase</td>
<td>IN</td>
</tr>
</tbody>
</table>
Table A.1: A summary of acronyms used in alphabetical order.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Markov Chain Monte Carlo</td>
<td>MCMC</td>
</tr>
<tr>
<td>Multi-Type Birth-Death</td>
<td>MTBD</td>
</tr>
<tr>
<td>Partial Differential Equation</td>
<td>PDE</td>
</tr>
<tr>
<td>Protease gene</td>
<td>PR</td>
</tr>
<tr>
<td>Pre-Exposure Prophylaxis</td>
<td>PrEP</td>
</tr>
<tr>
<td>Reverse Transcriptase gene</td>
<td>RT</td>
</tr>
<tr>
<td>Set Point Viral Load</td>
<td>SPVL</td>
</tr>
<tr>
<td>Treatment As Prevention</td>
<td>TasP</td>
</tr>
<tr>
<td>Viral Load</td>
<td>VL</td>
</tr>
</tbody>
</table>
A summary of all variables is documented in Table B.1.

Table B.1: A summary of parameters, variables, and metrics followed by their definitions in alphabetical order. The chapter in which they are utilized is also noted.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Fitness Within Infected Host</td>
<td>( \bar{w} )</td>
<td>3</td>
</tr>
<tr>
<td>Backward Transition Rate</td>
<td>( \delta^b(x) )</td>
<td>3,4</td>
</tr>
<tr>
<td>Base Transmission Rate</td>
<td>( \beta_0 )</td>
<td>3,4</td>
</tr>
<tr>
<td>Death Rate</td>
<td>( d(x), d_i )</td>
<td>3,4</td>
</tr>
<tr>
<td>Deterministic Pathogen Generations</td>
<td>( g_d )</td>
<td>2</td>
</tr>
<tr>
<td>Effective Population Size</td>
<td>( N_e )</td>
<td>2,3</td>
</tr>
<tr>
<td>Forward Transition Rate</td>
<td>( \delta^f(x) )</td>
<td>3,4</td>
</tr>
</tbody>
</table>
Table B.1: A summary of parameters, variables, and metrics followed by their definitions in alphabetical order. The chapter in which they are utilized is also noted.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Parameter Tensor</td>
<td>$\theta$</td>
<td>3</td>
</tr>
<tr>
<td>Generations to Fixation</td>
<td>$\bar{t}_1$</td>
<td>2</td>
</tr>
<tr>
<td>Host Level Selection Coefficient</td>
<td>$s_b$</td>
<td>2</td>
</tr>
<tr>
<td>Infected Host Population Size</td>
<td>$N_h$</td>
<td>2</td>
</tr>
<tr>
<td>Initial Frequency of Mutant Allele</td>
<td>$p$</td>
<td>2</td>
</tr>
<tr>
<td>Knot Count For Splines</td>
<td>$k$</td>
<td>3,4</td>
</tr>
<tr>
<td>Pathogen Bottleneck Size</td>
<td>$N_w$</td>
<td>2</td>
</tr>
<tr>
<td>Phylogenetic Tree</td>
<td>$\mathcal{T}$</td>
<td>3</td>
</tr>
<tr>
<td>Probability of Fixation</td>
<td>$\pi$</td>
<td>2</td>
</tr>
<tr>
<td>Selection Coefficient</td>
<td>$s$</td>
<td>2,3</td>
</tr>
<tr>
<td>Stochastic Pathogen Generations</td>
<td>$g_s$</td>
<td>2</td>
</tr>
<tr>
<td>Transition from Advection</td>
<td>$\gamma^A(x)$</td>
<td>3,4</td>
</tr>
<tr>
<td>Transition from Diffusion</td>
<td>$\gamma^D(x)$</td>
<td>3,4</td>
</tr>
<tr>
<td>Transition Kernel/Matrix</td>
<td>$\gamma, \gamma_{i,j}$</td>
<td>3,4</td>
</tr>
<tr>
<td>Transmission Kernel/Matrix</td>
<td>$\lambda(x, t), \lambda_{i,j}$</td>
<td>3,4</td>
</tr>
<tr>
<td>Within-Host Level Selection Coefficient</td>
<td>$s_w$</td>
<td>2, 3</td>
</tr>
</tbody>
</table>