Copyright
by
Lauren Ann Castro
2019

The Dissertation Committee for Lauren Ann Castro
certifies that this is the approved version of the following dissertation:

Modeling Uncertainty in Pathogen Transmission and Evolution for Infectious Disease Management

Committee:

Lauren Ancel Meyers, Supervisor

James J. Bull

Thomas Leitner

Claus O. Wilke

Modeling Uncertainty in Pathogen Transmission and Evolution for Infectious Disease Management

by

Lauren Ann Castro

DISSERTATION
Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN
May 2019

I dedicate this dissertation to my parents, who instilled in me the values of hard work and curiosity, and to all PhD students pursuing the first doctoral degree in their family.

Acknowledgments

The research in this dissertation would not have been possible without the support and encouragement of many individuals. First, I want to express my sincere gratitude to my PhD advisor, Dr. Lauren Ancel Meyers, for her mentorship and academic guidance during my dissertation. She gave me the freedom to develop my research interests while providing a wealth of new ideas, expertise in research design, and help in scientific communication.

I next want to thank my research collaborators, Dr. Thomas Leitner, Dr. Ethan Romero-Severson, Dr. Trevor Bedford, and Dr. Ned Dimitrov. My experiences with each of them have demonstrated the generosity and collegial nature of collaboration that I hope to emulate in my career. I am deeply grateful for their involvement in my projects. I would like to thank my committee members, Dr. Claus Wilke and Dr. James Bull for the time and effort they put into my research. Their insightful comments have greatly improved how I present and visualize science.

I'd like to thank former and current members of the Meyers Research Lab for stimulating discussions and collaborations. In particular, I would like to thank former PhD student Spencer Fox for being a steady source of advice in the lab, a dependable collaborator, and friend. I'd also like to thank Ravi Srinivasan, Jose Luis Herrera, and Steve Bellan for their encouragement and feedback early on in my PhD career. My conversations with them gave me the confidence to navigate the transition to graduate school.

This PhD would not have occurred without the support of key individuals from before my time at UT. I'd like to thank Dr. Alina Desphande for her mentorship during my post-baccalaureate year at Los Alamos National Laboratory and the many letters of recommendation she wrote for me that helped me secure my fellowships. Similarly, I would like to thank Dr. Andrew Dobson at Princeton University for first exposing me to the field of disease ecology and mathematical modeling.

I have been privileged to receive financial support for my research from a number of generous groups: The College of Natural Sciences Dean's Office, the Department of Defense through a National Defense Science and Engineering Graduate Fellowship, and the University of Texas Graduate School through Graduate Dean's Prestigious Fellowship Supplements. These funding sources provided the ability to engage in research right away and maintain steady progress throughout the five years of my PhD. I am also grateful to the Integrative Biology Department for the Startup Grant and Travel Award funding I received to travel to workshops and present at conferences. These experiences were critical for connecting with collaborators that contributed to this research. My acknowledgements would not be complete without thanking Tamra Rogers for helping me navigate the administrative intricacies of graduate school.

I would like to acknowledge the friends and family that have provided companionship, encouragement, advice, and rejuvenating breaks from research during the course of this dissertation. First I'd like to thank my parents, Richard and Debra Castro. I owe my successes in part to their sacrifices and never-ceasing belief in me. I also want to thank my sister, Amanda Castro, and brother-in-law Michael Brand, for always encouraging a healthy work-life balance. Thank you for the many visits to Austin – each one has been memorable and a highlight of my time in this city.

I want to thank all the members of the Flamencura Dance Studio for providing me with a creative home over the past five years. I will miss this group of wonderful women for their beauty and strength.

Finally, I owe a special thanks to my husband, Cole. Since we met on the sixth floor of Patterson five years ago you have been a continual source of laughter (puns), understanding, and partnership. You have helped me push myself beyond my perceived limits in all facets of life. This dissertation was undoubtedly improved because of your love and support.

Modeling Uncertainty in Pathogen Transmission and Evolution for Infectious Disease Management

Publication No. ____________________

Lauren Ann Castro, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Lauren Ancel Meyers

Public health practitioners rely on epidemiological case count and molecular sequence data for making decisions regarding the prevention, control, and treatment of infectious disease. The relationship between what practitioners can observe and what they wish to know, (i.e. the epidemiological or evolutionary state of the population) can be uncertain because of gaps or biases in the data collection and interpretation steps. In this dissertation, I integrate simulations of disease dynamics with statistical frameworks to link such observational data with probabilistic statements about the range of underlying possible outcomes. Chapter 2 addresses the need for real-time estimates of local Zika epidemic risk during an unfolding outbreak to inform county-specific public health response plans. In this chapter, I present a quantitative framework for estimating real-time ZIKV risk that captures uncertainty

in case reporting, importations, and vector-human transmission dynamics. I find that accurate estimates of the case reporting rate can reduce the uncertainty in perceived epidemic risk. In addition, local differences in both the environmental suitability for ZIKV transmission and in the numbers of new cases arriving influence how long a policy maker can wait before implementing response efforts to curb a growing epidemic. In Chapter 3, I address the theoretical and practical questions of how early can population level antigenic evolution of seasonal influenza A/H3N2 be predicted, and what are the best data metrics to inform such predictive models. Using a detailed simulation model of seasonal influenza transmission and evolution, I fit predictive logistic regression models of antigenic variant success to epidemiological and population genetic predictors derived from theoretical case and molecular data. The results show that the relative transmission rates of newly emerging influenza variants can robustly indicate future epidemic threats, up to 10 months prior to their widespread expansion. This chapter demonstrates that the early detection of emerging influenza viruses is limited by a tight race between the typical dynamics of antigenic turnover and the annual timeline for vaccine development. Chapter 4 examines how complex evolutionary processes, such as recombination and latency, leave detectable signals in HIV-1 molecular data when analyzed using methods that assume the absence of these processes. First, I develop a new method for simulating the evolutionary history of a set of molecular sequence samples using ancestral recombination graphs (ARG). Next, I use a new statistical framework for comparing simulated and observed

HIV-1 trees. I find evidence that the intensity of within-host evolutionary processes is detectable in binary trees constructed using hierarchical clustering methods. Furthermore, the latent reservoir size is likely to differ between individual patients. This model represents an important advance in the realism of HIV within-in host evolutionary modeling. Altogether, the results in this dissertation demonstrate how effective infectious disease management can be improved by using an interdisciplinary approach of computational epidemiology and statistics to strategically think about complex data collection and interpretation questions.

Table of Contents

Acknowledgments **v**

Abstract **viii**

List of Tables **xiv**

List of Figures **xviii**

Chapter 1. Introduction **1**

1.0.1 Thematic Overview 1

1.0.2 Chapter Summaries 4

1.0.3 Data Logistics and Challenges of Simulation Modeling . 9

1.0.4 Concluding Remarks 12

Chapter 2. Assessing real-time Zika risk in the United States **14**

2.1 Abstract 14

2.2 Introduction 15

2.3 Methods 18

2.3.1 Estimating Importation Rates 18

2.3.2 County Transmission Rates (R_0) 20

2.3.3 ZIKV Outbreak Simulation Model 23

2.3.4 Simulations 25

2.3.5 Outbreak Analysis 26

2.4 Results 27

2.5 Discussion 33

Chapter 3. Early Prediction of Antigenic Transitions for Influenza A H3N2 38

3.1 Abstract 38

3.2 Introduction 39

3.3 Methods 42

  3.3.1 Simulation Model 42

  3.3.2 Simulated Data 43

  3.3.3 Classifying Evolutionary Outcomes 44

  3.3.4 Predictive Models 44

  3.3.5 Candidate predictors 46

3.4 Results 48

3.5 Discussion 61

Chapter 4. Within-host HIV-1 biological processes reflected in deconstructed Ancestral Recombination Graphs 71

4.1 Abstract 71

4.2 Introduction 72

4.3 Methods 77

  4.3.1 Methodological overview 77

  4.3.2 Model Assumptions 77

  4.3.3 Simulation of the ARG 78

    4.3.3.1 Evolutionary Events 79

  4.3.4 Obtaining a Distance Matrix from the ARG 83

  4.3.5 Data 85

  4.3.6 Distance Function and Parameter Inference 86

4.4 Results 90

  4.4.1 Additional biological complexity results in trees consistent with HIV-1 phylogenies 90

  4.4.2 Evidence of biological processes in topological and distance statistics 93

    4.4.2.1 Recombination countervails the effects of latency 95

  4.4.3 Application to sequence data from HIV-1 infected individuals 98

    4.4.3.1 Distance algorithm recovers empirical tree features 98

4.4.3.2 Distance scores are low when N_L is small 100
4.5 Discussion 107

Appendices 113

Appendix A. Assessing real-time Zika risk in the United States 114
A.1 Extended Methods 114
A.1.1 Fitting the Generation Time 114

Appendix B. Early Prediction of Antigenic Transitions for Influenza A H3N2 123
B.1 Extended Methods 123
B.1.1 Influenza Phylodynamic Simulations 123
B.1.2 Candidate Predictors 124
B.2 Extended Results 126
B.2.1 Incorporation of historical data 126
B.2.2 Surveillance Threshold Results 127
B.2.3 Proxy Model Results 129
B.3 Alternate Surveillance Paradigms 131

Appendix C. Within-host HIV-1 biological processes reflected in deconstructed Ancestral Recombination Graphs 138
C.1 Extended Figures 138
C.2 Extended Tables 143

Bibliography 146

Vita 179

List of Tables

2.1 **Stochastic ZIKV outbreak model parameters.** We hold the disease progression parameters constant across all scenarios, estimate R_0 and importation rate for each individual county, and vary the reporting rate to investigate its impact on the uncertainty of ZIKV risk assessments. 22

2.2 **Import risk model variables.** These 10 variables were selected from 72 variables using a combination of representative variables selection and predictive backwards selection. The importance of each variable (from top to bottom) is determined by order of exclusion in backwards selection, with the most important variables remaining in the model the longest. 28

3.1 **Predictors selected by five-fold cross validation and forward selection.** The top four variables were selected in the identical order (as listed) across all surveillance threshold models. The fifth predictor, relative variance in transmissibility, was included in all models, but not always as the fifth chosen. In the formulas, c refers to cluster-level quantities. The rightmost column gives the full range of fitted coefficients (log-odds) across all models based on across the five-fold cross validation for each surveillance thresholds' final model. var(σ_c) was calculated across all hosts; var($\mathrm{S}_{\mathrm{eff}}$) was calculated across only infected hosts. I = number of infected hosts, N = total number of hosts, σ_c = effective susceptibility to infection by cluster c, β^{k} = the transmission rate of the virus carrying k deleterious mutations.Formulas to calculate each quantity are in Table B.1. 51

3.2 **Model 1 predicts the fate of a cluster using the top two predictors in our best fit model.** The two proxy models use data from two time points, when the cluster reached relative frequencies of 6% (t_1) and 10% (t_2). Model 2 considers the time elapsed between and the number of competing expanding clusters. Model 3 considers the relative fold change in the focal cluster between the two time points and the population-wide variance in fold change. Performance values are the median of five fold cross-validation. 60

4.1 **Within-host HIV evolutionary events in the ARG.** k_A and k_L represent the number of active and latent lineages present. λ is the global flow rate of a lineage into and out of the latent reservoir. σ is the per genome per day rate of recombination. *This range of λ corresponds to the values of eqn. 4.4 using values of $N_L \in (10^1, 10^3)$. 79

4.2 **Features used to match the simulated evolutionary history to a participant's empirical HIV-1 *env* phylogeny.** Some features measure tree characteristics (Topological) while others measure distance statistics (Distance). A simulation's score, d, is the sum of the normalized differences between the feature measured on reconstructed simulated tree and the empirical tree. 89

4.3 **Participant percentile of the ARG-simulated reconstructed phylogenies.** *Complete* corresponds to the $\sim 30,000$ simulations for each participant. *Best 5%* corresponds to the distribution of only the ARG simulations that produced the lowest 5% of distance scores. 100

B.1 **Full set of candidate predictors considered**. Values were taken at the moment a focal antigenic cluster reached a specified surveillance threshold. The columns *Population, Cluster, Relative* indicate the scale and measure (e.g. mean and/or variance) that a predictor was considered in the model. Depending on the scale of the predictor, the *formula* could refer to all strains in the population, i.e. the strains of infected hosts, or the subset of strains in a specific cluster. *For computational simplicity, these quantities were calculated using strains from a random sample of 10,000 infected individuals. N = number of hosts; t_{a_0} = the time of birth of virus a; λ = antigenic distance between two strains. The antigenic distance is the pairwise degree of cross-immunity between two strains determined by the size of antigenic mutations and parent-offspring relationships; $k(v_i)$ = the number of deleterious mutations on a virus v of infected host i; s_d = the fitness effect of a deleterious mutation; σ_v = the average individual population susceptibility to cluster c; $\sigma_{v,c(h,v)}$ = the probability of infection of a host with historical infection i by a strain of cluster v 125

B.2 **Best-fit model results for surveillance thresholds** $1-10\%$**.** The predictor variables are listed in the order by which they were selected using a forward selection algorithm. The coefficient estimate is the maximum and minimum coefficient (log-odds) from the five-fold cross validation of the final full-term model with the corresponding std. error. 128

B.3 **Evaluating proxy measures for different phases of a novel antigenic cluster's early expansion**. Model 1 shows the performance of the best-fit model using the actual values of relative fitness (relative growth rate) and competition (variance in the population growth rate) for clusters that reached the 5% surveillance thresholds (top two sections) and the 10% surveillance threshold (bottom two sections). Within each section, Model 2 substituted a time proxy for the fitness term and the absolute number of clusters that were growing for the competition term. Model 3 substituted a relative fold change for the fitness term and the population-wide variance in fold change for the competition term. t_1 is when a focal cluster reaches the lower surveillance threshold $(1\%, 3\%, 6\%, 8\%)$; t_2 is when the same cluster reaches the higher surveillance threshold $(5\%, 10\%)$ Performance metric values are the median across the five folds in cross-validation. Balanced accuracy measures the accuracy of the model, accounting for the imbalance in outcomes (i.e. number of transient versus established clusters) in the data set. 130

B.4 **Best-fit logistic regression results for predicting presence-absence of a cluster in X months time into the future.** Terms are listed in the order they were added to the model through forward-selection. 134

B.5 **Best-fit linear regression models for predicting frequency of a cluster in X months time into the future.** Terms are listed in the order they were added to the model through forward-selection. The R^2Adjusted and Root Mean Squared Error (RMSE) were measured on a testing data set of 5 random time samples over a 25-year period. 135

C.1 **The numbers of simulations used to calculate the mean and standard error for Figure 4.3** 143

C.2 **Mean distance score, $\bar{d}$, and standard deviation by recombination rate ρ for Participant 6 and Participant 7.** Participant 6 Top 5%: $\rho = 1e^{-12}, N = 163$; $\rho = 0.001, N = 242$; $\rho = 0.01, N = 884$; $\rho = 0.05, N = 235$; $\rho = 0.1, N = 22$. Participant 7 Top 5%: $\rho = 1e^{-12}, N = 263$; $\rho = 0.001, N = 421$; $\rho = 0.01, N = 855$; $\rho = 0.05, N = 5$. 143

C.3 **Kruskal-Wallis χ^2 ranked sum tests associated with Figure 4.3.** For a given *Metric*, and *Conditional Group*, the Kruskal-Wallis test is testing for significant differences in the means of the *Testing Parameter*. N_L refers to the latent reservoir size and ρ refers to the recombination rate of lineages per day. For example, there is a significant difference in the mean statistic of the Sackin index across possible values of N_L when $\rho = 0.1$. However, there is not a significant difference in the MLT across groups of N_L when $\rho = 0.1$. 144

C.4 **Insignificant pairwise comparisons between *Testing Group 1* and *Testing Group 2* for a given *Conditional Parameter* using Dunn's z-test-statistic.** The pairwise The p-value is adjusted using the Holm's adjusted method for multiple comparison correction. The null hypothesis for each pairwise comparison is rejected if the p-value is less than 0.5α, where $\alpha = 0.05$. All other pairwise comparisons are significant. . . . 145

List of Figures

2.1 **ZIKV emergence scenarios.** A ZIKV infection could spark (A) a self-limiting outbreak or (B) a growing epidemic. Cases are partitioned into symptomatic (grey) and asymptomatic (black). Arrows indicate new ZIKV importations by infected travelers and vertical dashed lines indicate case reporting events. On the 75th day, these divergent scenarios are almost indistinguishable to public health surveillance, as exactly three cases have been detected in both. By the 100th day, the outbreak (A) has died out with 21 total infections while the epidemic (B) continues to grow with already 67 total infections. Each scenario is a single stochastic realization of the model with R_0=1.1, reporting rate of 10%, and introduction rate of 0.1 case/day. 17

2.2 **ZIKV importation and transmission risk estimates across Texas for August 2016.** (A) Color indicates the probability that the next ZIKV import will occur in a given county for each of the 254 Texas counties. Probability is colored on a log scale. The 10 most populous cities in Texas are labeled. Houston's Harris County has 2.7 times greater chance than Austin's Travis County of receiving the next imported case. (B) Estimated county-level transmission risk for ZIKV (See Fig A.7 for seasonal differences). Harris county and Dallas County rank among the top 5 and top 10 for both importation and transmission risk respectively; counties in McAllen and Houston metropolitan area rank among the top 20. Bolded county border indicates counties with recorded local ZIKV transmission. 29

2.3 **Real-time risk-assessment for ZIKV transmission.** All figures are based on transmission and importation risks estimated for Cameron County, Texas. (A) Two thousand simulated outbreaks. (B) Total number of (current) autochthonous cases as a function of the cumulative reported autochthonous cases, under a relatively high (dashed) or low (solid) reporting rate. Ribbons indicate 50% quantiles. (C) The increasing probability of imminent epidemic expansion as reported autochthonous cases accumulate for a low (solid) and high (dashed) reporting rate. Suppose a policy-maker plans to trigger a public health response as soon as a second case is reported (vertical line). Under a 10% reporting rate, this trigger would correspond to a 49% probability of an ensuing epidemic. Under a 20% reporting rate, the probability would be 25%. 30

2.4 **Texas county ZIKV risk assessment.** (A) Probability of an outbreak with at least two reported autochthonous ZIKV cases. (B) The probability of epidemic expansion at the moment the second autochthonous ZIKV case is reported in a county. White counties never reach two reported cases across all 10,000 simulated outbreaks; light gray counties reach two cases, but never experience epidemics. (C) Recommended county-level surveillance triggers (number of reported autochthonous cases) indicating that the probability of epidemic expansion has exceeded 50%. White counties indicate that fewer than 1% of the 10,000 simulated outbreaks reached two reported cases. All three maps assume a 20% reporting rate and a baseline importation scenario for August 2016 (81 cases statewide per 90 days) projected from historical arbovirus data. 32

3.1 **Out-of-sample predictions of antigenic cluster evolutionary success at relative frequency thresholds of** 1% **(a) and** 10% **(b).** Grey shading indicates clusters that surpass the surveillance threshold, but do not establish. Other colors correspond to distinct antigenic clusters that eventually establish. The top time series graphs depict the absolute prevalence of antigenic clusters; the middle graphs give their relative frequencies. The bottom panels indicate the timing and accuracy of out-of-sample predictions based on the optimized model for each surveillance threshold. The top row of symbols indicate clusters predicted to succeed, with true positives indicated by circles and false positives indicated by crosses; the bottom row indicates clusters predicted to circulate only transiently, true negatives indicated by triangles and false negatives indicated by squares. The number of predictions in each category is provided in the legend. 50

3.2 **Relative growth rates predict future success.** (a) Clusters that eventually establish have significantly higher $R_c/\langle R\rangle$ than those that fail to establish. As clusters increase in relative frequency from 1% to 10%, their $R_c/\langle R\rangle$ generally declines but the distinction between future successes and future failures becomes more pronounced. (b) R is a composite value based on the mutational load and $\mathrm{S_{eff}}$. We compared the mutational load (left) and $\mathrm{S_{eff}}$ (right) of a cluster when it crossed the 1% and 10% thresholds by subtracting the former from the latter (orange distributions); we simultaneously calculated the difference in average mutational load and across the entire viral population (grey distributions). The top and bottom rows shows the distributions of change for clusters that establish and transiently circulate, respectively. The decrease in a cluster's fitness advantage is driven by both increasing mutational load and a decreasing $\mathrm{S_{eff}}$. The background mutational load does not change noticeably, while the background $\mathrm{S_{eff}}$ increases slightly. 53

3.3 **Viral competition predicts future success for clusters with borderline growth rates.** (a) Clusters with only a slight $R_c/\langle R\rangle$ advantage are more likely to establish if the background var(R) is low. Clusters with higher $R_c/\langle R\rangle$ successfully regardless of var(R). Contour lines indicate the density of values of $R_c/\langle R\rangle$ and var(R). The lines represent the correlation between the variables for successful and transient clusters. (Success-black: $r = 0.63$, $p < 2.2e^{-16}$; Transient-grey: $r = 0.18$, $p < 1.2e^{-06}$). The dots represent clusters with $R_c/\langle R\rangle$'s between 1.025-1.030, a range within the individual distributions of $R_c/\langle R\rangle$ for success and transient clusters do not statistically differ (Wilcox, p = 0.46). For clusters falling within this ambiguous range of $R_c/\langle R\rangle$, (b) var(R) is significantly higher in transient clusters than in established clusters, and (c) in comparison to transient clusters, successful clusters tend to face fewer co-circulating clusters (Wilcox, $p = 0.005$), with the current dominant cluster at higher frequency (Wilcox, $p = 0.023$). Points represent the number of circulating clusters and the frequency of the dominant cluster; shading represents the kernel density estimation of the distribution of points. Across all graphs, values are calculated when the focal clusters reach a 10% surveillance threshold. 54

3.4 **Empirical antigenic dynamics of influenza A/H3N2, 2006-2018.** (a) Relative frequencies of all antigenic clusters that reach the threshold of at least 10% of sampled viruses. Frequencies are calculated using a 60-day sliding window. Grey shading indicates clusters that surpass the 10% threshold, but do not eventually establish (i.e., reach relative frequency of at least 20% for at least 45 days). Other colors indicate distinct antigenic clusters that eventually establish. (b) A low relative number of epitope mutations when a cluster reaches the 1% relative frequency threshold is an early indicator of future success (Wilcox, $p < 0.003$). We divide the number of epitope mutations of a focal cluster by the average number of mutations of simultaneously circulating clusters. 57

3.5 **Model performance across surveillance thresholds.** (a) Area under the receiver operator curve (AUC) suggests that models can predict successful from unsuccessful clusters by the time they reach 1% of circulating viruses, with discriminatory power declining slightly as clusters rise in frequency. Bars represent the max, median, and minimum AUC values across 5-fold cross validation. (b-c) There is a trade-off between lead time and model performance. The horizontal bars represent the IQR of time between the moment the expanding antigenic cluster reaches the surveillance threshold and when it reaches the success criteria. Vertical bars represent the range and median positive predictive value (b) and sensitivity (c), across five-fold evaluation. Colors correspond to the best fit model for each surveillance threshold. Dashed gray lines indicate lead times of nine months, which represents the current time between the Northern Hemisphere vaccine composition meeting in February and the following start of the influenza season in October. . . 59

4.1 **Ancestral Recombination Graph (ARG) decomposition into a series of binary trees.** (a) The simulated ARG from two sampling events. Red dots represent recombination events, with the break point identified in the white text. Blue dots represent coalescent events between two lineages. Dashed grey lines represent when a lineage is in the latent reservoir. The branch lengths correspond to time. This ARG was simulated using $\rho = 0.08$, $N_L = 760$, and no bottlenecks. 5 samples were taken 3.5 months post transmission and 10 samples were taken 7 months post transmission. (b) Each panel represents the true genealogy of a specific site in the genome from (a). Despite being almost 400 residues apart, residue 61 and residue 418 share the same genealogy. Residue 507 and residue 697 are unique. 85

4.2 **Relationship among within-host evolutionary processes and virus phylogenies.** Each tree in (a-h) is the decomposed average tree from one ARG simulation of 12 years using 132 sampled lineages spread over 10 sampling events. Color represents tips sampled at the same time. The baseline model, a linear-growth coalescent model is in (A). Parameters are consistent across panels (a-h). For simulations in which the process is active: $\mathrm{B}_{\text{frequency}} = 300$ and $\mathrm{B}_{\text{strength}} = 15$; $N_L = 2500$; and $\rho = 0.01$. Panels (i-l) are reconstructed HIV-1 trees from participants in [145]. 91

4.3 **The magnitude of each biological process causes variability in distance and topological statistics.** Each dot represents the mean and standard error of the mean of the statistic. Colors represent different rates of recombination, ρ. Simulations are grouped into three levels of reservoir size of increasing magnitude, with the breakpoints depicted on the x-axis. This figure shows all simulations where $B_{\text{frequency}} < 200$ days, and $B_{\text{strength}} < 75$. Results over all bottleneck combinations of strength and frequency are similar. The number within each category are provided in Table C.1. The grey dashed lines indicate the maximum, median, and minimum value for the 9 empirical participants in [145]. On the y-axis we show the limits of the MLT and the Sackin index to highlight that both the empirical and simulation values are a subset of the possible range. 94

4.4 **Density distributions of topological and distance statistics derived from the complete collection of the ARG-simulated reconstructed phylogenies for Participant 6 and Participant 7.** Each panel shows the kernel density estimate of the statistic calculated from the reconstructed tree or distance matrix over 30,918 simulations for Participant 6 and 30,853 simulations for Participant 7. The blue and yellow lines show the empirical value for each participant. The grey lines show the maximum and minimum values from the 9 participants in [145]. 99

4.5 **Distance algorithm discriminates between good-fit and bad-fit trees.** The top row shows the empirical trees for Participant 6 and Participant 7 with their topological and distance statistics. The middle row shows the top 99th percentile of distance scores (the lowest 1% of scores). The bottom row shows trees at the 75th percentile of distance scores. Parameters for (c) $B_{\text{strength}} = 48.13$, $B_{\text{frequency}} = 207.87$, $\rho = 0.01$, $N_L = 357$; Parameters for (d) $B_{\text{strength}} = 95.28$, $B_{\text{frequency}} = 171.36$, $\rho = 0.01$, $N_L = 467$; Parameters for (e) $B_{\text{strength}} = 90.80$, $B_{\text{frequency}} = 29.53$, $\rho = 0.01$, $N_L = 6673$; Parameters for (f): $B_{\text{strength}} = 84.65$, $B_{\text{frequency}} = 294.54$, $\rho = 0.01$, $N_L = 4068$. . . 101

4.6 **Mean distance scores, $\bar{d}$, for Participant 6 (a) and Participant 7 (b) by recombination rate ρ and latent reservoir size N_L.** The color of each tile represents the mean normalized distance score for each combination of ρ and N_L. d is normalized separately for each participant. While we restrict our simulations to five discrete values of ρ, N_L is sampled from a continuous distribution. For this analysis, we bin N_L into increments of 1000. The number of individual simulations in each tile ranges from 3 when $\rho = 0.1$ to 1000 for $\rho \leq 0.01$. The mean distance score and standard deviation by recombination rate ρ are in Table S: C.2. 102

4.7 **Simulation distance scores for Participant 6 (a) and Participant 7 (b), restricted to the top 5% of simulations.** The number of individual simulations in each tile ranges from 1 to 100. Blank tiles indicate that no simulations with that combination of latent reservoir size and recombination rate are in the top 5% of simulations. The mean distance score and standard deviation by recombination rate ρ are in Table C.2. . 104

4.8 **Size of the latent reservoir accounts for qualitative but not quantitative differences between Participant 6 and Participant 7.** (a) Example reconstructed trees from ARG simulations based on prior distributions where only N_L varied between participants. (b) Violin plots show the kernel probability distribution of the topological and distance statistics from the distribution of reconstructed trees. Dashed lines show the empirical statistic, with the color corresponding to the Participant (blue - Participant 6; yellow - Participant 7). 106

A.1 **ZIKV Risk Assessment Framework.** The method consists of three steps. First, we use data-derived models to estimate county-level ZIKV introduction rates and ZIKV transmission rates. Each estimate is based on a combination of general and county-specific factors. Second, for every county-risk combination, we simulate 10,000 ZIKV outbreaks using a stochastic branching process ZIKV transmission model parameterized by the county-level importation and transmission rate estimates along with several other recently published disease progression estimates. The simulations track the numbers of autochthonous and imported cases (unreported and reported) and, based on the total size and maximum daily prevalence, classifies each outbreak as self-limiting or epidemic. Third, we analyze the simulations to determine (1) robust relationships between the number of reported cases in a county and the current and future ZIKV threat and (2) surveillance triggers (number of reported cases) indicative of imminent epidemic expansion. 116

A.2 **The 95% CI of R_0 Distributions for August.** From left to right, the 2.5%, 50% and 97.5% quantile R_0 values for August. The range of absolute values spans 0.02-6.90. Given the considerable uncertainty in socioeconomic and environmental drivers of ZIKV, we analyzed relative rather than absolute transmission risks. 117

A.3 **Diagram of ZIKV outbreak model.** The model tracks disease progression, transmission, and reporting of both imported and autochthonous ZIKV cases. Individuals progress through compartments via a daily Markovian process, according to the solid arrows in the diagram. The *Exposed* and *Infectious* periods consist of several (boxcar) compartments to simulate realistic outbreak timing. Unreported infected individuals have a daily probability of being reported. Imported cases are assumed to arrive daily according to a Poisson distribution (with mean σ) at the beginning of their infectious period, and otherwise follow the same infectious process as autochthonous cases. Autochthonous transmission occurs at rate $\beta(I_A + I_I)$, where I_A and I_I are the total number of infectious autochthonous and imported cases, respectively (dashed arrows). 118

A.4 **Determining outbreak simulation length.** If outbreak simulations are too short, self-limiting outbreaks may reach the maximum number of infections due to stochasticity. We chose to run our simulations to 2,000 cumulative infections as it conservatively differentiated the large outbreaks of simulations with R_0 just below 1 (0.95) from the epidemics of those with R_0 just above 1 (1.05). We therefore chose to run our simulations until a maximum number of 2,000 infections. 119

A.5 **Time between detection of locally transmitted cases during epidemics.** Across a range of R_0 values with an importation rate 0.1 cases/day, we plot the time between detection events of autochthonous cases for simulations out of the 10,000 trials in which epidemics occurred (black dots). The blue line indicates a two-week threshold as recommended by the CDC for follow-up of local transmission. Even under a high importation rate of 0.1 cases/day, epidemics do not occur when $R_0 = 0.8$, and rarely occur when $R_0 = 0.9$. As R_0 increases, a greater proportion of simulations have fewer days in between detection events as the number of infections rapidly increase. 120

A.6 **Selecting daily prevalence threshold for distinguishing self-limiting outbreaks from epidemics.** Across a range of R_0 values, we plot the maximum daily total autochthonous infectious individuals for 1,000 of our 10,000 trials (black dots). The blue line indicates the threshold (50) selected to differentiate epidemics with $R_0 > 1$ from outbreaks with $R_0 \leq 1$. At a low importation rate (0.01), the majority of simulations with $R_0 \leq 1$ are self-limiting and rarely progress into large sustained outbreaks. As R_0 increases, a greater proportion of simulations exceed the threshold. As the importation rate increases (panels from left to right) the separation between self-limiting outbreaks and epidemics becomes more pronounced. 121

A.7 **Monthly R_0 estimates based on seasonal changes in the temperature-dependent extrinsic incubation period of ZIKV in *Ae. aegypti* and the mosquito mortality rate of *Ae. aegypti*** 122

B.1 **The proportion of total infections caused by established clusters is more sensitive to a frequency criteria than the duration of time**. Established antigenic clusters account for the majority of the disease activity at any point in time. In our analysis, established clusters are those that circulate above 20% relative frequency for at least 45 days. Using this criteria, infections caused by future successful clusters account for a median of 81% of the disease activity at any point in time. . 124

B.2 **The fate of novel antigenic clusters** (A) Each point represents the number of antigenic clusters in our simulations that reach each increasing surveillance threshold (i.e., relative frequency in the population). As the surveillance threshold increases from 1 to 10%, the number of candidate clusters decreases from 7969 clusters at the 1% threshold to 1816 clusters at the 10% surveillance threshold. (B) Given a cluster has reached a surveillance threshold, the proportion of antigenic clusters that will establish (i.e. reaches > 20% for 45 days) increases with higher surveillance thresholds. 126

B.3 **Predictive models that rely on data from a single sample of time perform similarly to those that include data from multiple time points**. Each point represents the combination of candidate predictors that best predict the evolutionary fate of antigenic clusters at varying surveillance thresholds, whether the combination includes data from a single time point (yellow) or multiple time points (purple). Model performance is measured in terms of sensitivity and positive predictive value. Dots represent the median, and error bars span the range of performance values across the five folds of cross-validation of the best-fit model. 129

B.4 **The rate of change (a) and relative fold change (b) as proxy measures for the relative growth rate** $R/\langle R\rangle$. Contour lines indicate the density of cluster values for clusters that establish (black) and those that transiently circulate (grey). Values along the x-axis indicate the measured relative growth rate of a cluster the moment it reaches the 10% surveillance threshold. Values along the y-axis indicate the proxy measure (rate of change in (a) and relative fold change in (b) for the cluster, approximated for the time between the 6% and 10% thresholds. The rate of change, measured in the number of days between the two thresholds, is a better proxy measure than relative fold change. 131

B.5 **The distribution of the 2846 cluster frequencies from 620 random time samples across the 1500 years of influenza evolution.** Clusters that were below 1% relative frequency in the population or those that had already reached our establish criteria were excluded. 133

B.6 **Comparing the sensitivity and positive predictive value trade-off of two surveillance strategies: 1) surveillance threshold (circles, triangles) and 2) random sampling through time (squares).** Each line highlights the trade off between sensitivity and positive predictive value at different probability thresholds for what constitutes a positive prediction, i.e. a future successful cluster. All models converge in areas with low sensitivity and high positive predictive value, where the model has to predict with probability 0.90 that the cluster will establish in order to classify it as a positive prediction. However, in areas of greater sensitivity, the time sample model consistently under performs surveillance threshold models based on a 5% or higher surveillance threshold. The black stars represent the probability threshold that maximizes the F1 value. 136

B.7 **Model predictions of cluster frequency up to a year in advance in three-month increments.** Black dots represent clusters that will eventually establish, and grey dots are clusters that will transiently circulate. The black line represents perfect agreement between the actual and predicted log frequencies. The number of clusters present at the time of sampling, but expected to persist in the future, decreases with increasing month-ahead predictions. By 12 months, fewer than half of the starting clusters will still be circulating. As you predict further into the future, the model underestimates future-high frequency circulating clusters, which are usually clusters that will establish. 137

C.1 **The impact of bottleneck strength and frequency on distance and tree statistics.** Bottleneck frequency is measured in days *Rare* [200-365], *Periodic* [100-200) *Frequent* [14-100). Bottleneck strength is the size of N_e during a bottleneck. As N_e grows linearly with time, the bottleneck accounts for progressively greater proportions of the effective population size from the time of infection. Bottleneck Strength Categories: *Strong* [1-30), *Medium* [30-60), *Weak* [60-100]. The grey lines represent the empirical minimum, median, and maximum value of each statistic. 138

C.2 **Distributions of distance score, d, for each participant (P6: N= 30,918, P7: N = 30,853).** Within each participant, we normalize the distance scores across simulations so that a score $d = 0$ represents the closest-matching simulation and $d = 1$ represents the worst-matching simulation. The median of Participant's 7 score distribution is lower than that of Participant 6 (Wilcoxon Rank Sum, p ¡ 2.2e-16) because simulations with a $\rho = 0.01$ had the lowest scores compared to Participant 6, where $\rho = 0.05$ and $\rho = 0.1$ had lower scores. Because of computational cost and the reduced variability in simulation outcome at higher recombination rates, we sampled 10-100 times less at these recombination rates than at $\rho = 0.01$. Distance summary statistics: Participant 6 (median: 0.34, IQR: 0.24-0.44); Participant 7 (median: 0.17, IQR: 0.12-0.25). . . . 139

C.3 **Sampling diagnostics for Participant 6**. (Left) Pairwise relationships between parameters in the top 5% matching simulations for Participant 6. (Right) Pairwise relationships between statistical probes in the top 5% matching simulations for Participant 6. 140

C.4 **Sampling diagnostics for Participant 7**. Pairwise relationships between parameters in the top 5% matching simulations for Participant 7. (Right) Pairwise relationships between statistical probes in the top 5% matching simulations for Participant 7 140

C.5 **Marginal distributions of N_L, bottleneck frequency, and bottleneck strength.** Colors distinguish the density of parameters between simulations with the lowest 5% of distance scores (yellow) and highest 95% of distance scores (blue). Two-sided Kolmogorov-Smirnov tests indicated the distributions for the latent reservoir size and for bottleneck strength are significantly different between parameters in the lowest 5% of scores and those in the top 95% for Participant 6 (Latent Reservoir: $D = 0.735$, $p < 2.2e^{-16}$; Strength: $D = 0.110$, $p = 5.651^{-05}$; Frequency: $D = 0.063$, $p = 0.060$). Two-sided Kolmogorov-Smirnov tests indicated the distributions for all parameters are significantly different for Participant 7 (Latent Reservoir: $D = 0.270$, $p < 2.2e^{-16}$; Strength: $D = 0.308$, $p < 2.2e^{-16}$; Frequency: $D = 0.271$, $p < 2.2e{-}16$). The lowest 5% distributions are significantly different between Participant 6 and Participant 7 for each parameter (Kolmogorov-Smirno two-sided tests, $p < 2.2e^{-16}$). Summary statistics for the lowest 5% (yellow) in the form of Q1-Median-Q3 Patient 6: Latent Pool Size (0.12-0.30-0.63); Strength (32.44-55.43-78.99); Frequency (114.34-202.03-287.80). Patient 7: Latent Reservoir (0.58-1.14-1.85); Strength: (52.46-70.05-86.09); Frequency (187.18-253.63-312.11). 141

C.6 **Conditional distributions of N_L, bottleneck frequency, and bottleneck strength by recombination rate for the lowest 5% of distance scores.** The line through each ridge represents the 50th quantile. Within each participant, the median of the Latent Reservoir increases with the recombination rate. 142

Chapter 1

Introduction

1.0.1 Thematic Overview

Effective infectious disease management relies on understanding factors that promote pathogen transmission. These factors vary depending on the type of disease (e.g., endemic or novel, acute or chronic) and the public health goal (e.g., prevention, control, treatment). For novel diseases, or disease expanding into a new location (e.g., the 2014 Ebola and 2016 Zika virus outbreaks) key questions for disease management include: Where is the disease likely to spread? How many cases are expected? Will the outbreak wane without intervention? If not, what is the time scale for implementing control measures? Unfortunately, the pathogen-specific epidemiological data necessary to answer these questions often do not exist.

On the other hand, for an endemic disease that occurs annually, like seasonal influenza, the pathogen-specific epidemiological data does exist. Even so, the complex interactions between an endemic disease's epidemiology, human immunology, and environmental factors create uncertainty for future outbreaks [26, 91, 165]. For these diseases, a different set of questions should be asked: How does the magnitude and timing of this year's outbreak differ from previous years? Will it be influenced by climatic factors, such as increased rainfall? Will individuals infected last year be susceptible to the virus this year?

Analogous to controlling epidemic spread on the population level, clinicians make decisions to treat and cure patients with chronic infections. Chronic infections flare and wane according to a complex, within-host interaction between the immune system, treatment strategies, pathogen evolution, and co-infections. For a chronic disease such as HIV-AIDS, a clinician asks: How much pathogen is there in the patient and how will the level affect the projected pathogenesis? Will the individual respond to a specific treatment or is the pathogen resistant to it?

Infectious disease management makes evidence-based decisions by combining information from multiple data sources to gain a better picture of the current state of the outbreak or infection. These data sources include forms of disease surveillance that are based upon direct patient contact with public health officials, such as syndromic surveillance and biological tests, and data sources that capture individual behavior, such as school absenteeism or Google searches [5]. For example, disease incidence in a population might be measured either by the number of people seeking routine clinical care that match a syndromic case definition or by the proportion of biological tests that return positive confirmation [31, 32, 67]. Laboratory tests that measure the presence and quantity of antibodies within an infected individual can provide a picture of the within-host dynamics of an individual infection. Moving beyond these established methods, technological advances in computing power and next-generation sequencing (NGS) capabilities are generating a wealth of new data sources that can be harnessed for infectious disease management at both the within-host and between-host levels. Internet-based search activity can complement traditional epidemiological data to assess current disease incidence [4, 58, 88, 121, 141]. NGS pathogen genome data from infected patients holds great potential to inform diagnostics, patient care, and public health epidemiology as it can provide detailed information on the evolutionary history of a pathogen population within and between hosts [13, 57, 60], estimates of the size of the infected population [128], the rate of transmission [142] and many other informative variables.

Despite everything we can learn from our current data collection practices, pathogen evolution and transmission are partially visible processes. The captured data provides only a glimpse into the state of an outbreak or an individual infection. The conclusions derived from the data will be affected by the biases inherent to how the data is collected. Case reporting can be affected by the asymptomatic rate, the pathogen virulence, the quality of the diagnostics, and access to health care [56]. Because molecular sequence data primarily comes from individuals who seek routine clinical care, it is affected by the same biases. At a global level, molecular sequence sampling might be over represented in geographic areas with widespread sequencing capabilities while other areas remain underrepresented. In addition, the timing of the samples [55, 117, 152] and choice of analytical methods can influence conclusions derived from molecular sequence data. Therefore, one key infectious disease management challenge is translating sparse incidence and genetic data to actionable knowledge about the current and future state of disease dynamics.

Mathematical modeling is a useful tool for guiding the interpretation of incomplete and biased data [94]. By incorporating the ecological and evolutionary principles that govern disease dynamics and determine key epidemiological-immunological interactions, mathematical models can provide insight into the progression of an outbreak, assess the effectiveness of control efforts, and define pathogen epidemiological properties [77, 91, 129]. Mathematical models can further integrate the vagaries of data collection or observational processes and thus provide an intuitive framework for testing hypotheses while explicitly incorporating error. In particular, stochastic simulation methods—models in which defined events are not deterministically implemented but governed by probabilities—are well-suited to address these challenges because they provide a means of generating a range of plausible scenarios under defined starting conditions [70]. By coupling the outcome of stochastic simulation models with further statistical analysis, it is possible to quantify the probability of various outcomes, whether that is the likelihood of an outbreak or the elimination of an individual infection. These probabilities can then be used by policy makers to best decide where to commit limited resources, personnel, and time.

1.0.2 Chapter Summaries

In this dissertation, I combine realistic simulation models of disease transmission and evolution that produce either or both epidemiological and molecular data with statistical methods to address pressing infectious disease management questions. The questions fall into two categories: *predictive use*, those which explore outcomes under specific, pre-defined assumptions, and *statistical inference*, those which evaluate alternative hypotheses and estimate demographic and genetic parameters. The first two chapters use stochastic modeling for predictive infectious disease applications at the between-host level. In Chapter 2, I develop a new simulation model of Zika transmission that integrates the probability of an importation and the probability of local transmission to estimate a Texas county's risk of a Zika epidemic. In Chapter 3, using a rigorous experimental approach, I use a validated simulation of seasonal influenza AH3N2 to advance influenza surveillance forecasting strategies and quantify inherent limits of molecular forecasting. In my final chapter, I taken an inferential approach to examine how molecular data can inform knowledge about the within-host evolutionary processes of HIV-1 by developing a novel simulation model of HIV-1 evolution that incorporates key aspects of HIV-1 biology, specifically recombination, latency, and population demography.

In Chapter 2, I develop a county-level model of Zika transmission in Texas to help translate real-time reported case data into an estimate of future epidemic risk. These estimates are necessary because public health responses are difficult and expensive. During the widespread 2016 Zika epidemic, U.S public health officials were instructed to implement enhanced mosquito control efforts upon documenting two locally transmitted Zika cases. However, because of variable ecological and socioeconomic conditions, some areas might need to act upon recording the first Zika case as even a single case portends rapid spread. Other areas, in contrast, could delay allocating resources until a higher number of cases were reported. To support the decision of when to implement control efforts, I create a framework that combines local ecological suitability for Zika transmission with the probability of importing Zika cases from abroad to assess how the risk of an epidemic in a county changes as a function of the number of reported cases. Importantly, given that 80% of cases are asymptomatic, I account for transmission not captured in reported case counts. By incorporating each county's ecological and importation risk, I was able to assess the relative risk of a Zika outbreak across Texas' 254 counties. I find that during peak mosquito season, starting a response at two cases would not be sufficient to stop an outbreak in some areas (e.g., Rio Grande Valley and the Houston Metropolitan Area), while other areas of the state could wait until dozens of cases appeared. This framework provides a quantitative means for public health officials to develop county-specific guidelines for when to enact intervention efforts, and was presented to the Texas Department of Health and Human Services. This chapter demonstrates the importance of estimating location-specific risk and the challenge of obtaining reliable estimates of key parameters for novel diseases.

In Chapter 3, I identify important early predictors of Influenza A/H3N2 antigenic cluster transitions because such predictors can be incorporated into the decision making process for which strain to include in the seasonal vaccine.

Seasonal influenza poses a significant public health burden. Last year, over 185 pediatric deaths and roughly 800,000 hospitalizations were attributed to a severe strain of influenza A/H3N2. Seasonal influenza vaccines serve as a critical line of defense, but their efficacy depends on immunological matching with circulating viruses. During the annual selection of vaccine strains, public health agencies and vaccine manufacturers collaborate to forecast which influenza strains will dominate 9-12 months into the future. To support these decisions, scientists have begun to develop models that predict influenza evolutionary dynamics from molecular and epidemiological data collected by labs worldwide. I applied an empirically validated, high resolution simulation of influenza A/H3N2 evolution to address the following questions: How reliably and with how much lead time can we identify strains that will rise to dominance in future influenza seasons? What data should we be collecting to accelerate and improve the accuracy of such forecasts?

I found that, in a scenario in which we can precisely measure the mutational load and population-wide susceptibility of every newly appearing strain, we can reliably identify strains that will ultimately dominate with at least six months advanced warning. This is true even when a given strain is still at relative frequencies below 10% in the population. However, on a realistic vaccination production time scale of 9-12 months, our ability to identify strains is reduced. The results underscore the tradeoff between prediction accuracy and the lead-time for vaccine development. This chapter describes how numerous complimentary metrics can be derived from one data source and the broader potential for model-based evaluation and optimization of public health surveillance efforts.

In Chapter 4, I advance a within-host evolutionary model of Human Immunodeficiency Virus Type 1 (HIV-1) to include realistic biological features of HIV-1 evolution. Previously, this model had focused on modeling epidemiological links between two patients and had assumed no recombination, latent reservoir, or selection pressure on viral population size. To model HIV-1 evolution within a single host, I added all three processes in. The model serves to connect the longitudinal viral RNA sequence samples, (i.e. the data one collects over the course of an individual's infection) with the underlying evolutionary history of the viral sample. Since its origins in the human population in the mid-twentieth century, the HIV-AIDS epidemic has been one of the most devastating infectious diseases to affect modern human history. Despite antiretroviral therapy's (ART) ability to knock HIV-1 viremia to undetectable levels and extend the quality of life of infected individuals, drug therapy does not eradicate the virus. In part, a cure has been elusive because of the complexity and our limited understanding of the within-host HIV-1 biology. The growth of HIV-1 sequence databases has motivated the development of phylogenetic methods to gain a better understanding of the evolutionary dynamics of HIV-1. To answer questions about how the interactions between widespread evolutionary processes should manifest themselves on phylogenetic trees reconstructed from longitudinal samples, I integrated these key components of HIV-1 biology into a mathematical modeling framework. The phylogenetic patterns arising from this model can be used to infer estimates of important immune-evading mechanisms such as rates of recombination and the size of the latent reservoir pool, as well as to compare qualitative differences between individual patients. I find that high rates of recombination and a large latent reservoir size produce unrealistic HIV-1 trees. Second, the size of the latent reservoir might be variable between individuals. This chapter demonstrates the importance and challenge of developing realistic evolutionary models to improve inference of within-host evolutionary processes.

1.0.3 Data Logistics and Challenges of Simulation Modeling

The chapters in this dissertation show the utility of using stochastic simulation models for addressing questions of how to interpret or strategically collect data that public health officials and clinicians use for evidence-based decision making. However, there is another type of data present in this dissertation that warrants mention, namely the data that is used to parameterize and fit models. The quality and quantity of this data is an important factor to consider when developing simulation models and translating results from *in silico* to real-world settings. I will briefly describe how data availability impacted the model selection and presented challenges in all three research chapters.

The extent to which a model captures the real-world complexities that dictate disease dynamics is in part dictated by the ability to parameterize a complex model. If every component going into the model contains uncertainty, the results become less useful for public health officials, who often want a single quantitative statement to answer their question. Model complexity therefore becomes a balance between the quantity and quality of the data and the question of interest. This point is demonstrated by the three different simulation modeling techniques I used in this dissertation: an analytically tractable compartmental model of Zika transmission, a large-scale agent-based model of influenza A H3N2 transmission, and a coalescent-based simulation of HIV-1 evolution. When a disease is largely unknown, such as Zika, the wide range of uncertainty that comes from uninformative prior parameter estimates severely constrains the complexity of the model. In this case, even parameterizing a compartmental model can prove challenging. In Chapter 2, one way we sought to reduce the complexity of Zika transmission was by combining the two-step human-mosquito-human transmission interaction into one human-human step. Because our question concerned how quickly cases accumulated and because early parameter estimates were available for the Zika serial interval, we could integrate these two biological steps into one step in our model. In contrast, for disease that has been studied in depth, like seasonal influenza, there is sufficient information to parameterize an agent-based model that captures higher levels of realism. For Chapter 3, a fine-grained model was necessary to give us a detailed view into how small scale events affect population-level properties. Yet, there is still a limit on the complexity that this model could include because of both computational cost and data limitations. For example, it does not include spatial structure or more detailed immunological phenomenons that might influence antigenic evolution, such as original antigenic sin and short-term heterotypic immunity. In the final chapter, the premise of the coalescent model – simulating in reverse time — is in part to reduce complexity. Instead of forward simulating a growing population that would also require modeling the dynamic interplay between the virus population and the immune system, we only focus on a subset of the virus population and define a limited set of demographic events.

The realism in the model in turn impacts how easy it is to bridge the gap between using simulations as scenario-based tools to directly linking the simulation results to reality. In Chapter 2, although our results were able to provide reasonable relative risk assessments in the short term, we saw in the long-term that our expectations did not match the observed data. Our absolute estimates of county-level epidemiological risk, R_0, were too high and Zika cases did not materialize in Texas despite high numbers of importations in the most risky areas. This discrepancy was most likely due to a combination of a lack of Zika-specific data, a weak understanding of how socioeconomic conditions impact mosquito-human interaction in developed locations, and the county-level resolution of our model as mosquito transmission may be better modeled at the local and household levels. In Chapter 3, we saw the challenge of linking the simulation to reality when we tried to reproduce our results on an empirical data set. As previously mentioned, one of the strengths of the agent-based model framework is that it gives a high-resolution view of the system, that is often not reproducible in the real-world. However, the empirical data did not contain enough information on its own to test the predictors we had identified as most useful. Furthermore, the temporal resolution of the data was not sufficient to reproduce our proxy model results. In Chapter 4, we saw the challenge of using statistical inference to narrow in on plausible parameter combinations of individual patients. For complex models, such as ARGs, the likelihood function is computationally costly to evaluate, and so we need to look to likelihood-free statistical inference methods. While we were able to use our sampling estimation framework to quantitatively ascertain which end of the spectrum HIV-1 parameter values most likely reside, defining a narrow range of plausible parameters for different participants proved difficult. To adequately infer particular parameter values will require an immense amount of computational power to get sufficient statistical power.

1.0.4 Concluding Remarks

Stochastic simulation and statistical modeling provide a way to examine the complex interplay of the ecological, evolutionary, and behavioral forces that govern disease dynamics and surveillance. Using three in-depth analyses during my dissertation, I demonstrate how these methods can be applied to complex infectious disease management questions in varied biological settings. In this dissertation, I focused on two forms of real-time data, case counts and molecular sequence data. My results show how these two data sources can be powerful alone or in combination for providing decision support. Looking forward, infectious disease management is increasingly interested in fusing disparate forms of data sources, such as Internet data, weather, and mobility data, with case and molecular data to answer complex questions for different public health objectives. As each of these new data sources becomes integrated into infectious disease modeling, it will be paramount to assess the utility and limitations in how the data is collected and incorporated. This dissertation provides examples of how to approach those questions and outlines challenges to consider in methodology and translating insights from theoretical to real-word settings. Altogether, the results in this dissertation suggest the interface between the modeling community and the clinical and public health community will be critical to strategically thinking about data collection and interpretation for the purpose of addressing infectious disease management questions.

Chapter 2

Assessing real-time Zika risk in the United States

2.1 Abstract [1]

Confirmed local transmission of Zika Virus (ZIKV) in Texas and Florida have heightened the need for early and accurate indicators of self-sustaining transmission in high risk areas across the southern United States. Given ZIKV's low reporting rates and the geographic variability in suitable conditions, a cluster of reported cases may reflect diverse scenarios, ranging from independent introductions to a self-sustaining local epidemic. We present a quantitative framework for real-time ZIKV risk assessment that captures uncertainty in case reporting, importations, and vector-human transmission dynamics. We assessed county-level risk throughout Texas, as of summer 2016, and found that importation risk was concentrated in large metropolitan regions, while sustained ZIKV transmission risk is concentrated in the southeastern counties including the Houston metropolitan region and the Texas-Mexico border (where the sole autochthonous cases have occurred in 2016).

---

[1] This chapter was published as Castro LA, Fox SJ, Chen X, Liu K, Bellan SE, Dimitrov NB, et al. Assessing real-time Zika risk in the United States. BMC Infect Dis. BMC Infectious Diseases; 2017;17: 284. doi:10.1186/s12879-017-2394-9 [27]

We found that counties most likely to detect cases are not necessarily the most likely to experience epidemics, and used our framework to identify triggers to signal the start of an epidemic based on a policymakers propensity for risk. This framework can inform the strategic timing and spatial allocation of public health resources to combat ZIKV throughout the US, and highlights the need to develop methods to obtain reliable estimates of key epidemiological parameters.

2.2 Introduction

In February 2016, the World Health Organization (WHO) declared Zika virus (ZIKV) a Public Health Emergency of International Concern [64]. Though the Public Health Emergency has been lifted, ZIKV still poses a great threat for reemergence in susceptible regions in seasons to come [170]. In the US, the 268 reported mosquito-borne autochthonous (local) ZIKV cases occurred in Southern Florida and Texas, with the potential range of a primary ZIKV vector, *Aedes aegypti*, including over 30 states [49, 51, 162]. Of the 2,487 identified imported ZIKV cases in the US through the end of August, 137 had occurred in Texas. Given historic small, autochthonous outbreaks (ranging from 4 - 25 confirmed cases) of another arbovirus vectored by *Ae. Aegypti*—dengue (DENV) [159, 160, 162], Texas was known to be at risk for autochthonous arbovirus transmission, and the recent outbreaks have highlighted the need for increased surveillance and optimized resource allocation in the states and the rest of the vulnerable regions of the Southern United

States.

As additional ZIKV waves are possible in summer 2017, public health professionals will continue to face considerable uncertainty in gauging the severity, geographic range of local outbreaks, and appropriate timing of interventions, given the large fraction of undetected ZIKV cases (asymptomatic) and economic tradeoffs of disease prevention and response [3, 44, 90, 93]. Depending on the ZIKV symptomatic fraction, reliability and rapidity of diagnostics, importation rate, and transmission rate, the detection of five autochthonous cases in a Texas county, for example, may indicate a small chain of cases from a single importation, a self-limiting outbreak, or a large, hidden epidemic underway (Fig 2.1). These diverging possibilities have precedents. In French Polynesia, a handful of ZIKV cases were reported by October 2013; two months later an estimated 14,000-29,000 individuals had been infected [90, 93]. By contrast, Anguilla had 17 confirmed cases from late 2015 into 2016 without a subsequent epidemic, despite large ZIKV epidemics in surrounding countries [123]. To address the uncertainty, the CDC issued guidelines for state and local agencies; they recommend initiation of public health responses following local reporting of two non-familial autochthonous ZIKV cases [52].

Previous risk assessments of ZIKV have provided static *a priori* assessments based on historical incidence and vector suitability, but they do not provide dynamic risk assessments as cases accumulate in a region. Here, we present a framework to support real-time risk assessment, and demonstrate its application in Texas. Our framework accounts for the uncertainty regarding

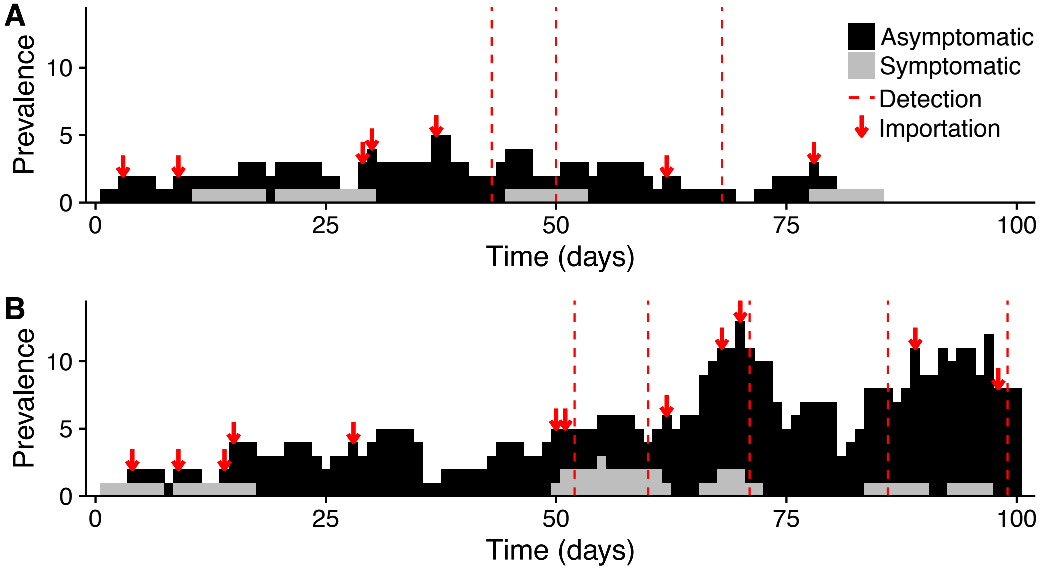

Figure 2.1: **ZIKV emergence scenarios.** A ZIKV infection could spark (A) a self-limiting outbreak or (B) a growing epidemic. Cases are partitioned into symptomatic (grey) and asymptomatic (black). Arrows indicate new ZIKV importations by infected travelers and vertical dashed lines indicate case reporting events. On the 75th day, these divergent scenarios are almost indistinguishable to public health surveillance, as exactly three cases have been detected in both. By the 100th day, the outbreak (A) has died out with 21 total infections while the epidemic (B) continues to grow with already 67 total infections. Each scenario is a single stochastic realization of the model with R_0=1.1, reporting rate of 10%, and introduction rate of 0.1 case/day.

ZIKV epidemiology, including importation rates, reporting rates, local vector populations, and socioeconomic conditions, and can be readily updated as our understanding of ZIKV evolves. To estimate current and future epidemic risk from real-time ZIKV case reports, the model incorporates a previously published method for estimating local ZIKV transmission risk and a new model for estimating local importation risk. Across Texas' 254 counties, we find that the estimated risk of a locally sustained ZIKV outbreak rises precipitously as autochthonous cases accumulate, and that counties at the southern tip of the Texas-Mexico border and in the Houston Metropolitan Area are at the highest risk for ZIKV transmission. This statewide variation in risk stems primarily from mosquito suitability and socio-environmental constraints on ZIKV transmission rather than heterogeneity in importation rates.

2.3 Methods

Our risk-assessment framework is divided into three sections: (1) county-level epidemiological estimates of ZIKV importation and relative transmission rates, (2) county-specific ZIKV outbreak simulations, and (3) ZIKV risk analysis (Fig A.1). To demonstrate this approach, we estimate county-level ZIKV risks throughout the state of Texas for August 2016, given that, by May 2016, Texas experienced dozens of ZIKV importations without subsequent vector-borne transmission.

2.3.1 Estimating Importation Rates

Our analysis assumes that any ZIKV outbreaks in Texas originate with infected travelers returning from active ZIKV regions. To estimate the ZIKV importation rate for specific counties, we (1) estimated the Texas statewide importation rate (expected number of imported cases per day) for August 2016, (2) estimated the probability (import risk) that the next Texas import will arrive in each county, and (3) took the product of the state importation rate and each county importation probability.

1. During the first quarter of 2016, 27 ZIKV travel-associated cases were reported in Texas [162], yielding a baseline first quarter estimate of 0.3 imported cases/day throughout Texas. In 2014 and 2015, arbovirus introductions into Texas increased threefold over this same time period, perhaps driven by seasonal increases in arbovirus activity in endemic regions and the approximately 40% increase from quarter 1 to quarter 3 in international travelers to the US [118]. Taking this as a baseline (lower bound) scenario, we projected a corresponding increase in ZIKV importations to 0.9 cases/day (statewide) for the third quarter.

2. To build a predictive model for import risk, we fit a probabilistic model (maximum entropy) [73] of importation risk to 183 DENV, 38 CHIKV, and 31 ZIKV Texas county-level reported importations from 2002 to 2016 and 10 informative socioeconomic, environmental, and travel variables (Table 2.2). Given the geographic and biological overlap between ZIKV, DENV and Chikungunya (CHIKV), we used historical DENV and CHIKV importation data to supplement ZIKV importations in the importation risk model, while recognizing that future ZIKV importations may be fueled by large epidemic waves in neighboring regions and summer travel, and thus far exceed recent DENV and CHIKV importations [109]. Currently, DENV, CHIKV, and ZIKV importation patterns differ most noticeably along the Texas-Mexico border. Endemic DENV

transmission and sporadic CHIKV outbreaks in Mexico historically have spilled over into neighboring Texas counties. In contrast, ZIKV is not yet as widespread in Mexico as it is in Central and South America, with less than 10 reported ZIKV importations along the border to date (October 2016). We included DENV and CHIKV importation data in the model fitting so as to consider potential future importation pressure from Mexico, as ZIKV continues its increasing trend since March 2016 [119]. To find informative predictors for ZIKV importation risk, we analyzed 72 socio-economic, environmental, and travel variables, and removed near duplicate variables and those that contributed least to model performance, based on out-of-sample cross validation of training and testing sets of data [100, 167], reducing the original set of 72 variables to 10. We validated our importation model by comparing the predicted distribution of cases across the state given a total number of imported cases (September 2016) as a linear predictor of the empirical distribution of cases across counties.

2.3.2 County Transmission Rates (R_0)

The risk of ZIKV emergence following an imported case will depend on the likelihood of mosquito-borne transmission. For emerging diseases like ZIKV, the public health and research communities initially face considerable uncertainty in the drivers and rates of transmission, given the lack of field and experimental studies and epidemiological data, and often derive insights through analogy to similar diseases. For our case study, we estimated county-level ZIKV transmission potential by *Ae. aegypti* using a recently published model [2], that derives some of its key parameters from DENV data (Table 2.1). The utility of our framework depends on the validity of such estimates and will increase as our knowledge of ZIKV improves. However, we expect our results to be robust to most sources of uncertainty regarding ZIKV and DENV epidemiology, as they may influence the absolute but not relative county-level risks.

We estimated the ZIKV reproduction number (R_0), the average number of secondary infections caused by a single infectious individual in a fully susceptible population, for each Texas county following the method described in Perkins et al. [2]. The method calculates R_0 using a temperature-dependent formulation of the Ross-Macdonald model, where mosquito mortality rate (μ) and extrinsic incubation period of ZIKV (n) are temperature dependent functions; the human-mosquito transmission probability ($b = 0.4$), number of days of human infectiousness ($\frac{c}{r} = 3.5$), and the mosquito biting rate ($a = 0.67$) are held constant at previously calculated values [2, 17, 21, 33, 106, 116]; and the economic-modulated mosquito-human contact scaling factor (m) is a function of county mosquito abundance and GDP data fit to historic ZIKV seroprevalence data [2]. To account for uncertainty in the temperature-dependent functions (the extrinsic incubation period (EIP) and mosquito mortality rate) and in the relationship between economic index and the mosquito-to-human contact rate, Perkins et al. generated functional distributions via 1000 Monte

Parameter	Description	Values Investigated (or median 95%)	Source
Exposed compartments (e)	Number of exposed compartments	6	Fit (See Section A.1.1)
Incubation Rate (ν)	Daily probability of progressing from one exposed compartment to the next	0.584	[90, 96]
Infectious compartments (n)	Number of infectious compartments	3	Fit (See Section A.1.1)
Recovery Rate (δ)	Daily probability of progressing from one infectious compartment to the next	0.3041	[90, 96]
Reproduction Number (R_0)	The expected total number of secondary infections from one infectious individual in a fully susceptible population	$0 - 3.1$	County R_0 estimates
Daily Reporting Rate (η)	The daily probability of an infectious individual being reported	Daily: $0.011 - 0.0224$ Overall: $10 - 20\%$	[44]
Daily Importation Rate (σ)	The expected number of infectious ZIKV importations per day	$0.0 - 1.21$	County importation rate estimates
Generation Time	The average length of time between consecutive exposures $GT = \frac{e}{\nu} + (\frac{1}{2})\frac{n}{\delta}$	15 (9.5-23.5) days	[96]

Table 2.1: **Stochastic ZIKV outbreak model parameters.** We hold the disease progression parameters constant across all scenarios, estimate R_0 and importation rate for each individual county, and vary the reporting rate to investigate its impact on the uncertainty of ZIKV risk assessments.

Carlo samples from the underlying parameter distributions. We assume DENV estimates for these temperature-dependent functions, since we lack such data for ZIKV and these *Flaviviruses* are likely to exhibit similar relationships between temperature and EIP in *Ae. aegypti* [17]. We used the resulting distributions to estimate R_0 for each county, based on county estimates for the average August temperature, mosquito abundance from Kraemer et al. [84], and GDP [17]. Our R_0 estimates were similar to those reported by Perkins et al. [2] with 95% confidence intervals spanning from 0 to 3.1 (Fig A.2). Given this uncertainty, and that our primary aim is to demonstrate the risk assessment framework rather than provide accurate estimates of R_0 for Texas, we use these estimates to estimate relative county-level transmission risks (by scaling the county R_0 estimates from 0 to 1). In each simulation, we assume that a county's R_0 is the product of its relative risk and a chosen maximum R_0. For our case study, we assume a maximum county-level R_0 of 1.5 This is consistent with historical arbovirus activity in Texas (which has never sustained a large arbovirus epidemic) and demonstrates the particular utility of the approach in distinguishing outbreaks from epidemics around the epidemic threshold of $R_0 = 1$.

2.3.3 ZIKV Outbreak Simulation Model

Assuming mosquito-borne transmission as the main driver of epidemic dynamics, to transmit ZIKV, a mosquito must bite an infected human, the mosquito must get infected with the virus, and then the infected mosquito must bite a susceptible human. Rather than explicitly model the full transmission cycle, we aggregated the two-part cycle of ZIKV transmission (mosquito-to-human and human-to-mosquito) into a single exposure period where the individual has been infected by ZIKV, but not yet infectious, and do not explicitly model mosquitoes. For the purposes of this study, we need only ensure that the model produces a realistic human-to-human generation time of ZIKV transmission, and the simpler model is more flexible to disease transmission pathways. We fit the generation time of the ZIKV model to early ZIKV Epidemiological estimates, with further fitting details described in section A.1.1.

The resultant model thus follows a Susceptible-Exposed-Infectious-Recovered (SEIR) transmission process stemming from a single ZIKV infection using a Markov branching process model (Fig A.3). The temporal evolution of the compartments is governed by daily probabilities of infected individuals transitioning between disease states. New cases arise from importations or autochthonous transmission. We treat days as discrete time steps, and the next disease state progression depends solely on the current state and the transition probabilities. We assume that infectious cases cause a Poisson distributed number of secondary cases per day (via human to mosquito to human transmission), but this assumption can be relaxed as more information regarding the distribution of secondary cases becomes available. We also assume infectious individuals are introduced daily according to a Poisson distributed number of cases around the importation rate. Furthermore, Infectious cases are categorized into reported and unreported cases according to a reporting rate. We assume that reporting rates approximately correspond to the percentage ($\sim 20\%$) of symptomatic ZIKV infections [44] and occur at the same rate for imported and locally acquired cases. Additionally, we make the simplifying assumption that reported cases transmit ZIKV at the same rate as unreported cases. We track imported and autochthonous cases separately, and conduct risk analyses based on reported autochthonous cases only, under the assumption that public health officials will have immediate and reliable travel histories for all reported cases [29].

2.3.4 Simulations

For each county risk scenario, defined by an importation rate, transmission rate, and reporting rate, we ran 10,000 stochastic simulations. Each simulation began with one imported infectious case and terminated either when there were no individuals in either the Exposed or Infectious classes or the cumulative number of autochthonous infections reached 2,000. Thus the total outbreak time may differ across simulations. We held R_0 constant throughout each simulation, as we sought to model early outbreak dynamics over short periods (relative to the seasonality of transmission) following introduction. We classified simulations as either epidemics or self-limiting outbreaks; epidemics were simulations that fulfilled two criteria: reached 2,000 cumulative autochthonous infections and had a maximum daily prevalence (defined as the number of current infectious cases) exceeding 50 autochthonous cases (Fig A.4 and A.6). The second criterion distinguishes simulations resulting in large self-sustaining outbreaks (that achieve substantial peaks) from those that accumulate infections through a series of small, independent clusters (that fail to reach the daily prevalence threshold). The latter occurs occasionally under low R_0s and high importation rates scenarios.

To verify that our simulations do not aggregate cases from clear temporally separate clusters, we calculated the distribution of times between sequential cases (Fig A.5). In our simulated epidemics, almost all sequentially occurring cases occur within 14 days of each other, consistent with the CDC's threshold for identifying local transmission events (based on the estimated maximum duration of the ZIKV incubation period) [52].

2.3.5 Outbreak Analysis

Our stochastic framework allows us to provide multiple forms of real-time county-level risk assessments as reported cases accumulate. For each county, we found the probability that an outbreak will progress into an epidemic, as defined above, as a function of the number of reported autochthonous cases. We call this epidemic risk. To solve for epidemic risk in a county following the xth reported autochthonous case, we first find all simulations that experience at least x reported autochthonous cases, and then calculate the proportion of those that are ultimately classified as epidemics. For example, consider a county in which 1,000 of 10,000 simulated outbreaks reach at least two reported autochthonous cases and only 50 of the 1,000 simulations ultimately fulfill the two epidemic criteria; the probability of detecting two cases in the county would be 10% and the estimated epidemic risk following two reported cases in that county would be 5%. This simple epidemic classification scheme rarely misclassifies a string of small outbreaks as an epidemic, with the probability of such an error increasing with the importation rate. For example, epidemics should not occur when $R_0 = 0.9$. If the importation rate is high, overlapping series of moderate outbreaks occasionally meet the two epidemic criteria. Under the highest importation rate we considered (0.3 cases/day), only 1% of outbreaks were misclassified.

This method can be applied to evaluate universal triggers (like the recommended two-case trigger) or derive robust triggers based on risk tolerance of public health agencies. For example, if a policymaker would like to initiate interventions as soon as the risk of an epidemic reaches 30%, we would simulate local ZIKV transmission and solve for the number of reported cases at which the probability of an epidemic first exceeds 30%. Generally, the recommended triggers decrease (fewer reported cases) as the policymaker threshold for action decreases, (e.g. 10% versus 30% threshold) and as the local transmission potential increases (e.g. $R_0 = 1.5$ versus $R_0 = 1.2$).

2.4 Results

ZIKV importation risk within Texas is predicted by variables reflecting urbanization, mobility patterns, and socioeconomic status (Table 2.2), and is concentrated in metropolitan counties of Texas (Fig 2.2A). In comparing the predictions of this model to out-of-sample data from April to September 2016,

Variables ordered by importance
Total Amount of County Direct Spending on Traveling ($K)
Percentage Population holding Graduate or professional degree
Total Amount of Visitor Tax Receipts(Local) ($K)
County Male Population
Population Commuting to Work with Other Means
Max Temperature of Warmest Month
Percentage Population below Poverty Level
Precipitation of Wettest Quarter
Population without Health Insurance
Population holding Graduate or professional degree

Table 2.2: **Import risk model variables.** These 10 variables were selected from 72 variables using a combination of representative variables selection and predictive backwards selection. The importance of each variable (from top to bottom) is determined by order of exclusion in backwards selection, with the most important variables remaining in the model the longest.

the model underestimated the statewide total number of importations (81 vs 151), but robustly predicted the relative importation rates between counties ($\beta = 0.97$, $R^2 = 0.74$, $p < 0.001$). The two highest risk counties – Harris, which includes Houston, and Travis, which includes Austin – have an estimated 27% and 10% chance of receiving the next imported Texas case respectively and contain international airports.

ZIKV transmission risk is concentrated in southeastern Texas (Fig 2.2B), partially overlapping with regions of high importation risk (Fig 2.2A). Our county-level estimates of R_0 range widely (from 0.8 to 3.1 for the highest-risk county), reflecting the uncertainty in socioeconomic and environmental drivers of ZIKV (Fig A.2). We therefore analyzed the relative rather than

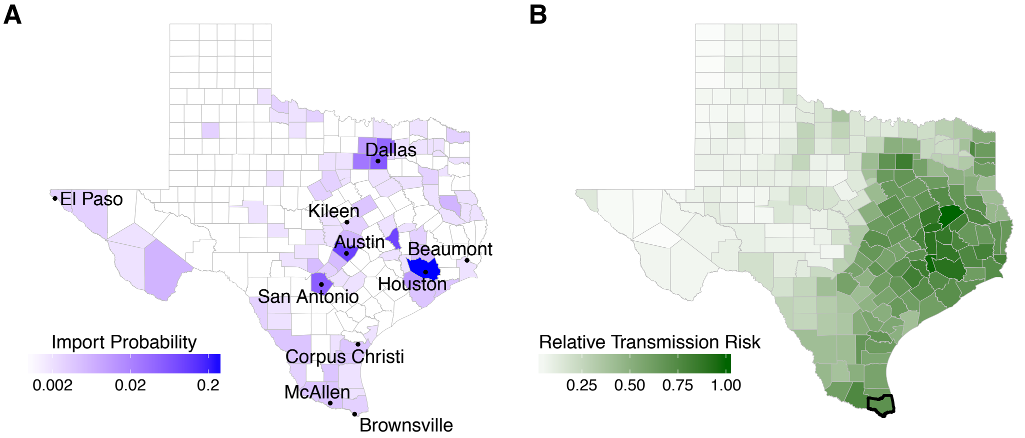

Figure 2.2: **ZIKV importation and transmission risk estimates across Texas for August 2016.** (A) Color indicates the probability that the next ZIKV import will occur in a given county for each of the 254 Texas counties. Probability is colored on a log scale. The 10 most populous cities in Texas are labeled. Houston's Harris County has 2.7 times greater chance than Austin's Travis County of receiving the next imported case. (B) Estimated county-level transmission risk for ZIKV (See Fig A.7 for seasonal differences). Harris county and Dallas County rank among the top 5 and top 10 for both importation and transmission risk respectively; counties in McAllen and Houston metropolitan area rank among the top 20. Bolded county border indicates counties with recorded local ZIKV transmission.

absolute transmission risks. For purposes of demonstration, we assumed a plausible maximum county-level R_0 of 1.5, which closely followed our median estimates, and scaled the transmission risk for each county accordingly. The following risk analyses can be readily refined as we gain more precise and localized estimates of ZIKA transmission potential.

Wide ranges of outbreaks are possible under a single set of epidemiological conditions (Fig 2.3A). The relationship between what policymakers

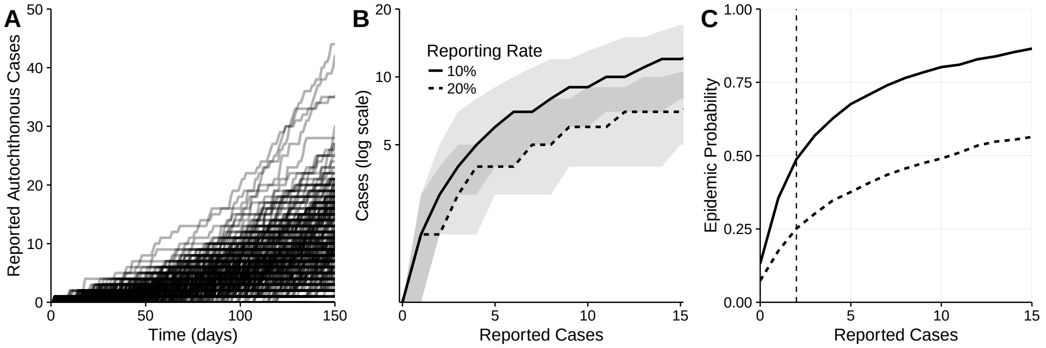

Figure 2.3: **Real-time risk-assessment for ZIKV transmission.** All figures are based on transmission and importation risks estimated for Cameron County, Texas. (A) Two thousand simulated outbreaks. (B) Total number of (current) autochthonous cases as a function of the cumulative reported autochthonous cases, under a relatively high (dashed) or low (solid) reporting rate. Ribbons indicate 50% quantiles. (C) The increasing probability of imminent epidemic expansion as reported autochthonous cases accumulate for a low (solid) and high (dashed) reporting rate. Suppose a policy-maker plans to trigger a public health response as soon as a second case is reported (vertical line). Under a 10% reporting rate, this trigger would correspond to a 49% probability of an ensuing epidemic. Under a 20% reporting rate, the probability would be 25%.

can observe (cumulative reported cases) and what they wish to know (current underlying disease prevalence) can be obscured by such uncertainty, and will depend critically on reporting rates (Fig 2.3B). Under a scenario estimated for Cameron County which experienced the only autochthonous ZIKV transmission in Texas and with a 20% reporting rate, ten linked and reported autochthonous cases correspond to 6 currently circulating cases with a 95% CI of 1-16 from inherent, early-stage outbreak stochasticity. From this wide range of outbreak trajectories, we can characterize time-varying epidemic risk as cases accumulate in a given county. We track the probability of epidemic expansion following each additional reported case in high and low reporting rate scenarios (Fig 2.3C).

These curves can support both real-time risk assessment as cases accumulate and the identification of surveillance triggers indicating when risk exceeds a specified threshold. For example, suppose a policymaker wanted to initiate an intervention upon two reported cases, this would correspond with a 49% probability of an epidemic if 10% of cases are reported, but only 25% if the reporting rate is doubled.. Alternatively suppose a policy maker wishes to initiate an intervention when the chance of an epidemic exceeds 50%. In the low reporting rate scenario, they should act immediately following the third autochthonous reported case, but could wait until the eleventh case with the high reporting rate.

To evaluate a universal intervention trigger of two reported autochthonous cases, we estimate both the probability of two reported cases in each county and the level of epidemic risk at the moment the trigger event occurs (second case reported). Assuming a baseline importation rate extrapolated from importation levels in March 2016 to August 2016, county R_0 scaled from a maximum of 1.5, and a 20% reporting rate, only a minority of counties are likely to experience a trigger event (Fig 2.4A). While 247 of the 254 counties (97%) have non-zero probabilities of experiencing two reported autochthonous cases, only 86 counties have at least a 10% chance of such an event (assuming they experience at least one importation), with the remaining 168 counties having a median probability of 0.0038 (range 0.0005 to 0.087). Assuming that

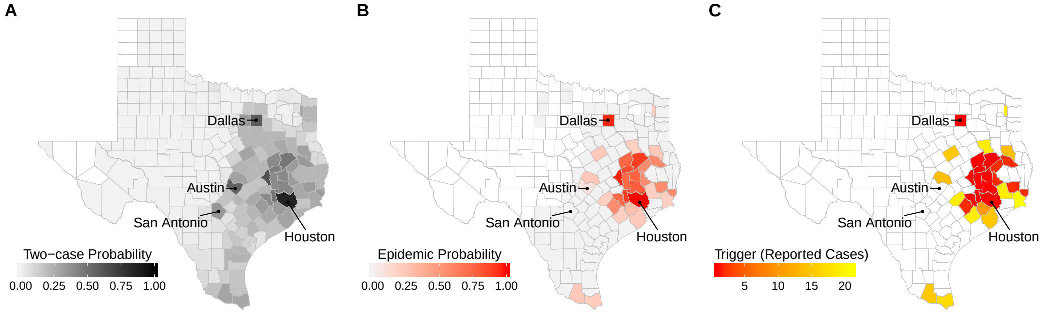

Figure 2.4: **Texas county ZIKV risk assessment.** (A) Probability of an outbreak with at least two reported autochthonous ZIKV cases. (B) The probability of epidemic expansion at the moment the second autochthonous ZIKV case is reported in a county. White counties never reach two reported cases across all 10,000 simulated outbreaks; light gray counties reach two cases, but never experience epidemics. (C) Recommended county-level surveillance triggers (number of reported autochthonous cases) indicating that the probability of epidemic expansion has exceeded 50%. White counties indicate that fewer than 1% of the 10,000 simulated outbreaks reached two reported cases. All three maps assume a 20% reporting rate and a baseline importation scenario for August 2016 (81 cases statewide per 90 days) projected from historical arbovirus data.

a second autochthonous case has indeed been reported, we find that the underlying epidemic risk varies widely among the 247 counties, with most counties having near zero epidemic probabilities and a few counties far exceeding a 50% chance of epidemic expansion. For example, two reported autochthonous cases in Harris County, correspond to a 99% chance of ongoing transmission that would proceed to epidemic proportions without intervention, with the rest of the Houston metropolitan also at relatively high risk ranging from 0 (Galveston) to 90% (Waller) (Fig 2.4B).

Given that a universal trigger may signal disparate levels of ZIKV risk, policy makers might seek to adapt their triggers to local conditions. Suppose a policymaker wishes to design triggers that indicate a 50% chance of an emerging epidemic (Fig 2.4C). Under the baseline importation and reporting rates, an estimated 31 of the 254 counties in Texas are expected to reach a 50% epidemic probability, with triggers ranging from one (Harris County) to 21 (Jefferson County) reported autochthonous cases, with a median of two cases. Counties who detect cases simply due to high importation rates do not have triggers, and the magnitude of a trigger helps quantify a county's absolute risk for an epidemic as a function of the reported autochthonous cases.

2.5 Discussion

Our framework provides a data-driven approach to estimating ZIKV emergence risks from potentially sparse and biased surveillance data [19, 28]. By mapping observed cases to current and future risks, in the face of considerable uncertainty, the approach can also be used to design public health action plans and evaluate the utility of local versus regional triggers. We demonstrate its application across the 254 ecologically and demographically diverse counties of Texas, one of the two states that has sustained autochthonous ZIKV outbreaks [159, 160]. The approach requires local estimates of ZIKV importation and transmission rates. For the Texas analysis, we developed a novel model for estimating county-level ZIKV importation risk and applied published methods to estimate relative county-level transmission risks (Fig 2.2). We expect that most Texas counties are not at risk for a sustained ZIKV epidemic (Fig 2.4), and find that many of the highest risk counties lie in the southeastern region surrounding the Houston metropolitan area and the lower Rio Grande valley. However, R_0 estimates are uncertain, leaving the possibility that the R_0 could be as high as other high risk regions that sustained epidemics [2, 86, 131]. Our analysis is consistent with historic DENV and CHIKV outbreaks and correctly identifies Cameron county, the only Texas county to have reported local transmission, as a potential ZIKV hot-spot, especially when November estimates are used [161] (Fig A.7).

Surveillance triggers, guidelines specifying situations that warrant intervention, are a key component of many public health response plans. Given the urgency and uncertainty surrounding ZIKV, universal recommendations can be both pragmatic and judicious. To assist Texas policymakers in interpreting the two-case trigger for intervention guidelines issued by the CDC [52], we used our framework to integrate importation and transmission risks and assess the likelihood and implication of a two-case event for each of Texas' 254 counties, under a scenario projected from recent ZIKV data to August 2016. Across counties, there is enormous variation in both the chance of a trigger and the magnitude of the public health threat if and when two cases are reported. Given this variation, rather than implement a universal trigger, which may correspond to different threats in different locations, one could design local surveillance triggers that correspond to a universal risk threshold. Our modeling framework can readily identify triggers (numbers of reported cases) for indicating any specified epidemic event (e.g., prevalence reaching a threshold or imminent epidemic expansion) with any specified risk tolerance (e.g., 10% or 50% chance of that the event has occurred), given local epidemiological conditions. We found close agreement between the recommended two-case trigger and our epidemic derived triggers based on a 50% probability of expansion. Of the 30 counties with derived triggers, the median trigger was 2, ranging from one to 21 reported autochthonous cases. These findings apply only to the early, pre-epidemic phase of ZIKV in Texas, when importations occur primarily via travel from affected regions outside the contiguous US.

These analyses highlight critical gaps in our understanding of ZIKV biology and epidemiology. The relative transmission risks among Texas counties appear fairly robust to these uncertainties, allowing us to identify high risk regions, including Cameron County in the Lower Rio Grande Valley. Public health agencies might therefore prioritize such counties for surveillance and interventions resources. Given the minimal incursions of DENV and CHIKV into Texas over that past eleven years since the first DENV outbreak in Cameron County, and the high number of importations into putative hotspot counties without autochthonous transmission, we suspect that, if anything, we may be underestimating the socioeconomic and behavioral impediments to ZIKV transmission in the contiguous US. Our analysis also reveals the significant impact of the reporting rate on the timeliness and precision of detection. If only a small fraction of cases are reported, the first few reported cases may correspond to an isolated introduction or a growing epidemic. In contrast, if most cases are reported, policymakers can wait longer for cases to accumulate to trigger interventions and have more confidence in their epidemiological assessments. ZIKV reporting rates are expected to remain low, because an estimated 80% of infections are asymptomatic, and DENV reporting rates have historically matched its asymptomatic proportion [40, 44]. Obtaining a realistic estimate of the ZIKV reporting rate is arguably as important as increasing the rate itself, with respect to reliable situational awareness and forecasting. An estimated 8-22% of ZIKV infections were reported during the 2013-2014 outbreak in French Polynesia [86]; however estimates ranging from 1 to 10% have been reported during the ongoing epidemic in Columbia [131, 170]. While these provide a baseline estimate for the US, there are many factors that could increase (or decrease) the reporting rate, such as ZIKV awareness among both the public and health-care practitioners, or active surveillance of regions with recent ZIKV cases. Our analysis assumes that all counties have the same case detection probabilities. However, only 40 of the 254 Texas counties maintain active mosquito surveillance and control programs, potentially leading to differences in case detection rates and surveillance efficacy throughout the state [147]. Thus, rapid estimation of the reporting rate using both traditional epidemiological data and new viral sequenced based methods [142] should be a high priority as they become available.

Our framework can support the development of response plans, by forcing policymakers to be explicit about risk tolerance, that is, the certainty needed before sounding an alarm, and quantifying the consequences of premature or delayed interventions. For example, should ZIKV-related pregnancy advisories be issued when there is only 5% chance of an impending epidemic? 10% chance? 80%? A policymaker has to weigh the costs of false positives–resulting in unnecessary fear and/or intervention–and false negatives–resulting in suboptimal disease control and prevention–complicated by the difficulty inherent in distinguishing a false positive from a successful intervention. The more risk averse the policymaker (with respect to false negatives), the earlier the trigger should be, which can be exacerbated by low reporting rates, high importation rate, and inherent ZIKV transmission potential. In ZIKV prone regions with low reporting rates, even risk tolerant policymakers should act quickly upon seeing initial cases; in lower risk regions, longer waiting periods may be prudent.

Chapter 3

Early Prediction of Antigenic Transitions for Influenza A H3N2

3.1 Abstract [1]

Influenza A/H3N2 is a rapidly evolving virus which experiences major antigenic transitions every two to eight years. Anticipating the timing and outcome of transitions is critical to developing effective seasonal influenza vaccines. Using simulations from a published phylodynamic model of influenza transmission, we identified indicators of future evolutionary success for an emerging antigenic cluster. The eventual fate of a new cluster depends on its initial epidemiological growth rate, which is a function of mutational load and population susceptibility to the cluster, along with the variance in growth rate across co-circulating viruses. Logistic regression can predict whether a cluster at 5% relative frequency will eventually succeed with $\sim$ 80% sensitivity, providing up to eight months advance warning. As a cluster expands, the predictions improve while the lead-time for vaccine development and other interventions decreases. By focusing surveillance efforts on estimating population-wide sus-

---

[1] At the time of submitting this dissertation, this chapter was published as Castro L, Bedford T, Meyers L. Early Prediction of Antigenic Transitions for Influenza A H3N2. bioRxiv. Cold Spring Harbor Laboratory; 2019; doi:10.1101/558577 [26]

ceptibility to emerging viruses, we can better anticipate major antigenic transitions.

3.2 Introduction

Seasonal influenza A/H3N2 causes significant annual morbidity and mortality worldwide, as well as severe economic losses [102]. In the United States, the 2017-2018 season was unusually long and severe, lasting over 16 weeks and causing over 900,000 hospitalizations and 80,000 fatalities, including 183 pediatric deaths [30, 31, 110]. The global health community continually tracks H3N2 and annually updates H3N2 vaccines. However, annual influenza epidemics continue to impart a significant public health burden. The rapid antigenic evolution of the influenza virus via mutations in hemagglutinin (HA) glycoproteins and neuraminidase (NA) enzymes [66, 115], and logistical requirement of selecting vaccine strains almost a year prior to the flu season pose a significant challenge. Vaccines target the antigen-binding regions of dominant influenza subtypes. While a particular subtype may circulate for a few years, strong positive selection for new antigenic variants will eventually produce antigenic drift [15, 23, 36], rendering a vaccine less effective if new mutations in the antigen-binding regions are not included in vaccine chosen strains [14, 24]. The typical reign of a dominant subtype ranges from two to eight years [13, 80]. From 2004 to 2018, seasonal influenza vaccines have had an estimated average efficacy of 40.56% against all influenza strains included in the vaccine [14].

The World Health Organization's Influenza Surveillance and Response System (GISRS) coordinates influenza surveillance efforts to survey and characterize the diversity of influenza viruses circulating in humans. Viral samples are rapidly analyzed via sequencing of HA and NA genes, serologic assays, and other laboratory tests to identify newly emerging antigenic clusters. Within the past decade, the number of complete HA gene sequences in the GISAID EpiFlu [20, 146] database has increased tenfold, from fewer than 1,000 in 2010 to over 10,000 in 2017 [104]. Molecular data at high spatiotemporal resolution could potentially revolutionize influenza prediction. However, the research and public health communities have just begun to determine effective strategies for extracting and integrating useful information into the vaccine selection process.

Phylodynamic models describe the interaction between the epidemiological and evolutionary processes of a pathogen [61]. The availability of molecular data coupled with the recent development of detailed, data-driven phylodynamic models has galvanized the new field of viral predictive modeling [53, 81, 87, 137]. These models aim to predict the future prevalence of specific viral subtypes based on past and present molecular data. For example, one approach generates one-year ahead forecasts of clade frequency using a fitness model parameterized by the number of antigenic and genetic mutations that dictate the virus' antigenicity and stability respectively [95]. Another method maps antigenic distance from hemagglutination inhibition (HI) assay data onto an HA genealogy to determine whether the changes in antigenicity among high-growth clades necessitate a vaccine composition update [153]. A third model predicts which clade will be the progenitor lineage of the subsequent influenza season by estimating fitness using a growth rate measure derived from topological features of the HA genealogy [114]. All three approaches have been tested on historical predictions. The Łuksza & Lässig model [95] predicted positive growth for 93% of clades that increased in frequency over one year. Steinbruk et al. [153] predicted the predominant HA allele over nine influenza seasons with an accuracy of 78%. Both Łuksza's & Lässig's [95] and Neher et al. [114] model predictions of progenitor strains to the next season's performed similarly. Since 2015, both these models have been used to provide recommendations on vaccine composition for the upcoming flu seasons [11, 65, 111].

Taken together, this body of work points to the promise of predictive evolutionary models. Phylodynamic simulation models provide a complementary window into the molecular evolution of emerging viruses. By observing influenza evolution *in silico*, we can take rigorous experimental approach to test hypotheses about early indicators of cluster [29,30] success and design surveillance strategies to inform vaccine strain selection. Here, we simulate decades of H3N2 phylodynamics using a published model [12, 82] and analyze the simulated data to identify early predictors of a cluster's evolutionary fate. Viral growth rates, both for an emerging cluster and its competitors, are the most robust predictors of future ascents. When a new antigenic cluster first appears at low frequency (e.g., 1% of sampled viruses), our models can predict whether it will eventually rise to dominance (e.g., maintain a relative frequency greater than 20% of sampled viruses for at least 45 days) with reasonable confidence and advanced warning. To translate these findings into actionable guidance for global influenza surveillance, we also evaluate proxy indicators that can be readily estimated from current data, quantify limits in the accuracy, precision and timeliness of predictions, and construct models to predict future frequencies of emerging clusters.

3.3 Methods

3.3.1 Simulation Model

We implemented a published stochastic individual-based susceptible-infected (SI) phylodynamic model of influenza A/H3N2 [12, 82] to repeatedly simulate 30 years of transmission in a constant population of 40 million hosts with birth and death dynamics (Fig. 3.1). In brief, each individual host is characterized by its infection status – susceptible or infected — and a history of prior viral infections. Viruses are defined by a discrete antigenic phenotype, which determines the degree of immune escape from other phenotypes, and a deleterious genetic mutation load (k) which affects the virus' transmissibility. Antigenic mutations occur stochastically and confer advanced antigenicity to the virus. The probability that a given virus will infect a given host is determined by how similar the antigenic phenotype of the challenging virus is to the antigenic phenotype of the host's most related previous infection. This probability, or degree of immune escape, is tracked through the simulation by the evolutionary history of clusters (parent-child relationships). Antigenic and deleterious non-antigenic mutations occur only during transmission events; the model assumes that viruses within a single individual host are genotypically homogeneous. The model also assumes no co-infection, no seasonal forcing [45, 144], and no short-term immunity that would broadly prevent reinfection after recovering from infection. We used the baseline parameters chosen in Koelle & Rasmussen [82] based on empirical, epidemiological, and virological estimates [8, 16, 25, 140].

3.3.2 Simulated Data

We ran 100 replicate simulations and selected a subset that produced realistic global influenza dynamics. Specifically, we excluded 38 simulations in which endemic transmission died out prior to the 30 years. We treated the first five years of each simulation as burn-in periods. In total, we analyzed 1550 years of simulated influenza transmission and evolutionary dynamics.

Throughout each simulation, we tracked 23 metrics reflecting the epidemiological state of the host population (i.e., number of susceptible and infected individuals) and evolutionary state of the viral population (Table B.1) at 14 day intervals. When possible, we monitored these quantities for both individual antigenic clusters and the entire viral population, and then calculated their ratio. For example, we monitored the average number of deleterious mutations within each antigenic cluster and across all viruses, as well as the relative mutational load of each cluster with respect to the entire viral population. Henceforth, we refer to the metrics as candidate predictors.

3.3.3 Classifying Evolutionary Outcomes

We classified each novel antigenic cluster in each simulation into one of three categories: (1) rapidly eliminated clusters that never reach 1% relative frequency in the population, (2) transient clusters that surpass 1% relative frequency but do not qualify as established clusters, and (3) established clusters that circulate above 20% relative frequency for at least 45 days. With this criteria, transient and established clusters constituted on average 81% of the infections at any point in time (Figures B.1 and B.2).

3.3.4 Predictive Models

Restricting our analysis to transient and established clusters, we used generalized linear modeling to identify important early predictors of evolutionary fate. For each antigenic cluster, we predicted its evolutionary future (i.e., whether it ultimately becomes established) at specified surveillance thresholds, such as 5% relative frequency. Specifically, we recorded all candidate epidemiological and evolutionary predictors at the moment each cluster crossed the threshold. We analyzed all ten surveillance thresholds ranging from 1% to 10% at 1% increments. For each surveillance threshold, we centered and scaled candidate predictors and removed collinear factors. Using five-fold cross validation, we partitioned the data into five subsets, keeping data from individual simulations in the same subsets. We fit mixed-effects logistic regression models using four subsets for training and controlling for differences between independent simulations. Predictors were added sequentially based on which term most significantly lowered the average Akaike Information Criterion of the five training folds.

We evaluated model performance by predicting the evolutionary outcomes of clusters in the held-out test subset. We calculated three metrics: the area under the receiver operating curve (AUC), the sensitivity (the proportion of all positives predicted as positive), and the positive predictive value (the proportion of true positives of all predicted positives). The model predicts the probability that a cluster will establish. To translate these outputs into discrete binary predictions of future success, we applied a probability threshold which maximized the F1 score [150], which is the harmonic average of a model's positive predictive value and sensitivity (Table B.2). When we included historical data of candidate predictors, i.e the value of a candidate predictor at an earlier surveillance threshold, model performance did not have a significant difference (Fig. B.3).

We also considered an opportunistic sampling regime, where samples are tested as they arise regardless of their relative frequency. We fit models aimed at two prediction targets: (1) the evolutionary success of a cluster sampled at an arbitrary relative frequency and (2) the frequency of a cluster up to twelve months into the future. We built models based on data sampled from ten random time points in each of the 62 25-year simulations. We considered all clusters present above 1% relative frequency but not yet established as a dominant cluster. The frequency of a cluster at the time of sampling was included as an additional predictor. To predict the frequency of an antigenic cluster X months into the future, we fit a two-part model that first predicted whether the cluster would be present at the specified date, and, if so, then estimated the frequency of the cluster at that date. We used forward variable selection and cross validation model, as described above. We used the R statistical language version 3.3.2 [158] for all analyses, and the afex package for generalized linear models [149].

3.3.5 Candidate predictors

Reproductive Rates In our simulated data, we can calculate the instantaneous reproductive rate for particular clusters and the entire viral population. As described in Koelle & Rasmussen [82], the reproductive rate of a virus is given by,

$$R(v) = \frac{\beta_0\big(1 - s_d\big)^{k(v)}}{\mu + \nu}\Big(\frac{S_{\text{eff}}(v)}{N}\Big) \tag{3.1}$$

where β_0 is the inherent transmissibility, s_d is the fitness effect for each of the virus' $k(v)$ deleterious mutations, μ and ν are the per capita daily death and recovery rates, respectively, and N is the host population size. We assume that β_0, s_d, μ and ν are constant across all viruses. $S_{\text{eff}}(v)$ denotes the population-wide susceptibility to the virus accounting for cross-immunity from prior infections, herein referred to as the effective susceptibility, and the population level effective susceptibility is estimated for a virus as,

$$S_{\text{eff}}(v) = \frac{S}{N}\sum_{h=1}^{N}\sigma_v(h) \tag{3.2}$$

where $\sigma_v(h)$ is the immunity of host h towards virus v based on the antigenic similarity between v and the virus in host h's infection history most antigenically similar to virus v. A $\sigma_v(h) = 1$ indicates full susceptibility, while $\sigma_v(h) = 0$ indicates complete immunity.

The growth rate of an antigenic cluster is then the average over all viruses in that cluster, given by,

$$R_c = \frac{1}{I_c} \sum_{i=1}^{I_c} R(v_i) \tag{3.3}$$

where I_c is the number of hosts infected by a virus from cluster c and v_i is the virus infecting host i. Likewise the population-wide average ($\langle R \rangle$) and variance ($\mathrm{var}(R)$) in are computed across all current infections, and the relative reproductive rate of a cluster is given by$R_c/\langle R \rangle$.

Practical approximations Equations (3.1)–(3.3) are not easily calculated from current surveillance data. Therefore, we considered two proxy measures of viral growth rates and two proxy measures of viral competition. We first choose two surveillance thresholds, for example, 6% and 10%. When the relatively frequency of a cluster crosses the second threshold, we calculate both the time elapsed since it crossed the first threshold and the relative fold change, as given by,

$$\chi_c(t_1, t_2) = \frac{\Delta_c(t_1, t_2)}{\frac{1}{N_c} \sum_{j=1}^{N_c} \Delta_j(t_1, t_2)} \tag{3.4}$$

where $t(1)$ and $t(2)$ are the times at which cluster c crossed the first and second threshold, respectively, $\Delta_c(s, t)$ is its relative frequency at time t divided by its relative frequency at time s and N_c is the number of distinct clusters present at both time $t(1)$ and $t(2)$. For the competition proxy measures, we calculate the the variance in $\chi_c(t_1, t_2)$ and the N_c where $\Delta_c(s,t) > 1$.

We evaluate the performance of these approximations by comparing logistic regression models that predict whether a cluster will establish from either the true $R_c/\langle R \rangle$ at the 10% surveillance threshold, the relative fold change between the 6% and 10%, or time elapsed between reaching the 6% and 10% thresholds. As before, we evaluated model performance based on AUC, positive predictive value, and sensitivity.

3.4 Results

Our simulations roughly reproduce the global epidemiological and evolutionary dynamics of H3N2 influenza over a 25 year period. Without seasonal forcing, prevalence rises and falls, peaking every 3.2 years on average (s.d. = 1.6). These dynamics reflect the turnover and competition of antigenic clusters. The median of the most recent common ancestor (TMRCA) in our simulations was 5.9 years (IQR 4.62 - 7.9), which is higher than empirical estimates of 3.89 years [13]. The median life span of established clusters was 1128 days (s.d. = 480), corresponding to roughly 3.5 years. However, the annual incidence of influenza in our model (4.0%, 95% $CI 0.37 - 9.7\%$) was lower than empirical annual incidence estimates of $9 - 15\%$ [13]. Given the model only simulates the transmission of H3N2 and not all circulating flu types, our annual incidence is comparable to empirical estimates [120]. We assume that clusters become detectable once they cross a relative frequency threshold of 1% and are fully established if they maintain a relative frequency above 20% for at least 45 weeks. In our simulations, 2% of the approximately 200 novel antigenic clusters per year overcome early stochastic loss to reach detectable levels. As the relative frequency of a newly emerging cluster increases, the probability that the cluster will ultimately establish also increases. There is an inverse relationship between the number of clusters that reach a threshold and the probability of future success. For example, far fewer clusters reach a relative frequency of 10% than 1%. If a cluster succeeds in reaching relative frequency thresholds of $1\%, 6\%$, and 10%, its probability of establishing increases from 13% to 50% to 67% (Fig. B.2). Our model classifies clusters as either positives that are likely to establish or negatives that are expected to circulate only transiently. As we increase the surveillance threshold, the fraction of successful clusters that are misclassified as negatives decreases. In a representative out-of-sample 25-year simulation, 17 of 132 detectable clusters eventually rose to dominance (Fig. 3.1). Of these, 65% and 88% were correctly predicted when they reached the 1% and 10% surveillance threshold, respectively. The number of true negative events decreased considerably, from 109 at the 1% surveillance threshold to only 11 at the 10% surveillance threshold, while the other types of events held relatively constant.

Across all surveillance thresholds, the first four predictors chosen through forward model selection are the relative growth rate of the focal cluster $(R_c/\langle R\rangle)$, the background variance $(\mathrm{var}(R))$ and mean $\langle R\rangle$ of viral growth rates,and the

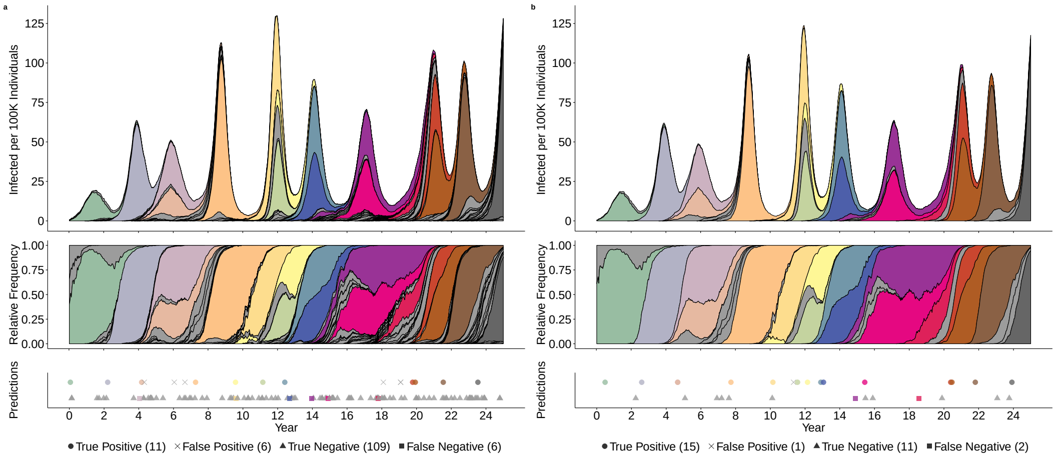

Figure 3.1: **Out-of-sample predictions of antigenic cluster evolutionary success at relative frequency thresholds of 1% (a) and 10% (b).** Grey shading indicates clusters that surpass the surveillance threshold, but do not establish. Other colors correspond to distinct antigenic clusters that eventually establish. The top time series graphs depict the absolute prevalence of antigenic clusters; the middle graphs give their relative frequencies. The bottom panels indicate the timing and accuracy of out-of-sample predictions based on the optimized model for each surveillance threshold. The top row of symbols indicate clusters predicted to succeed, with true positives indicated by circles and false positives indicated by crosses; the bottom row indicates clusters predicted to circulate only transiently, true negatives indicated by triangles and false negatives indicated by squares. The number of predictions in each category is provided in the legend.

	Predictor	**Symbol**	**Models Included (Surveillance Threshold %)**	**Coefficient Estimate**
All Models	1. Relative Growth Rate (R)	$R_c/\langle R\rangle$	1-10	[2.3, 2.64]
	2. Variance in population R	var(R)	1-10	[-0.72, - 0.49]
	3. Population R	$\langle R\rangle$	1-10	[0.32, 0.42]
	4. Relative mutational load	$k_c/\langle k\rangle$	1-10	[-0.34, -0.21]
Some Models	Relative variance in transmissibility	var(β_c)/var(β)	1-10	[0.17, 0.34]
	Variance in susceptibility to cluster c	var(σ_c)	1-6	[0.16, 0.20]
	Frequency of current dominant cluster	I_c/I	3,5,8,9	[0.14, 0.21]
	Proportion of individuals infected	I/N	2	[-0.17]
	Total number of individuals infected	I	3,4	[-0.17, -0.16]
	The most recent common ancestor	tMRCA	10	-0.16
	Relative variance in susceptibility*	var(σ_c)/var(S_{eff})	1	[0.12, 0.16]

Table 3.1: **Predictors selected by five-fold cross validation and forward selection.** The top four variables were selected in the identical order (as listed) across all surveillance threshold models. The fifth predictor, relative variance in transmissibility, was included in all models, but not always as the fifth chosen. In the formulas, *c* refers to cluster-level quantities. The rightmost column gives the full range of fitted coefficients (log-odds) across all models based on across the five-fold cross validation for each surveillance thresholds' final model. var(σ_c) was calculated across all hosts; var$(\mathrm{S_{eff}})$ was calculated across only infected hosts. I = number of infected hosts, N = total number of hosts, σ_c = effective susceptibility to infection by cluster c, β^{k} = the transmission rate of the virus carrying k deleterious mutations.Formulas to calculate each quantity are in Table B.1.

relative deleterious mutational load of the focal cluster ($k_c/\langle k\rangle$). Population-level epidemiological quantities were only selected for models at low surveillance thresholds ($2-4\%$); in these models, overall prevalence had a slightly negative correlation with future viral success (Table 1). The median number of predictors chosen was 6.5, with a range of 5 to 7. The best fit models are described in Table B.2.

We examine the dynamics of the top two predictors. As newly emerging clusters rise in relative frequency from 1% to 10%, their relative growth rate declines towards one. That is, they approach the population average fitness (Fig. 3.2). The relative growth rate is significantly higher for clusters that will eventually establish than those will burn out, with the separation between the two groups increasing as the clusters ascend in frequency (Fig. 3.2a). This predictor is a composite quantity, estimated based on both mutational load and effective susceptibility. We compare these two quantities at two time points, when the clusters reach 1% and 10% frequencies. Mutational load increases and effective susceptibility decreases in ascending clusters, with more extreme changes occurring in clusters that ultimately fail to establish. We also measure the changes in these two quantities for the entire population, and find that the background mutational load remains relatively constant and background effective susceptibility increases slightly. The background effective susceptibility peaks when a new cluster begins to constitute a major proportion of the circulating types — at this point the immunity from previous infections is not strongly protective against the newly dominant cluster. The decline in cluster fitness likely stems from the accumulation of deleterious mutations and exhaustion of the susceptible population (Fig. 3.2b). While this occurs within both established and transient clusters, the mutational loads in established and transients increase by averages of 1.4 and 2.04 mutations, respectively (Wilcox, $p < 2.2e^{16}$).

The background variance in viral growth rates, var(R), is the second most informative predictor. The lower the variance, the more likely a cluster is to establish. However, it is a weaker predictor than $R_c/\langle R \rangle$'s; the estimated

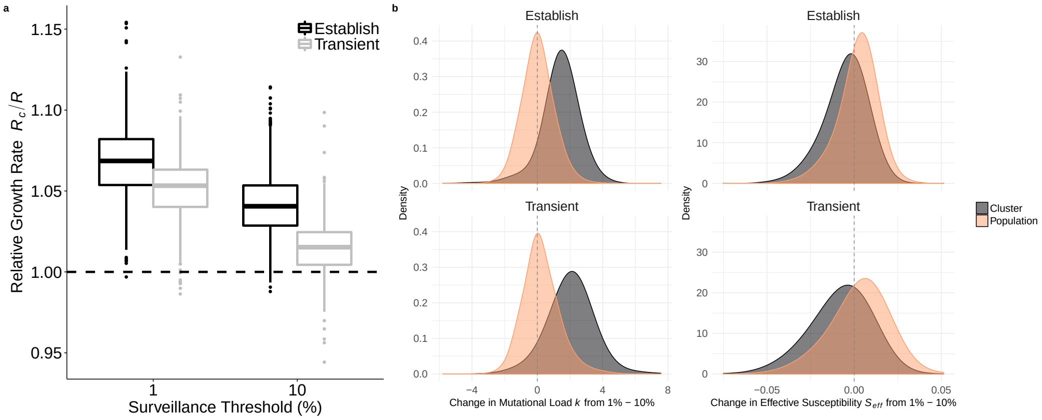

Figure 3.2: **Relative growth rates predict future success.** (a) Clusters that eventually establish have significantly higher $R_c/\langle R\rangle$ than those that fail to establish. As clusters increase in relative frequency from 1% to 10%, their $R_c/\langle R\rangle$ generally declines but the distinction between future successes and future failures becomes more pronounced. (b) R is a composite value based on the mutational load and S_{eff}. We compared the mutational load (left) and S_{eff} (right) of a cluster when it crossed the 1% and 10% thresholds by subtracting the former from the latter (orange distributions); we simultaneously calculated the difference in average mutational load and across the entire viral population (grey distributions). The top and bottom rows shows the distributions of change for clusters that establish and transiently circulate, respectively. The decrease in a cluster's fitness advantage is driven by both increasing mutational load and a decreasing S_{eff}. The background mutational load does not change noticeably, while the background S_{eff} increases slightly.

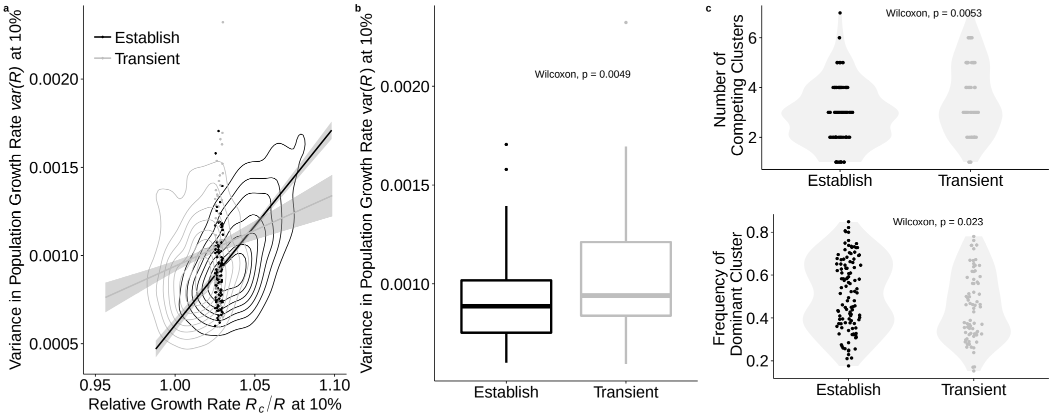

Figure 3.3: **Viral competition predicts future success for clusters with borderline growth rates.** (a) Clusters with only a slight $R_c/\langle R\rangle$ advantage are more likely to establish if the background var(R) is low. Clusters with higher $R_c/\langle R\rangle$ successfully regardless of var(R). Contour lines indicate the density of values of $R_c/\langle R\rangle$ and var(R). The lines represent the correlation between the variables for successful and transient clusters. (Success-black: $r = 0.63$, $p < 2.2e^{-16}$; Transient-grey: $r = 0.18$, $p < 1.2e^{-06}$). The dots represent clusters with $R_c/\langle R\rangle$'s between 1.025-1.030, a range within the individual distributions of $R_c/\langle R\rangle$ for success and transient clusters do not statistically differ (Wilcox, p = 0.46). For clusters falling within this ambiguous range of $R_c/\langle R\rangle$, (b) var(R) is significantly higher in transient clusters than in established clusters, and (c) in comparison to transient clusters, successful clusters tend to face fewer co-circulating clusters (Wilcox, $p = 0.005$), with the current dominant cluster at higher frequency (Wilcox, $p = 0.023$). Points represent the number of circulating clusters and the frequency of the dominant cluster; shading represents the kernel density estimation of the distribution of points. Across all graphs, values are calculated when the focal clusters reach a 10% surveillance threshold.

logit coefficient of the $R_c/\langle R\rangle$ is approximately four times that of var(R) (Table 1). The var(R) tends to increase as a cluster expands from 1% to 10% relative frequency (Wilcox, $p < 2.2e^{-16}$). This may stem from diverging fitnesses of the newly expanding cluster and the receding dominant cluster, which has likely accumulated a considerable deleterious load and burned through much of its susceptible host population. A higher var(R) decreases the probability of a cluster being successful, particularly when a cluster has only a modest growth rate. Clusters with high $R_c/\langle R\rangle$'s are successful even when emerging in highly variant environments (Fig. 3.3a). High variance may reflect high levels of inter-viral competition. If we consider both transient and established clusters with similar $R_c/\langle R\rangle$ (ranging from 1.025 to 1.03), successful clusters encounter significantly fewer co-circulating clusters, and the frequency of the resident dominant cluster is significantly higher (Fig. 3.3c). This may reflect suppression of competition by the dominant cluster, creating a vacuum for a moderately fit cluster to fill.

Using 6271 geographically diverse influenza A/H3N2 sequences sampled from 2006 and 2018, we assessed whether our predictive models can be directly applied to influenza surveillance efforts. Clusters were distinguished by single mutations to epitope sites on the HA1 sequence and successful clusters were those that reached a relative frequency of at least 20% for at least 45 days. Despite sparse sampling, the dynamics of antigenic transitions resemble those produced by our simulations (Fig. 3.4a). Over the 12-year period, dominant clusters circulated for an average of 2.25 years (s.d. 1.17); 44 clusters reached a relative frequency of 10%; 18 of the 44 were eventually successful. For each emerging cluster, we calculated the relative number of epitope mutations by dividing the average number of epitope mutations in viruses within the cluster by the average number found in other co-circulating viruses. For clusters that reached 1% relative frequency, this quantity was less than one for clusters that eventually established (N=18) and greater than one for transient clusters that did not establish (N=1516); this difference is statistically significant (Wilcox, p = 0.0003) (Fig. 3.4b). This difference was not significantly different when measured when clusters reached the 5% relative frequency surveillance threshold. We also fit classifier models to the empirical data using proxies for fitness (e.g., fold change and growth rate between sequential sampling of a cluster) and competition, (e.g., the number of co-circulating clusters). While some of these factors are significant predictors of future evolutionary success, our best models had sensitivity and positive predictive values below 50%.

When forecasting influenza dynamics, there may be trade offs between prediction certainty, the extent advanced warning, and the surveillance effort required to detect and characterize emerging viruses. Across our ten models, there is a marked trade-off between lead-time and reliability, with low surveillance thresholds providing earlier but less accurate indication of future threats (Fig. 3.5). Across simulations, the median time difference between a cluster reaching the 1% and 10% surveillance thresholds was approximately 7 months (IQR: 154-294 days).

Classifier models have substantial discriminatory and predictive power

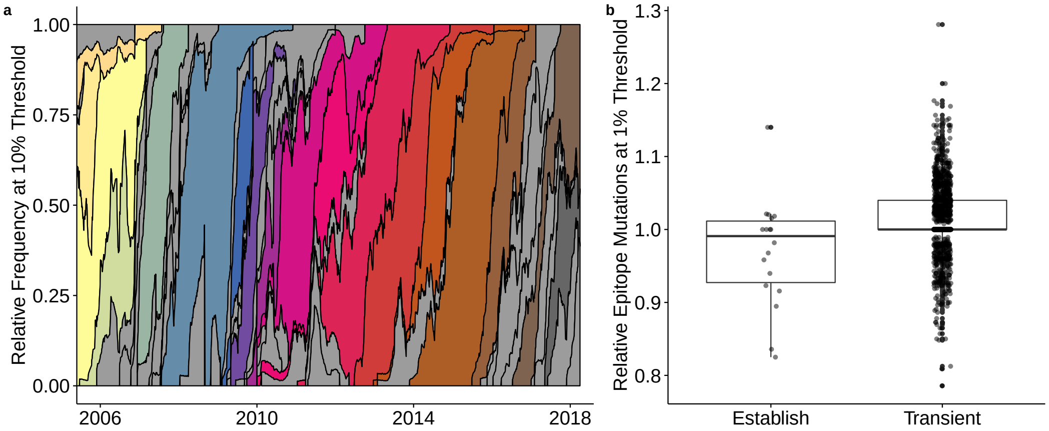

Figure 3.4: **Empirical antigenic dynamics of influenza A/H3N2, 2006-2018.** (a) Relative frequencies of all antigenic clusters that reach the threshold of at least 10% of sampled viruses. Frequencies are calculated using a 60-day sliding window. Grey shading indicates clusters that surpass the 10% threshold, but do not eventually establish (i.e., reach relative frequency of at least 20% for at least 45 days). Other colors indicate distinct antigenic clusters that eventually establish. (b) A low relative number of epitope mutations when a cluster reaches the 1% relative frequency threshold is an early indicator of future success (Wilcox, $p < 0.003$). We divide the number of epitope mutations of a focal cluster by the average number of mutations of simultaneously circulating clusters.

even when an antigenic cluster is present at low frequencies (Fig. 3.5a). Model AUC's tend to decrease as the frequency of the candidate clusters increases. Conversely, the positive predictive value (PPV) and sensitivity increase at higher surveillance thresholds. The gains in sensitivity and PPV per month decrease at higher surveillance thresholds. Between the 1% and 5% surveillance thresholds, there is on average a 4% increase in sensitivity and 4.5% increase in PPV per month lost in lead-time. However, between the 6% and 10% surveillance thresholds, sensitivity gains drop to 1.2% and PPV to 3.6% per month lost in lead-time. This decreasing trade off between gain in certainty and loss of lead-time reflects shorter intervals between surveillance thresholds as the cluster begins to rapidly expand and the model's prediction capabilities reach upper capacity.

The primary predictor across all models—the relative growth rate of a cluster—cannot easily be estimated from available surveillance data. Thus, we built and evaluated bivariate logistic regression models that predict future success using more easily attained proxies (3.4). One considers the time taken for the cluster to rise from 6% to 10% relative frequency and the total number of clusters that grew during this period; the other considers the fold-change in the relative frequency of the cluster between these time points and the background variance in fold-change. Of the four proxies, all but the relative fold-change of the cluster were statistically significant predictors, with negative effects on the probability of cluster success (Fig. B.4). These resulting models have higher sensitivity than positive predictive values. We also tested

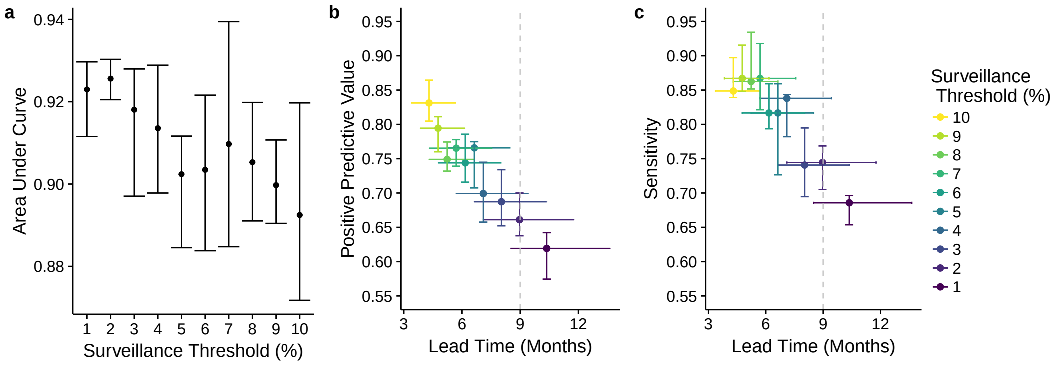

Figure 3.5: **Model performance across surveillance thresholds.** (a) Area under the receiver operator curve (AUC) suggests that models can predict successful from unsuccessful clusters by the time they reach 1% of circulating viruses, with discriminatory power declining slightly as clusters rise in frequency. Bars represent the max, median, and minimum AUC values across 5-fold cross validation. (b-c) There is a trade-off between lead time and model performance. The horizontal bars represent the IQR of time between the moment the expanding antigenic cluster reaches the surveillance threshold and when it reaches the success criteria. Vertical bars represent the range and median positive predictive value (b) and sensitivity (c), across five-fold evaluation. Colors correspond to the best fit model for each surveillance threshold. Dashed gray lines indicate lead times of nine months, which represents the current time between the Northern Hemisphere vaccine composition meeting in February and the following start of the influenza season in October.

Model	**Type**	**AUC**	**PPV**	**Sensitivity**
1. $R_c/\langle R\rangle$ + $\mathrm{var}(R)$	Actual	0.88	0.81	0.89
2. $\delta_c(t_1, t_2)$ + $N_{\Delta_j(t_1,t_2)>1}$	Proxy	0.78	0.74	0.87
3. $\chi_c(t_1, t_2)$ + $\mathrm{var}(\Delta_j(t_1, t_2))$	Proxy	0.67	0.66	0.95

Table 3.2: **Model 1 predicts the fate of a cluster using the top two predictors in our best fit model.** The two proxy models use data from two time points, when the cluster reached relative frequencies of 6% (t_1) and 10% (t_2). Model 2 considers the time elapsed between and the number of competing expanding clusters. Model 3 considers the relative fold change in the focal cluster between the two time points and the population-wide variance in fold change. Performance values are the median of five fold cross-validation.

analogous models using statistics calculated at alternative surveillance checkpoints (1% to 5%, 3% to 5% , and 8% to 10%), and found that the 6% – 10% comparison performed best (Table B.3).

Finally, we fit models to predict the presence and frequency of clusters based on opportunistic sampling of clusters, rather than waiting for specified surveillance thresholds. Cluster frequencies tend to skew towards low frequencies (Fig. B.5). Our best fit model for predicting the future success of all clusters present at a random time point performs comparable to our best models for low surveillance thresholds (Fig. B.6). We fit a second two-part model that sequentially predicts the presence-absence and the frequency of a cluster in three month intervals out to one year ahead. The model predicted up to twelve-month ahead presence-absence with 92% discriminatory power (AUC). However, the accuracy of the frequency predictions declined after six months, with a tendency to underestimate the frequencies of future dominant clusters

(Fig. B.7,Tables B.4 and B.5). The top predictors included the frequency of the cluster at the time of sampling and most of the top predictors selected for the surveillance threshold models.

3.5 Discussion

Until we develop an effective universal flu vaccine, seasonal vaccines will remain the frontline of flu prevention. The severe 2017-2018 flu season was a stark reminder that anticipating dominant strains with sufficient lead time for incorporation into vaccines is paramount to public health. Here, we analyzed over 1500 years of simulated influenza phylodynamics to explore the predictability of antigenic emergence and identify early predictors of future evolutionary success that can be plausibly monitored via ongoing surveillance efforts.

Phylodynamic models provide insight into both the interplay of evolutionary and epidemiological processes and how these dynamics are manifested in observable data. Our simulations revealed a stereotypical path to antigenic turnover consistent with those described in Koelle & Rasmussen [82]. An antigenic mutation appears on a virus. If its fitness is high relative to the competition, it can gain a foothold. In general, the lower the deleterious load and higher the susceptibility in the host population, the higher the fitness of the new virus. Thus, antigenic mutations occurring on good genetic backgrounds are more likely to gain traction [82]. The dynamics of susceptibility are a bit more complex. Although all hosts will be partially susceptible to a new antigenic type, the level of susceptibility will depend on past infections by antigenically similar viruses. We find that some successful mutants arise with markedly higher fitness than other co-circulating viruses, which propels them towards dominance, while others enter with only moderately high fitness but are able to ascend in the wake of a prior antigenic sweep, which suppresses other potentially competing viruses. Many of the mutants that eventually establish as dominant clusters first appear as another dominant cluster is cresting [1, 122]. The rampant transmission affords opportunities for such mutations to arise and quashes other potentially competing clusters. Transmission of the previous cluster type begins to decreases as the population gains immunity through infection. With fewer susceptible hosts available to the previously dominant cluster, the expanding cluster, as well as any competing clusters, begin to constitute a larger fraction of all circulating viruses.

As a cluster expands, it accumulates deleterious mutations. By the time new clusters reach a relative frequency of 10%, their mutational loads begins to approach the population average. Simultaneously, the number of susceptible hosts decreases as the cluster sweeps through the host population. If a cluster reaches a relative frequency of 10%, its probability of future dominance will be influenced by how much of a fitness advantage it retains, and by the level of competition from other clusters.

The strongest predictor of future dominance across all of our models is the relative effective reproductive number of a cluster, that is, the growth rate of the cluster compared to the average growth rate across the viral population. This measure of viral fitness incorporates both the real-time competitive advantage (vis-a-vis the immunological landscape) and deleterious mutational load. Intuitively, faster growing clusters are more likely to persist and expand. Our ability to predict the fate of an emergent virus improves as the cluster increases in relative frequency. Both sensitivity — the proportion of successful clusters detected by the model — and positive predictive value — the proportion of predicted successes that actually establish — surpass 80% by the time a cluster has reached 10% relative frequency.

The second most informative predictor selected across all models — the population-wide variance in the effective growth rate, var(R) — requires a more nuanced interpretation. The greater the background variance at the time a cluster is emerging, the less likely the cluster is to succeed. To unpack this result, we analyzed the competitive environment of emerging clusters with only modest growth rates; rapidly growing clusters are likely to succeed regardless of their competition. Within this class of slowly emerging viruses, those that initially face a single high frequency dominant cluster and fewer co-emerging competitors are more likely to succeed [59, 154]. A recent sweep by a dominant cluster leaves a wake of immunity that can be exploited by antigenically-novel clusters that stochastically battle for future dominance. We hypothesize that these two conditions—a reigning dominant cluster and reduced competition with emerging novelty—reduce the overall variance in viral growth rate and explain the negative correlation between this quantity and the future ascent of an emerging cluster.

While the certainty of our predictions improves as clusters increase in relative frequency, there is a trade-off with lead time. The longer we wait to assess a rising cluster, the less time there will be to update vaccines and implement other intervention measures. For a successful cluster detected at a relative frequency of 1%, there will be, on average, 10 months before the cluster becomes established (maintains a relative frequency over 20% for 45 days). If detected only after reaching a relative frequency of 10%, the expected lead time shrinks to four months. Although real-world surveillance is noisy and dependent on sufficient sampling depth and geographic coverage, our results suggest that, with a perfect knowledge of the host and viral populations, predictions can be made with at least 85% sensitivity and confidence before a cluster rises to 10% of all circulating strains.

As policy-makers consider new strategies for antigenic surveillance and forecasting, the trade-off between prediction accuracy and lead time has practical implications. For example, a detection system targeting new viruses as soon as they reach 1% relative frequency has the benefit of early warning and drawback of low accuracy, which translate into economic and humanitarian costs and benefits. On the positive side, early warning increases the probability that seasonal vaccines will provide a good match with circulating strains, and thus lowers the expected future morbidity and mortality attributable to seasonal flu. Based on the vaccine production and delivery schedule, the surveillance window for emerging clades is from October to February for the Northern Hemisphere and vaccine composition is determined at an international meeting in February [104]. Our analysis suggests that, at nine months before an emerging cluster sweeps to dominance, it is likely to have be circulating at a low relative frequency in the range of 1% to 4%. On the negative side, the low surveillance threshold for candidate clusters and consequent lower accuracy require far more surveillance and vaccine development resources than higher surveillance thresholds. In our simulations, for example, the number of clusters screened at the 1% threshold is an order of magnitude higher than at the 10% surveillance threshold and the number of false positive predictions potentially prompting further investigation is also manifold greater.

Our top predictors of viral emergence require a comprehensive sampling of the viral and host population. Although exact measurements of these quantities are practically infeasible, our results suggest that targeting molecular surveillance towards precise and accurate estimation of viral growth rates, both for newly emerging clusters and the resident circulating viruses, may enhance flu prediction. One approach is to target the two key components of growth rate separately—mutational load and effective susceptibility. Changes in the mutational load can be estimated from sequence data, comparing the number of differences that occur in non-epitope portions of the genome over time [65, 95, 112]. Our parameterization of R_c follows the empirical method of [95], with fitness costs based on nonsynonymous amino acid differences between a given strain and its most recent common ancestor. Estimating the effective susceptibility is more challenging, as it depends on the interaction between an individual's exposure history [50, 91, 125] and new amino acid substitutions in epitode coding regions [18, 66]. Nonetheless, several studies introduce innovative methods for estimating susceptibility from the historic distribution of flu subtypes, seasonal flu prevalence, and HI-titers. For example, Neher et al. [112] predict antigenic properties of novel clades by mapping both serological and sequence data to a phylogenetic tree structure of HA sequences. Łuksza & Lässig estimate effective susceptibility by first estimating the historic frequency of clades in six-month intervals and then estimating cross-immunity between those clades and the focal cluster based on amino acid differences in epitope regions [95]. However, both methods only consider clusters that have already surpassed 10% relative frequency, at which point strains are thought to be geographically well-mixed and less prone to geographic sampling bias.

Another approach to estimating the growth rate of an emerging cluster is to treat it as a composite quantity. We evaluated several proxy measures of cluster growth rate, including the relative fold-change in frequency between two time points. Models based on fold change rather than the true growth rate actually have greater sensitivity, that is, they are more likely to detect clusters destined for dominance when they first emerge. However, the positive predictive values of our best models drop from from 0.81 to 0.67, meaning that replacing the true growth rates with an approximation increases the rate of false alarms. Importantly, the proxy model improves with the addition of a second predictor, the variance in fold-change across the viral population, which can also be readily estimated from surveillance data. Thus, variance in fitness appears to be a robust secondary predictor of future sweeps, regardless of how fitness is quantified. Surprisingly, a model based on seemingly naive approximations of growth rate — the time elapsed between two frequency thresholds and the number of other co-circulating clusters rising in frequency — was even more accurate, though still inferior to the true growth rate models. We did not evaluate a promising alternative strategy for approximating fitness, based on the phylogenetic reconstruction of currently circulating sequences [39, 114]. Unlike the proxies we considered, this does not require historical data but does rely on pathogen sequencing. Finally, although not as informative as predictors that quantify the evolutionary and immunological state of the population, easily quantifiable predictors such as the total number of infected individuals or the frequency of the circulating dominant cluster, can be incorporated into future predictive models.

Our attempts to directly apply the optimized models to empirical data were of limited success. While the global evolutionary dynamics of influenza A/H3N2 clusters visually resemble those observed in our simulations, the sparse genotypic data available do not permit estimation of the phenotypic predictors identified in our study. The number of epitope mutations in a newly emerging cluster relative to co-circulating viruses provides early indication of future success. This provides proof of concept that the evolutionary viability of influenza viruses is predictable, but will require better models for estimating viral fitness from sequence data and the expansion of surveillance efforts [33] to collect phenotypic data reflecting the mutational loads viruses and dynamic trends in population susceptibility.

While our study provides actionable suggestions for improving both the surveillance and forecasting of antigenic turnover, it is limited by several assumptions. One caveat of our method is that we do not capture the explicit phylogenetic structure of the influenza population. Therefore, we do not distinguish between clusters that are successful because of one mutation and clusters that are successful because of a series of mutations. If for instance, a novel antigenic mutation caused the emergence of a new cluster (phenotype) that circulated briefly before a second novel antigenic mutation caused a second phenotype that eventually achieved our defined criteria, we ignore the fact that the established cluster is a subclade of the first and that the antigenic mutation that conferred the first phenotype is fixed along with the second antigenic mutation [59, 143]. This scenario follows Koelle & Rasmussen's description of a two-step antigenic change molecular pathway that leads to antigenic cluster transitions [82]. Our analysis is therefore relevant for scenarios that depict their described jackpot strategy – a combination of one large antigenic mutation occuring on a low deleterious background. Second, the simulation represents global H3N2 dynamics and ignores differences in temperate and tropical transmission dynamics [1, 144, 156]. Prior studies have revealed considerable global variation in transmission rates, which should positively correlate with the frequency of cluster transitions. Furthermore, viruses that emerge in tropical regions are more likely to be the source of viruses that eventually circulate in temperate regions [10, 89]. Temperate regions produce more extreme seasonal bottlenecks, potentially leading to greater stochasticity in viral dynamics, which makes it more difficult for novel strains that emerge in temperate regions to spread globally [164]. We also do not consider selective pressures imposed by seasonal vaccination. Its impact on antigenic turnover depends on vaccination rates and the immunological match between the vaccine and all co-circulating viruses. Seasonal vaccination could differentially modify the effective susceptibility of clusters, suppressing some while creating competitive vacuums for others. Theoretical study suggests that antigenic drift should slow down [7, 155] and the circulation of co-dominant clusters may become more common [85]. Given these caveats, we emphasize our qualitative rather than the quantitative results. Our study highlights promising predictors of viral success, characterizes robust trade-offs between the timing, costs and accuracy of such predictions, and serves as proof-of-concept that model-derived surveillance strategies can accelerate and improve forecasts of antigenic sweeps. If we fit similar models directly to historic surveillance data, the resulting predictions will likely reflect greater uncertainty but perhaps naturally reflect global variation in influenza dynamics and vaccination pressures.

Our study demonstrates that the early detection of emerging flu viruses is limited by a tight race between the typical dynamics of antigenic turnover and the annual timeline for influenza vaccine development. It also provides a foundation for analyzing the costs and benefits of expanding surveillance capacities and shortening the vaccine production pipeline. As we strive to expedite and improve molecular surveillance for vaccine strain selection, even incremental progress is valuable. Earlier detection of antigenic sweeps, regardless of vaccine efficacy, can inform better predictions of severity, public health messaging regarding personal protective measures, and clinical preparedness for seasonal influenza.

Chapter 4

Within-host HIV-1 biological processes reflected in deconstructed Ancestral Recombination Graphs

4.1 Abstract

Pathogen phylogenies are used to infer transmission dynamics of HIV-1 among hosts. However, inference of within-host evolutionary dynamics has proved challenging because common phylogenetic reconstruction methods and tools assume that there is no recombination, an important feature in HIV-1's complex biology. Recombination contributes significantly to HIV-1 within-host diversity, has been suggested to accelerate HIV-1 adaptation, and facilitates the persistence of HIV-1 lineages that display a signal of latency. To address this gap, we developed a new within-host coalescent model of HIV-1 based on a network approach, the Ancestral recombination Graph. Using our within-host HIV-1 ARG model to simulate evolutionary histories from a set of molecular sequence samples, we first investigated whether complex processes such as recombination, cycling in and out of a latent reservoir, and population demographic changes leave detectable signals in reconstructed phylogenies using traditional tools. Second, we implemented patient-specific sampling schemes to generate plausible evolutionary histories of two specific individuals, and

used a sampling estimation framework that used topological and distance-based tree statistics for comparison between simulated and observed trees to identify differences in recombination rate and latent reservoir size between the two. Overall, our within-host ARG model showed agreement with HIV-1 trees when the recombination rate is at the current biological estimate, and when the latent reservoir is at the low end of current estimates. We also found that as the recombination rate increases, the latent reservoir size has to increase to generate simulated trees that align with observed trees. Finally we find the possibility that the latent reservoir size differs between individual patients. This study represents an important step in adding realism to HIV-1 within-host evolutionary modeling and is the first study in developing a rigorous inference method based on approximate Bayesian computation (ABC) that will jointly estimate recombination and latent reservoir parameters for 11 patients.

4.2 Introduction

HIV-1 molecular sequence data, collected as part of routine clinical care, has recently been shown to capture epidemiological dynamics [42, 75, 79] at both the within-host and population levels. Phylogenetic inference has been used to elucidate the epidemiological relationships between transmission pairs [133], the direction of transmission, and the diversity of the founding population [134]. At the population level, studies in HIV phylodynamics, the coupling of phylogenetic analysis with epidemiological models, have estimated stage-specific HIV transmission rates, the estimated time of cross-species spillover into humans [169], and HIV-1 prevalence in at risk populations [128]. Despite the advances in inferring epidemiological dynamics, applications to within-host HIV-1 evolutionary dynamics have not kept pace. Within a host, HIV-1 evolves rapidly due to interplay of the selective pressure exerted by the host's immune response and a combination of HIV-1's error prone reverse transcription process, short generation time, and pervasive recombination. The frequency of recombination in HIV has serious clinical and epidemiological consequences, including driving drug resistance, immunological escape, and disease progression. Despite its importance as a key feature of HIV-1 evolution, its inclusion in modeling studies remains limited due to the fact that almost all of the tools and models developed to study molecular evolution assume that there is no recombination.

A recombinant is a genetic sequence that contains regions from genetically distinct parental strains. A recombinant genotype occurs when distinct parental strains coinfect a CD4 T+ cell, are encapsulated into a heterozygous virion, the virion infects a new cell, and the reverse transcriptase enzyme synthesizing new retroviral DNA switches between the distinct RNA templates. Both participant [113] and simulation-based studies [9] estimate that the effective recombination rate, which incorporates both the probability of coinfection of a single host cell [76] and the template switching rate, is on the order of 1.4×10^{-5} to 1.38×10^{-4} per base per generation, comparable to the estimated point mutation rate of $2.2 - 5.4 \times 10^{-5}$ per base per generation [54, 97].

Within a host, recombination provides HIV-1 with a means to increase the genetic variation for selection to operate on [37, 46, 48, 101, 107], can lead to rapid emergence of antiviral drug resistance [63, 78, 105], and can help shed deleterious mutations. However recombination can also slow the rate of adaptation by separating beneficial alleles in the same genome. Theoretical studies have shown that the magnitude of benefits brought by recombination depends on the interaction between factors such as population size and epistatic interactions [83, 101, 103]. In the case of HIV-1, a study of mother-child HIV-1 transmission demonstrated that recombination beneficially contributes to the early diversification of HIV and elevates the effective evolutionary rate [139].

The most common methods for reconstructing the evolutionary history from genetic sequence data assume a dichotomous tree such that each virus is the descendant of one ancestral virus. However, recombination, especially on the pervasive scale of HIV-1, violates this assumption. Under the assumption of recombination, a single sampled genome can have different evolutionary and genealogically relationships depending on what part of the genome you use to reconstruct those relationships. Therefore, representing the evolutionary history of a set of HIV samples with a single genealogy may lead to erroneous inference in both the estimation of parameters and inference of relationships [6, 127]. This source of error can be factored out of analysis by using recombination detection tools [98, 99, 126, 151] to identify and exclude recombinants from phylogenetic analysis. Yet, the rate at which recombination occurs in HIV suggests that most lineages have a recombination event in their recent past and are comprised of multiple ancestral lineages. One way to incorporate recombination into phyogenetic methods is to use a more general class of networks to model the evolutionary history of a sampled set of individuals. Griffiths and Majoram proposed such a method in their 1996 paper based on Ancestral recombination graphs (ARGs) [62]. ARGs allow for the complete record of coalescent and recombination events of a set of aligned DNA sequences. However, they have not yet been widely integrated into population genetics because of the computational intensity of inferring an ARG from more than a handful of sequences. Furthermore, unlike standard phylogenetic reconstruction, we have less intuition for how to interpret the structure of an ARG to derive meaningful conclusions about the evolutionary history of a sample.

Further complicating the development of phylogenetic inference methods for within-host HIV-1 evolution that take into account recombination is the presence of the latent reservoir, transcriptionally inactive HIV-1 provirus integrated into the genome of resting memory CD4+ T cells. The latent reservoir may establish as early as the first week of infection[34, 166], will persist through antiretrovrial therpy (ART)[136], and can be reactivated through exposure to recall antigen or various cytokines, or the cessation of ART [135]. Once ancestral sequences are reactivated, recombination facilitates the persistence of archival sequences [71], HIV sequences that display less divergence from the transmitting virus than contemporary sequences, further increasing the diversity of the population that selection can act upon. Therefore, a key barrier to functionally curing HIV is the presence of the latent reservoir.

”Shock and kill” methods that hope to cure HIV through first shocking CD4+ T cells into activation and then using a second agent to kill the infected cells and block new infection cannot be effectively evaluated without an estimate of the reservoir size. Estimates of the size of the latent reservoir are difficult to measure, and have ranged from 1 in every 10^6 resting CD4+ T cells [35, 47, 168] to $\sim$ 60-fold higher [69].

HIV-1 biology and evolutionary dynamics are vastly more complex than standard models permit, including recombination, latency, and population demography. The purpose of this study is to determine if it possible to detect the signal of that complex biology in standard binary tree form. We incorporated these within-host biological processes into a modified coalescent framework to generate an ARG. We investigated the effect of including these processes on a reconstructed binary tree using topological and distance summary statistics. Finally, we sought to describe qualitative differences between two participant phylogeny’s from Shankarappa et. al [145] by using a matching algorithm to search combinations of parameters that produce similar trees. We find that our biologically realistic model is able to reproduce the qualitative and quantitative characteristics of HIV-1 evolution as measured in a binary tree and that our trees are comparable to empirical patients when using the biological estimation of recombination and smaller latent reservoir sizes.

4.3 Methods

4.3.1 Methodological overview

Our approach is based on a modified coalescent theory method that simulates an ARG conditioned on a defined sampling scheme, i.e. number of samples taken at designated time points since infection. A forward-time simulation of the entire viral population's evolutionary history would be computationally costly as an infected host can generate up to 10^{11} viral particles per day, whereas the coalescent approach efficiently simulates the history of a sample of lineages. The simulated ARG thus represents the entire evolutionary history of our sample, where branch lengths correspond to the time between coalescent and recombination events. We decompose the ARG into a series of binary trees based on the genealogies of specific residues, taking into account recombination breakpoints and periods of latency. Using a sample of the residue genealogies, we calculate an average distance matrix between extant tips and construct an average tree based on a hierarchical clustering algorithm. Finally, we compare the average distance matrix and tree to empirical data to identify plausible parameter values and combinations for evolutionary processes.

4.3.2 Model Assumptions

We make several assumptions about the within-host evolutionary dynamics of HIV-1. First, we assume that an individual is infected with a single transmitted HIV-1 variant. From that lineage, the HIV-1 population diverges and diversifies linearly with time, with intermittent demographic bottlenecks caused by the host immune response. In addition, we assume that there is a reservoir latent population that is established three weeks after infection and remains constant in size for the duration of HIV-1 that we are simulating. The lineages present in the reservoir population turnover during the course of the infection. For all processes, we assume neutrality, specifically that any lineage is equally likely to coalesce, recombine, or go into or out of the latent reservoir. Lastly, we assume that the evolutionary rate remains constant over the course of the infection.

4.3.3 Simulation of the ARG

Our simulation models four possible events in reverse time: (1) a recombination event between two active lineages, (2) a virus entering the latent reservoir, (3) a virus in the latent reservoir reactivating, and (4) a coalescent event between two active lineages. We assume that waiting times to each event are independent and conditional only on the extant number of lineages in either an active or latent state and the time since transmission. To move forward in the simulation along the reverse time axis, we draw random waiting times to each of the four possible events according to the equations described below and proceed with the event of the minimum waiting time.

Event	Reaction	Parameter Value	Reference
Coalescence	$2k_A \to k_A$	$\alpha = 0; \beta = 1$	[132]
Recombination	$k_A \to 2k_A$	$\rho \in (0, 0.001, 0.01, 0.05, 0.1)$	[9, 74, 113, 171]
Latency/Deposition	$k_A \to k_L$	$\lambda = [0.5 - 5.25]$*	[148]
Activation	$k_L \to k_A$	$\lambda = [0.5 - 5.25]$*	[148]

Table 4.1: **Within-host HIV evolutionary events in the ARG.** k_A and k_L represent the number of active and latent lineages present. λ is the global flow rate of a lineage into and out of the latent reservoir. σ is the per genome per day rate of recombination. *This range of λ corresponds to the values of eqn. 4.4 using values of $N_L \in (10^1, 10^3)$.

4.3.3.1 Evolutionary Events

- **Coalescence:** When a coalescent event occurs, two extant active lineages are chosen randomly and joined to form a parent lineage. The branch length between the chosen and parent lineages is the time since each extant lineage's last event in the ARG and the new coalescent event. The expected waiting time for two random lineages to coalesce is dependent on the effective population size and the number of active lineages. During periods of linear growth, the effective population size is modeled as $\mathrm{N(t)} = \alpha + \beta \mathrm{t}$, where α is the number of transmitted lineages, β is the growth rate, and t, is the time since infection in days. For all simulations we assume that $\alpha = 0$ and that $\beta = 1$. During bottleneck events, the population remains at a constant size, B_{strength}, for a period of 5 days.

  To simulate the time to the next coalescent event, we use two functions corresponding to the two regimes of population demography. During periods of linear growth, we use the inverse cumulative function derived in [132]:

$$F_{C(l)}^{-1}(u) = \left(1 - (1-u)^{\frac{\beta}{\binom{k_A}{2}}}\right)(\alpha + \beta t_1)\beta^{-1} \tag{4.1}$$

where u is a unit uniform random variate.

During periods of bottlenecks that occur every $B_{\text{frequency}}$ days, we use the Kingman n-coalescent model:

$$F_{C(b)}(k_A) \sim \text{Exp}\left(\frac{k_A(k_A-1)}{2N}\right) \tag{4.2}$$

In this way, we approximate the effects of directional selection without explicitly incorporating the reproduction probability of individual lineages. Following the period of decreased population size, the population resumes linear growth, maintaining the overall gradual linear increase in genetic diversity over the course of the infection.

- **Recombination:** A recombination event adds a lineage to the total number of extant active lineages. When a recombination event occurs, one lineage is chosen from the set of extant active lineages and two lineages are added to the ARG, representing the two parental lineages in forward time that each contributed a portion of its genome. Each recombination event has a randomly generated specific breakpoint at one of 700 sites, corresponding to the *env* gene we are simulating. The branch lengths between the recombinant lineage and each parent lineage is the time since the recombinant lineage's last event in the ARG and the new recombination event.

We assume recombination to be a homogeneous process where recombination events occur at rate ρ per lineage, and thus draw the expected time to the next recombination event from:

$$F_R(\rho, k_A) \sim \text{Exp}\big(k_A \rho\big) \tag{4.3}$$

- **Latent Reservoir Deposition and Reactivation:** When a deposition event occurs we randomly select one of k_A lineages, calculate its branch length as the time from its last event in the ARG until the time of the deposition event, and put the lineage into a latent state. While a lineage is in a latent state, which we refer to as *latency*, it does not experience mutational processes. When an activation event occurs, we randomly select one of the k_L lineages in the latent reservoir, calculate its branch length from its time of deposition to the time of the next activation event, and put the lineage into an active state. On the whole, through deposition and activation events, the total number of lineages in the population, k, remains constant, with k_L and k_A increasing and decreasing by one respectively.

  We use a single parameter, λ, as the global rate per day for entering and leaving the latent population. The value of λ is calculated by:

$$\lambda = N_L \frac{log(2)}{44 * 30}, \tag{4.4}$$

where 44 corresponds to the estimated half life in months of the latent reservoir population [148] and N_L is the total reservoir population size that contains replication competent provirus. We use the inverse cumulative function,

$$F_L^{-1}(u) = \left(1 - (1-u)^{\frac{\beta}{k_A \lambda}}\right)(\alpha + \beta t_1)\beta^{-1} \tag{4.5}$$

and the exponential distribution,

$$F_A(\lambda, k_L) \sim \text{Exp}\left(\frac{k_L}{N_L}\lambda\right) \tag{4.6}$$

to govern the dynamics in and out of the latent reservoir population. Equation 4.5 is structurally similar to Equation 4.1 because the population size of active lineages k_a is changing over time. From the moment of transmission, the HIV-1 population size is growing. During the early stage of infection when the population is still small, there is a greater probability that a lineage entering the latent reservoir will be in the evolutionary history of our sample. Conversely, as the population size grows, the probability that the next lineage to enter latency is in the evolutionary history of our sample decreases. Equation 4.6 governs the joint probability that a lineage in the latent population will reactivate in reverse time and that the lineage will be in our sample.

Longitudinal Sampling To simulate longitudinal sampling we designate times along the reverse time-axis when new active lineages are added to the simulation. At each additional sampling event, the branch lengths of any remaining active and latent lineages from the previous sample are extended in time up to the next sampling time. From there the simulation proceeds as before, with k and k_A updated to reflect the additional new samples. The simulation ends after the last sampling time (first in forward-time) when $k_A = 1$ before $t = 0$.

4.3.4 Obtaining a Distance Matrix from the ARG

Decomposing the ARG into a series of binary trees To use distance and topological statistics that are applicable to dichotomous trees, we decompose the ARG into a series of binary trees that each represent the genealogies of individual residues. First, for residue i in the genome, we remove recombination events by selecting one path through the ARG according to the individual residue's history. Moving along the reverse time axis, for a recombination event and a given residue, i, if residue position i is smaller than the recombination break point, we select the path of the most recent parent in the ARG. Conversely, if residue position i is greater than the recombination break point, we select the path of the more ancestral parent in the ARG. We remove the branch in the ARG between the recombination event and the non-selected parent and create a new branch between the descendant of the recombination event and the selected parent. The new branch length equals the length of time between the descendant node and the recombination event plus the length of time between the recombination event and the next event of its parent node. Second, for each residue binary tree, we adjust the branch lengths to account for periods of latency. We remove the inner events that represent the start and end of a lineage's time in the latent reservoir recalculate the branch length as only the time spent in an active state. Because of computational limitations, we complete this decomposition process for a random sample of 25 residues.

Calculating a distance matrix Once each ARG has been decomposed into a series of i binary trees that consistent only of coalescent events with branch lengths reflecting time spent in an active state, we calculate an average distance matrix. We first use the igraph package [38] in R to turn the edge list of each residue i binary tree into a distance matrix, where entries record the pairwise distance in time between two tips. Second, we average the pairwise distances between all combinations of tips over all i residue distance matrices. Finally we use a minimum evolution principle [41] to generate a hierarchical clustering representation of the average distance matrix. To visualize a hierarchical clustered tree, we root it at tMRCA of the samples from the first (in forward-time) sampling event and order the internal structure of the tree to highlight tree imbalance.

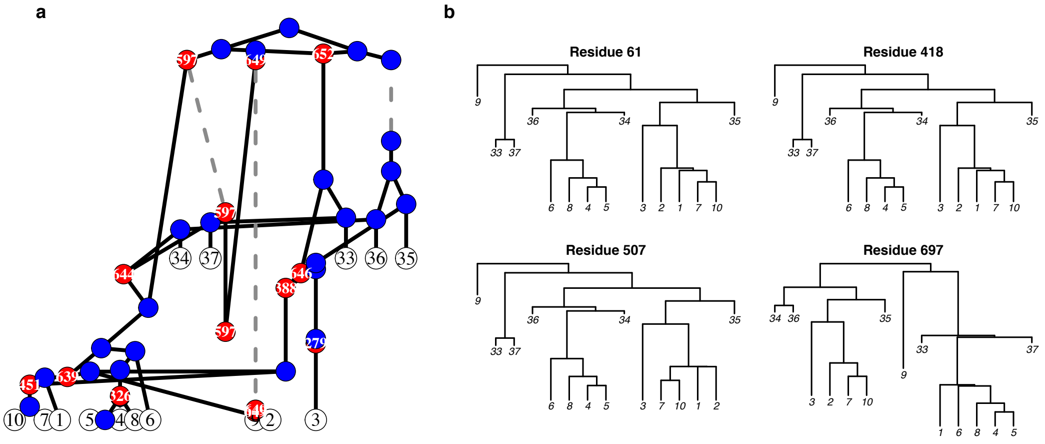

Figure 4.1: **Ancestral Recombination Graph (ARG) decomposition into a series of binary trees.** (a) The simulated ARG from two sampling events. Red dots represent recombination events, with the break point identified in the white text. Blue dots represent coalescent events between two lineages. Dashed grey lines represent when a lineage is in the latent reservoir. The branch lengths correspond to time. This ARG was simulated using $\rho = 0.08$, $N_L = 760$, and no bottlenecks. 5 samples were taken 3.5 months post transmission and 10 samples were taken 7 months post transmission. (b) Each panel represents the true genealogy of a specific site in the genome from (a). Despite being almost 400 residues apart, residue 61 and residue 418 share the same genealogy. Residue 507 and residue 697 are unique.

4.3.5 Data

We applied this framework to the analysis of longitudinal HIV-1 DNA sequences sets from 2 empirical participants [145]. In the study, sequences corresponding to the HIV-1 *env* gene were taken from each participant over a course of 6 to 12 years starting at 3 months from the time of seroconversion. There was an average of 11.875 individual sampling events with an average of 9.83 (s.d. 1.66) samples taken per event.

For our study, we aligned the sequences of each participant using MAFFT v7.305b2 and generated a tree again using the minimum evolution principle [41]. To translate the sequences into distance matrices we used the TN93 model[157]. The overall distribution of distances were similar using other evolutionary models. We obtained the sampling scheme from each participant, i.e. the number of samples (clones) taken at how many months post transmission. Thus for each participant, we have a distance matrix, a consensus tree, and a sampling scheme.

4.3.6 Distance Function and Parameter Inference

For five strata of recombination rates, we sampled from three proposed distributions for the frequencies of demographic bottlenecks, $\mathrm{U}(14, 365)$, the bottleneck strength $\mathrm{U}(1, 100)$, and the latent reservoir size $\mathrm{U}(1, 10000)$. Thus, for each θ_i, representing the four parameters of the simulation, we simulated and analyzed the ARG. We chose to stratify the recombination rate into five fixed rates instead of a continuous distribution because of computational efficiency. Likewise, we sampled fewer particles at the higher recombination rates because of computational cost. We sampled on the order of 10,000 particles each for $\rho = \{1e^{-12}, 0.001, 0.01\}$; on the order of 1,000 particles for $\rho = 0.05$ and on the order of 100 for $\rho = 0.1$. We do not believe sparse sampling at high recombination rates significantly impacts our results because we saw less variation in simulation outcome and summary scores at higher recombination values.

We compared the fit of an ARG simulation to an empirical participant through a combined distance score of five equally weighted features, including three distance statistics and two topological statistics (Table 4.2).

For the Sackin index, MLT, and the EI Ratio, the score of a statistic, s_i was the squared difference between the raw statistic and the empirical statistic. The CV score was calculated as the sum of the absolute differences in CV value at each sampling time. While all features may be sensitive to the sampling scheme, the ranked distance matrix, is calculated by making one-to-one comparison between the empirical and simulated distance matrices. We calculate the score of the ranked distribution of pairwise distances as follows: $y_{i,j,k,n}$ is the observed genetic distance between the jth sequence in the ith sample time and the nth sample in the kth sample time, and $s_{i,j,k,n}$ is the simulation time distance between the jth sequence in the ith sample time and the nth sample in the kth sample time. (1) Sort the distances for all time points pairs i, k ($y_{i,.,k,.}$ and $s_{i,.,k,.}$) (2) Calculate $f(y, s) = \sum^{i,k}(y_{i,j,k,n} - y_{i,j,k,n} \times \tau)^2$. (3) Minimize $f(y, s)$ over $\tau < 0$. τ is a scaling factor which serves to scale the simulated distance matrix, measured in time, to the participant distance matrix, measured in genetic distance.

To get the final ranking of trees across all parameter combinations for a participant, we normalized each individual statistic across all simulations and added the normalized components for a total score. Thus, for a simulation with parameter set Θ_i, giving a vector of statistics $s_1, ..., s_5$ the distance score would be $d = s_1 + ... + s_5$ where s_i is the normalized statistic. Lower values of the distance score d indicate closer matches to the empirical tree.

We calculated the Sackin Index normalized by the Yule Model [138] using the CollessLike package in R. A higher Sackin Index indicates a less symmetrical and more ladder-like tree. We calculated the number of lineages through time by dividing the number of monophyletic groups by the number of extant tips at each sampling event [72]. We used the R package ape (Analysis of Phylogenetics and Evolution) [124] to create and plot all binary trees.

Feature	Feature Type	Description
Sackin Index	Topological	The average number of splits from a tip to the root of a tree and captures asymmetry over the evolutionary history of the sample
Mean Lineages Through Time (MLT) [72]	Topological	The diversification of lineages normalized by the number of extant tips
External:Internal Branch Ratio (EI Ratio)	Distance	The ratio between the mean external branch length (branch that ends with a sampling event) and the mean internal branch length (branch between coalescence events)
Ranked Distance Matrix	Distance	The difference in the empirical and simulated distributions of branch lengths ranked by size within each sampling event
Coefficient of Variation in Pairwise Distance (CV)	Distance	The ratio between the standard deviation of pairwise distances between branches of the same sampling event and the mean pairwise distance between branches of the same sampling event

Table 4.2: **Features used to match the simulated evolutionary history to a participant's empirical HIV-1 *env* phylogeny.** Some features measure tree characteristics (Topological) while others measure distance statistics (Distance). A simulation's score, d, is the sum of the normalized differences between the feature measured on reconstructed simulated tree and the empirical tree.

4.4 Results

4.4.1 Additional biological complexity results in trees consistent with HIV-1 phylogenies

The within-host ARG simulation model that includes population demography, a latent reservoir, and recombination produces trees consistent with HIV-1 phylogenies over a 12 year period of within-host evolution (Fig. 4.2). We illustrate an example of adding each of these features to the within-host model to visualize their effects on the reconstructed viral phylogeny and build up to our most complex biological model in Fig. 4.2h. For comparison, Panels (i-l) are four participant HIV-1 phylogenies from the Shankarappa et al. study [145]. The empirical trees reveal several defining characteristics of within-host HIV-1 trees. Generally, they have a strong ladder-like structure, meaning the tree is asymmetrical, the terminal branch lengths are longer than the internal branches, and tips from the same sampling event cluster together. However, this last feature, which we will refer to as chronological grouping, is not strictly adhered to. In Participants 7, 1, and 11 there are instances where tips from later sampling events are clustered with tips from previous sampling events.

Each panel in (a-h) shows a reconstructed phylogeny from one ARG simulation using the same sampling scheme but realized under different levels of biological complexity. First, we consider the effect of each process in isolation. In (a), we see an example of a tree generated using only a linear-growth coalescent model that accounts for linear growth in genetic diversity from the time of transmission. This condition represents the baseline upon

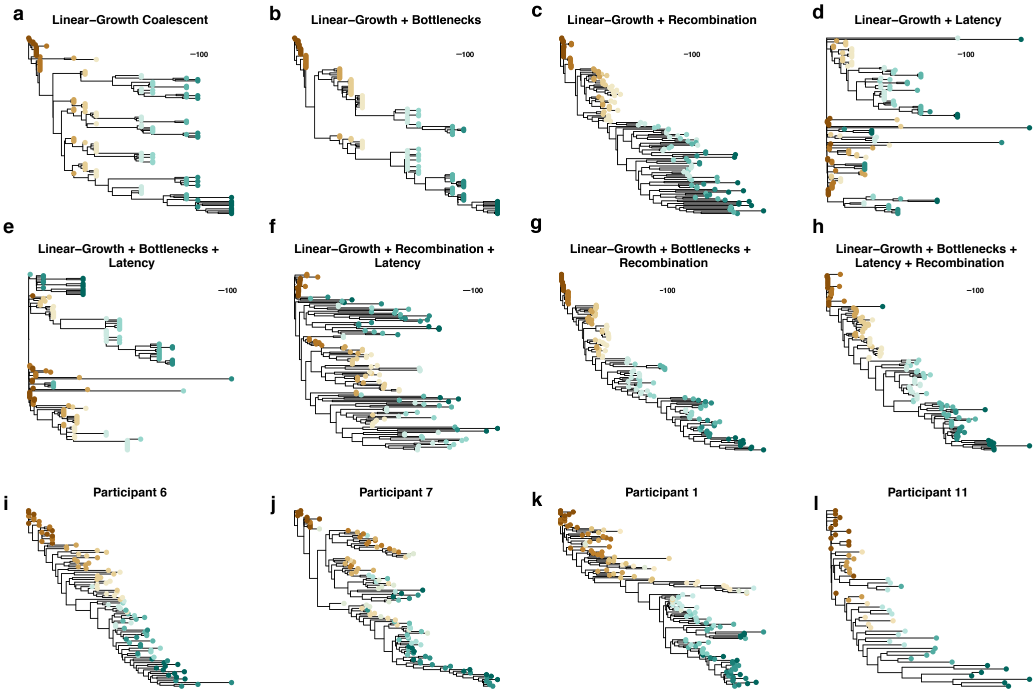

Figure 4.2: **Relationship among within-host evolutionary processes and virus phylogenies.** Each tree in (a-h) is the decomposed average tree from one ARG simulation of 12 years using 132 sampled lineages spread over 10 sampling events. Color represents tips sampled at the same time. The baseline model, a linear-growth coalescent model is in (A). Parameters are consistent across panels (a-h). For simulations in which the process is active: $\mathrm{B}_{\text{frequency}} = 300$ and $\mathrm{B}_{\text{strength}} = 15$; $N_L = 2500$; and $\rho = 0.01$. Panels (i-l) are reconstructed HIV-1 trees from participants in [145].

which we advance the biological realism of the coalescent model. In (a) terminal branches are elongated due to the underlying growing population. Samples from later time points form polyphyletic groupings and the tree is moderately symmetrical. In (b) we introduce population demography with intermittent bottlenecks, emulating the impact of the host immune response. The effective population size is reduced during bottlenecks, causing lineages to quickly coalesce according to standard coalescent theory that the rate of coalescence will be proportional to the inverse of the population size. This process results in a stronger ladder-like structure that is consistent with the signature of positive selection. However, periods of rapid coalescence cause shorter terminal branches. In (c) we add recombination to the linear-growth coalescent model. Recombination increases the ladder-like structure and chronological grouping over that of (a). However, the length of the terminal branch lengths results in an extreme external to internal branch ratio. Adding the presence of a latent reservoir (d) permits lineages to spend time in a dormant stage where they do not experience outside evolutionary pressures. Including this feature has the most visually disruptive effect. The ladder-like structure is lost, terminal branches of later time points extend back to the root, and the chronological structure is lost.

Panels (e-g) show the effects of adding features in tandem to the linear-growth coalescent model. The presence of a latent reservoir in (e) and (f) results in a flatter tree despite either intermittent bottlenecks or recombination. The tree in (g), the product of recombination and intermittent bottlenecks, begins to look similar to empirical trees, in particular (i). For trees where all tips chronologically group together by sampling event, these two features are sufficient to reproduce the qualitative characteristics of an HIV-1 phylogeny. However, if there is any signal of latency, such as in panels (k-l), then the model must also include a latent reservoir. In panel (h) we show our most biologically complete model. Fig. 4.2h shows a clear picture of how all three processes work together – recombination maintains the ladder-like structure which intermittent bottlenecks further enhance, recombination and latency cause variation among and lengthening of the terminal branches, and latency allows for deviations from strict chronological groupings.

4.4.2 Evidence of biological processes in topological and distance statistics

To establish that recombination, a latent reservoir, and population demography leave detectable quantitative signals on reconstructed phylogenies, we measured how two topological and two distance statistics change with varying intensities of each process.

Both the frequency and strength of bottlenecks have significant impacts on topological and distance statistics (Fig. C.1). As bottlenecks become more frequent and reduce the effective population size to a smaller number, trees become more imbalanced, fewer lineages survive between sampling events, branch lengths between lineages sampled during the same event become more varied, and the EI ratio decreases. These results are consistent with theoretical and

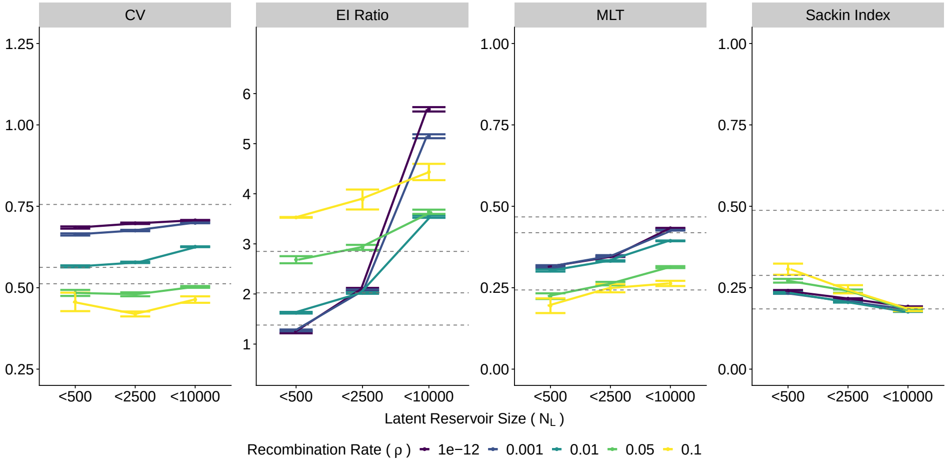

Figure 4.3: **The magnitude of each biological process causes variability in distance and topological statistics.** Each dot represents the mean and standard error of the mean of the statistic. Colors represent different rates of recombination, ρ. Simulations are grouped into three levels of reservoir size of increasing magnitude, with the breakpoints depicted on the x-axis. This figure shows all simulations where $B_{\text{frequency}} < 200$ days, and $B_{\text{strength}} < 75$. Results over all bottleneck combinations of strength and frequency are similar. The number within each category are provided in Table C.1. The grey dashed lines indicate the maximum, median, and minimum value for the 9 empirical participants in [145]. On the y-axis we show the limits of the MLT and the Sackin index to highlight that both the empirical and simulation values are a subset of the possible range.

experimental population genetics research that show how bottlenecks reduce population genetic variation.

4.4.2.1 Recombination countervails the effects of latency

We can see the varying impact of the latent reservoir size, N_L, and recombination rate, ρ, across all four statistics in Fig. 4.3. Increasing N_L has a monotonic effect on each statistic for $\rho \leq 0.01$. As N_L increases, trees have more variable distance between tips of the same sampling event, greater EI ratios, a higher number of lineages surviving through time (MLT), and are more symmetrical. These effects are most strongly seen when the recombination rate is lower, and may not be observed at higher rates of ρ because of smaller sample sizes. Kruskal-Wallis rank sum tests between the mean statistic at each level of N_L are highly significant for all values of ρ except for $\rho = 0.1$ ($p < 2.2e^{-16}$ for all Kruskal-Wallis tests). At $\rho = 0.01$, the mean Sackin index at $N_L < 500$ and $N_L < 10000$ is significantly different (Dunn's Test, Z = -2.63, Holm's Adjusted p = 0.01) and the mean CV is significantly different between $N_L < 2500$ and $N_L < 10000$ (Dunn's test, Z=2.57, Holm's Adjusted p=0.012). Higher MLTs and lower Sackin Indices imply more balanced trees. This is because with a non-negligible latent reservoir, lineages that are deposited soon after transmission are reactivated later in the infection and will be genetically more similar to lineages from earlier sampling events than contemporary lineages. This deviation from a strong chronological structure creates more variation in the branch lengths of any given sampling event.

If one of the ancestral lineages of a recombinant genome was ever in the latent reservoir, recombination can remove the latency signal in a single residue genealogy if the residue descended from the non-latent parent. Thus recombination offsets the effects of latency on all four statistics. For example when recombination is absent ($\rho = 1e^{-12}$) and N_L is large, the tree is more comb-like with short internal branches and long external branches, and high levels of symmetry. As ρ increases, reactivated ancestral lineages are increasingly forming recombinant genomes with contemporaneous lineages, leading to stronger chronological structure and a more homogeneous distribution of distances between extant tips of the same sampling event as indicated by lower CV means. In the same way, the MLT decreases with larger values of ρ. For all statistics, Kruskal-Wallis rank sum tests between values of ρ are significant at all values of N_L ($p < 2.2e^{-16}$ for all Kruskal-Wallis tests see Table C.3). Narrowing our focusing on the pairwise differences between $\rho = 0.01$ and $\rho = 1$, as an example comparison between a mid-range biological estimate and an extreme value of ρ, there are significant differences for all statistics at $N_L < 2500$, and at $N_L < 10000$ with the exception of the Sackin index (Table C.4 Dunn's test, Holm's adjustment). There are no significant differences at $N_L < 500$, most likely due to small sample sizes (Table Dunn's test, Holm's adjustment). The EI ratio shows an interaction dependency between N_L and ρ. At small values of N_L, higher values of ρ have an elongating effect on external branches. The reverse is true at higher values of N_L, where extreme EI ratios are from simulations with $\rho = 1e^{-12}$ and $\rho = 0.001$. This is because the processes of recombination and latency each individually lead to longer external branches. However, the effect of latency is stronger on its own than recombination. Thus when recombination is included, it has the ability to remove some of the latency signal.

Figure 4.3 shows that a range of recombination, latency, and bottleneck intensities may be consistent with within-host HIV-1 evolutionary dynamics. However, while we represent the mean statistic in Fig. 4.3 to describe quantitative trends in topological and distance metrics as function of intensity, there is considerable variation in the statistical values. This variability might come from integrating over the variability in strength and frequency of the bottlenecks and latent reservoir size. Indeed, combinations with a mean statistic outside the empirical range can occasionally produce trees with statistics that fall within the empirical range. Furthermore, different statistics give conflicting support for which parameter values of recombination and latent reservoir size are most consistent with HIV-1. Specifically, the MLT and EI Ratio give inconsistent conclusions for the likely end of the latent reservoir population spectrum; the MLT statistic approaches the empirical median at higher latent reservoir sizes, while the EI Ratio far overshoots the empirical range at higher reservoir sizes. To account for such dissonant information, we use a distance algorithm that looks at an equally weighted combined measure of all statistics to gauge the fit of a simulation to an empirical tree.

4.4.3 Application to sequence data from HIV-1 infected individuals

4.4.3.1 Distance algorithm recovers empirical tree features

We implemented the specific sampling schemes of Participants 6 and 7 and simulated possible ARG evolutionary histories over 30,000 parameter sets of Θ. We choose Participants 6 and 7 because despite similar sampling schemes, their HIV-1 trees are dissimilar in the strength of the chronological grouping and symmetry of the trees. Using our distance algorithm to rank decomposed trees derived from the ARG simulations, we identified trees that are close fits to each participant.

Figure 4.4 shows the simulated distributions of all statistics, the empirical range from the 9 participants in the [145] study, and the specific statistic value of Participants 6 and 7. Kolmogorov-Smirnov two-sided tests indicate that the distributions of the statistical values are different for every statistic. (CV: D = 0.5, $p < 2.2e^{-16}$; EI Ratio: D = 0.10, $p < 2.2e^{-16}$; MLT: D = 0.14, $p < 2.2e^{-16}$; Sackin: D = 0.06, $p < 2.2e^{-16}$. Wilcoxon rank sum tests indicate that the distribution means are significantly different for all statistics except for the CV (CV: W = 4.78×10^8, $p = 0.57$; EI Ratio: W = 5.41×10^8, $p < 2.2e^{-16}$; MLT: W = 5.70×10^9, $p < 2.2e^{-16}$; Sackin: W = 5.165×10^9, $p < 2.2e^{-16}$). These results hold when we tested differences between a smaller sample that included 10% of simulations at each value of ρ. The topological and distance statistics of the simulated trees span the empirical ranges and with the exception of the Sackin index produce more extreme values at either end. For the Sackin index and the MLT, the empirical measurements fell at

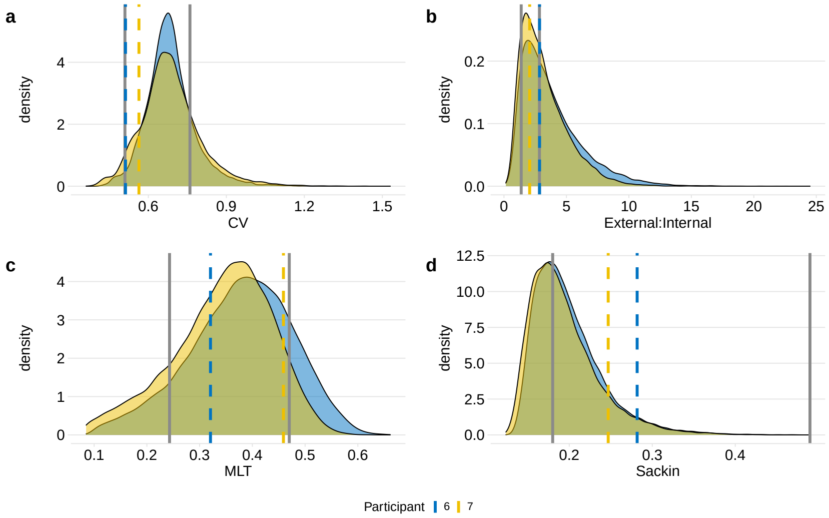

Figure 4.4: **Density distributions of topological and distance statistics derived from the complete collection of the ARG-simulated reconstructed phylogenies for Participant 6 and Participant 7.** Each panel shows the kernel density estimate of the statistic calculated from the reconstructed tree or distance matrix over 30,918 simulations for Participant 6 and 30,853 simulations for Participant 7. The blue and yellow lines show the empirical value for each participant. The grey lines show the maximum and minimum values from the 9 participants in [145].

or above the 90% percentile of all simulations (Table 4.3). We suggest hypotheses for why our simulations generally produce more balanced trees in the discussion.

We illustrate the performance of our algorithm for matching ARG-simulations to empirical data in Figure 4.5. Our algorithm was able to recover the distinguishing features between Participants 6 and 7 in the best-matching trees. The empirical tree for Participant 6 is conventional (Fig. 4.5a) and the best-fit simulation for Participant 6 has similar characteristics, specifically a

Participant		Sackin %	MLT %	EI Ratio%	CV%
6	Complete	95	27	43	22
	Best 5%	77	67	92	14
7	Complete	90	91	32	11
	Best 5%	86	96	65	43

Table 4.3: **Participant percentile of the ARG-simulated reconstructed phylogenies.** *Complete* corresponds to the $\sim 30,000$ simulations for each participant. *Best 5%* corresponds to the distribution of only the ARG simulations that produced the lowest 5% of distance scores.

long EI ratio, a clear ladder-like structure, and consistent chronological structure. Conversely, the empirical tree for Participant 7 is noisy (Fig. 4.5b). For later sampling events, there are three distinct clusters that comprise tips from multiple time points. Fig. 4.5b is an example of a tree from the lowest 1% of distance scores for Participant 7, where $d = 0$ represents the best score. It has multiple surviving clusters, heterogeneous branch lengths between tips of the same-time point, and is more symmetrical than Participant 6. Trees that return higher distance scores are dissimilar to the participant trees. Panels (c-d) show that trees at the 25th percentile of scores display deviation from the empirical trees.

4.4.3.2 Distance scores are low when N_L is small

The distribution of normalized distance scores differs between the two participants. The median score for Participant 6 is 0.3379 (IQR: 0.2407-0.4395) and for Participant 7 is 0.1697 (IQR: 0.1159-0.2494) (Fig. SC.2). The distribution of scores for Participant 7 is more skewed to the right than for Participant

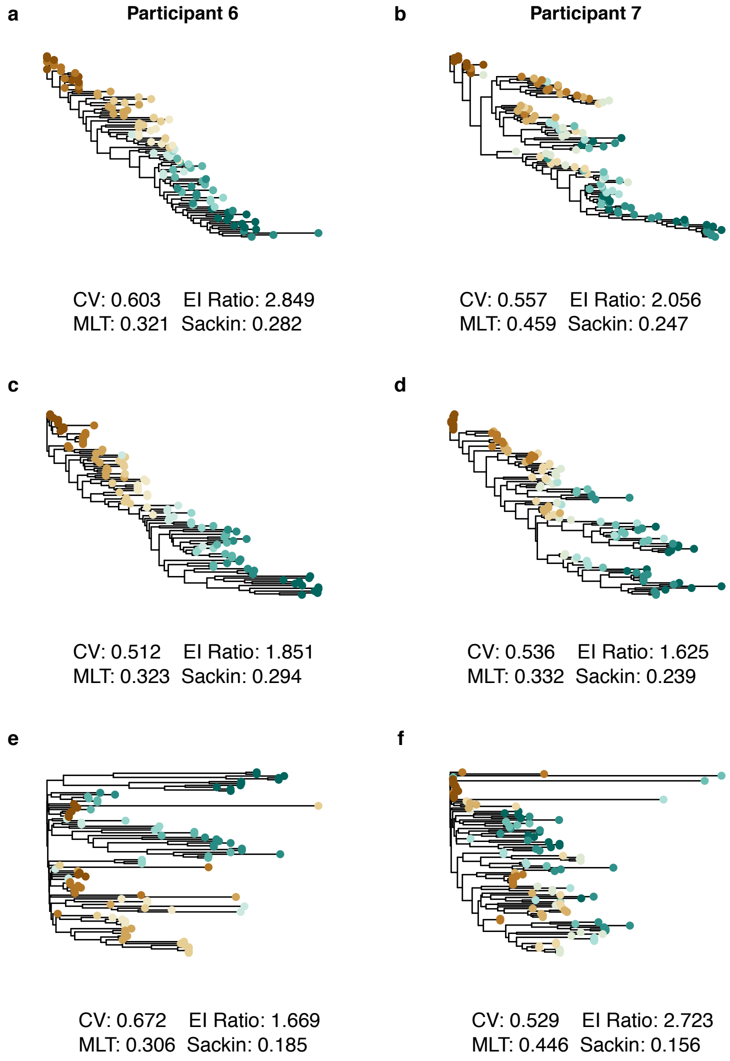

Figure 4.5: **Distance algorithm discriminates between good-fit and bad-fit trees.** The top row shows the empirical trees for Participant 6 and Participant 7 with their topological and distance statistics. The middle row shows the top 99th percentile of distance scores (the lowest 1% of scores). The bottom row shows trees at the 75th percentile of distance scores. Parameters for (c) $B_{\text{strength}} = 48.13$, $B_{\text{frequency}} = 207.87$, $\rho = 0.01$, $N_L = 357$; Parameters for (d) $B_{\text{strength}} = 95.28$, $B_{\text{frequency}} = 171.36$, $\rho = 0.01$, $N_L = 467$; Parameters for (e) $B_{\text{strength}} = 90.80$, $B_{\text{frequency}} = 29.53$, $\rho = 0.01$, $N_L = 6673$; Parameters for (f): $B_{\text{strength}} = 84.65$, $B_{\text{frequency}} = 294.54$, $\rho = 0.01$, $N_L = 4068$.

6. This is because for Participant 7 $\rho = 0.01$ simulations produce the lowest distance scores (Fig 4.6) while higher values of ρ produce lower distance scores for Participant 6 (see below for further discussion). Because of the computational cost of simulating and decomposing the ARG at $\rho \geq 0.01$, there are 10 and 100 fold fewer simulations at $\rho = 0.05$ and $\rho = 0.1$ respectively than $\rho = 0.01$.

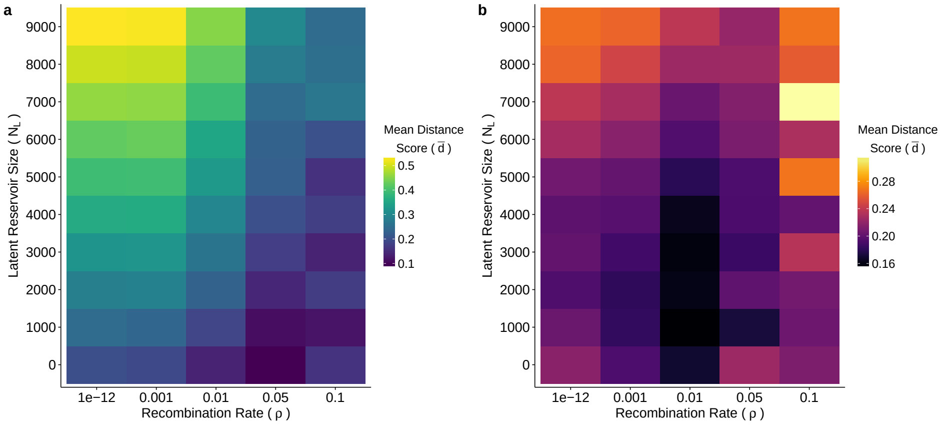

Figure 4.6: **Mean distance scores, $\bar{d}$, for Participant 6 (a) and Participant 7 (b) by recombination rate ρ and latent reservoir size N_L.** The color of each tile represents the mean normalized distance score for each combination of ρ and N_L. d is normalized separately for each participant. While we restrict our simulations to five discrete values of ρ, N_L is sampled from a continuous distribution. For this analysis, we bin N_L into increments of 1000. The number of individual simulations in each tile ranges from 3 when $\rho = 0.1$ to 1000 for $\rho \leq 0.01$. The mean distance score and standard deviation by recombination rate ρ are in Table S: C.2.

To evaluate which θ_i produced good fits for Participants 6 and 7 we measure the mean distance score for different combinations of recombination ρ and latent reservoir size N_L (Fig. 4.6).

Comparing across the entire range of simulations for Participants 6 and 7, several patterns emerge. For Participant 6, the mean distance score is closer to 0, indicating a better fit, when the recombination rate is high and the reservoir population is small. In addition, we see that as the recombination rate increases, higher reservoir populations produce lower distance scores compared to smaller reservoir population sizes. For example, for a $N_L = 6000$, at $\rho = 1e^{-12}$, $\bar{d} = 0.423$; at $\rho = 0.1$, $\bar{d} = 0.204$. This suggests that in Participant 6, there is an interaction effect between the recombination rate and the reservoir population size. The highest $\bar{d}$ are at low recombination rates with high latent reservoir sizes. Participant 7 has a more defined range of good-fit combinations. The recombination rate $\rho = 0.01$ consistently produces better fit trees for a range of N_L. In contrast to Participant 6, there is no relationship between higher rates of ρ, N_L, and the mean distance score.

These patterns become more extreme as we restrict our analysis to simulations in the best 5% of distance scores (Fig. 4.7). When considering these top simulations, there are significant differences in the latent reservoir size at different values of ρ (Kruskal-Wallis rank sum test, $\chi^2 = 62.14$, d.f. = 4, $p = 1.03e^{-12}$.) The lowest mean scores for Participant 6 are concentrated in regions with higher recombination rates and smaller reservoir sizes. We stress the qualitative patterns rather than the absolute differences in the mean distance score because of the few numbers of individual simulations in the $\rho =$ 0.1 column. In Participant 7, there is strong support for a maximum $\rho = 0.01$ with a range of possible latent reservoir sizes. Across recombination rates,

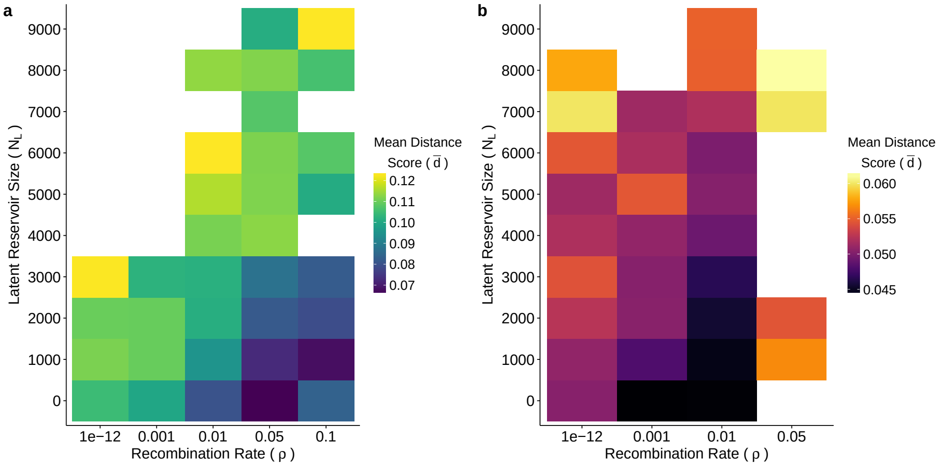

Figure 4.7: **Simulation distance scores for Participant 6 (a) and Participant 7 (b), restricted to the top 5% of simulations.** The number of individual simulations in each tile ranges from 1 to 100. Blank tiles indicate that no simulations with that combination of latent reservoir size and recombination rate are in the top 5% of simulations. The mean distance score and standard deviation by recombination rate ρ are in Table C.2.

the latent reservoir distribution is significantly different between $\rho = 0.001$ and $\rho = 0.01$ (Dunn's Test, Z = -2.80, Holm's adjusted $p = 0.0153$). Both participants show less support for extreme latent reservoir sizes close to 10000.

For both participants, we do not find a strong effect of the bottleneck frequency and strength on the distance score (Figs. C.5, C.6). The top 5% of simulations for Participant 6 are nearly uniformly sampled across the range of both frequency and strength and did not have a strong correlation with the final distance score (B_{strength}: Pearson's r = -0.04, p = 0.099, d.f. = 1544; $B_{\text{frequency}}$: Pearson's r = -0.015, p = 0.554, d.f. = 1544). Participant 7 showed a slight negative correlation between the strength of the bottleneck and the distance score (Pearson's r = -0.048, p = 0.06, d.f. = 1541,) but no significant correlation with bottleneck frequency (Pearson's r = -0.002, p = 0.931, d.f = 1541).

Our results thus far demonstrate that Participants 6 and 7 are the most sensitive to the recombination rate and the latent reservoir size. Although the top 5% of simulations for Participant 6 favored recombination rates $\rho \geq 0.05$, simulations with $\rho = 0.01$ are included, which Participant 7 heavily favored. Furthermore, $\rho = 0.01$ is a recombination rate within the range of current estimates. Given the overlap of the model parameters (Fig C.5, we sought to parsimoniously attribute the differences between Participants 6 and 7 to only the latent reservoir size. To test this hypothesis, we set equal prior distributions on all parameters except for the latent reservoir size. Because Patient 7 showed some sensitivity to the bottleneck specifications (Fig. C.6),

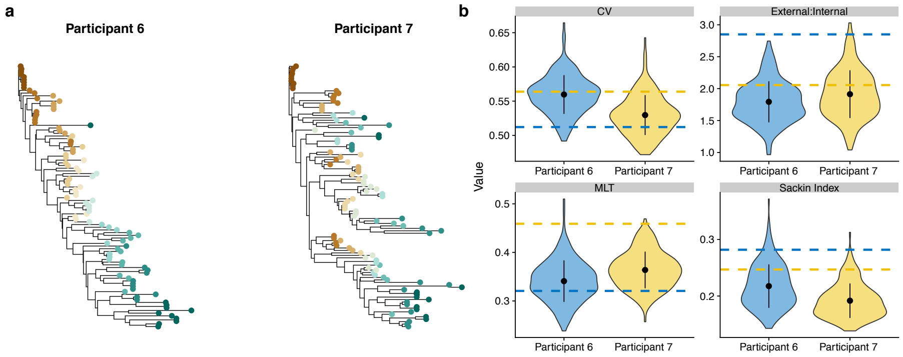

Figure 4.8: **Size of the latent reservoir accounts for qualitative but not quantitative differences between Participant 6 and Participant 7.** (a) Example reconstructed trees from ARG simulations based on prior distributions where only N_L varied between participants. (b) Violin plots show the kernel probability distribution of the topological and distance statistics from the distribution of reconstructed trees. Dashed lines show the empirical statistic, with the color corresponding to the Participant (blue - Participant 6; yellow - Participant 7).

we took the mean (μ) and standard deviation (σ) of the bottleneck strength ($\mu = 68.65, \sigma = 21.52$) and bottleneck frequency ($\mu = 237.92, \sigma = 79.16$) for the simulations in the top 2.5% where $\rho = 0.01$. Likewise, for each participant we found the mean and standard deviation of the latent reservoir size when $\rho = 0.01$ from the top 2.5% of simulations (P6 $\mu = 816$, $\sigma = 759$; P7 $\mu = 2190$, $\sigma = 1675$). We set a prior on each parameter θ such that $\theta \sim N(\mu_\theta, \frac{\sigma_\theta}{3})$ for all parameters except for the recombination rate, which we set to $\rho = 0.01$. We generated 200 ARG simulations under these prior distributions, where the only difference between Participant 6 and Participant 7 was the prior placed on the latent reservoir size. Through this experiment, we are able to produce the qualitative but not quantitative differences between Participants 6 and

7. When we looked at the reconstructed trees from these distributions, we found that in general Participant 6 trees showed higher tree-like structure and fewer lineages through time. However, the relative ranking in the CV and EI ratio was backwards: trees from Participant 7 had on average higher EI ratio, when the empirical measures indicate Participant 6 has a higher EI Ratio. Reconstructed trees for Participant 7 also had lower variation between tips of the same sampling event, when in reality Participant 6 has a lower CV.

4.5 Discussion

The importance of recombination and the latent reservoir in within-host HIV-1 evolutionary dynamics has been described in detail since the mid 1990's [47, 135]. Although these two processes have been modeled individually and together using forward mathematical models and simulations [43, 68, 71, 108], both processes have been largely ignored in phylodynamic investigations because of the theoretical and computational problems in jointly inferring phylogeny and recombination patterns. As efforts continue towards finding a functional cure of HIV-1 that involves eliminating or reducing the size of the latent reservoir, studying the interaction between the latent reservoir and recombination remains key. In this study we extend a previously developed within-host coalescent model [132] into an ARG simulation model that allows for lineages to coalesce, recombine, and cycle in and out of a latent state. We find that these complex evolutionary dynamics leave tractable signals in binary trees reconstructed using hierarchical clustering and can be quantified in readily measurable topological and distance features. Additionally, we use a sampling estimation framework to evaluate the intensity of recombination and the latent reservoir size in HIV-1 patient trees.

Our within-host HIV-1 ARG simulation model and decomposition framework produces trees that qualitatively and quantitatively capture HIV-1 evolutionary dynamics. The topological and distance-based statistics from reconstructed trees are within the range of empirical HIV-1 phylogenies, and our distance matching algorithm shows the ability to recover distinguishing tree and topological characteristics when applied to patients with visually distinct evolutionary histories. We find that in some cases, there is a dependency between the intensity of recombination and the size of the latent reservoir. In these cases, faster recombination rates are associated with larger latent reservoirs. Because these are compensatory features, parameter identifiability is a problem as multiple combinations of the two might be observationally equivalent. Moving forward, experimental studies might help define an informative prior distribution for one or both of these parameters. In Participant 7 we see an instance of how strong evidence for one parameter can reduce the inference problem to one dimension. Furthermore, a rigorous Approximate Bayesian Computation (ABC) inference framework could better test between alternate hypotheses and indeed our future directions include implementing one for more participants.

The analysis of the two participants from the Shankarappa study demonstrates the above-mentioned difficulties in parameter identification. Even so, we can derive conclusions and generate hypotheses about the relative strength and consistency of the recombination rate and size of the latent reservoir. We find consistency between the recombination rates that produce the best-fitting trees to the empirical data and the current biological estimates of the effective recombination rate. The estimated effective recombination rate, 1.4×10^{-5} to 1.38×10^{-4} per base per generation, corresponds to a rate of $0.008 - 0.081$ recombination events per lineage per day in our simulation of a 700bp genomic fragment. For Participant 7, reconstructed trees from simulations with $\rho = 0.01$ have on average the lowest distance scores. For Participant 6, the lowest distance scores correspond to simulations where $\rho = 0.05$ or $\rho = 0.1$, but favor 0.05 when considering only the top fitting simulations (Table C.2). Trees with negligible recombination are ill-fits for both participants, emphasizing the need to develop phylodynamic methods that account for recombination.

In terms of the latent reservoir size, both participants show strong support for latent reservoirs that are on the order of 10^2-10^3, although simulations with values up 10^4 occasionally produce good fits. Large latent reservoirs closer to 10^4 produce trees with extreme external to internal branch ratios, at times almost 9-fold larger than the maximum empirical measurement (Fig. 4.4). A 10^2 latent reservoir corresponds to the estimate of 1 per 10^6 million CD4+ T cells. While our results suggest the latent reservoir is near the smaller end of the estimated spectrum and might vary between individuals, there remain unanswered questions relating the absolute size of the reservoir with how much virus it produces and quantifying the adaptive impacts it has on generating genetic diversity available for immune escape.

The third biological feature we include in our within-host ARG simulation model is population demography in the form of bottlenecks. We find that bottlenecks are important for increasing the asymmetry of an HIV-1 tree, an indicator of directional selection driven by immune escape, and producing realistic topological and distance metrics (Fig. C.1). However, we do not find strong evidence for any one type of bottleneck event (Fig. C.5). This is unsurprising given that we model bottlenecks as a crude approximation of the selection pressure coming from the immune system. A lack of fitness differences between lineages might explain the tendency of our model to produce more balanced trees than real HIV-1 trees. In between bottlenecks, the coalescence rates of lineages are neutral. In reality, low-level positive selection from the immune system will create fitness differences between lineages, particularly for reactivated latent lineages that would be less fit because of long-term immunological memory[22, 130, 163]. These fitness differences in turn will increase the ladder-like structure of the tree. In addition to creating more balanced trees, this neutrality likely causes the variability in possible evolutionary histories for a given Θ, particularly when recombination is low. Lastly, this model artifact might be creating tension in the distance scoring. Because the Sackin index of simulated trees is often far below the empirical value, when the Sackin index is close, it has a disproportionate weight on the overall scoring. To address these shortcomings, future work will consider alternative formulations of changes in demography.

We show that when we simulate different patients under the same parameter assumptions except for the average latent reservoir size we can recover qualitative differences but not the expected ranking differences in topological and distance statistics (Fig. 4.8). One potential reason is that rather than a single cause that distinguishes the differences because individuals, it might be the complex dynamic interaction of multiple biological processes. Other possible explanations include a missing biological component, such as an explicit model of selection, or a poorly defined prior of the parameters.

In our distance algorithm we introduce the ranked distance matrix as a new statistical probe, which has two notable features. First, because the ARG models the time between events, the distance matrix obtained from a ARG-decomposed binary tree accounts for the true evolutionary distance in time between the extant tips. Second, distance matrix does not depend on any topological structure and thus the assumptions of phylogenetic reconstruction that omit recombination. However, a potential source of error comes from matching the time scale of the ARG to the genetic distance of the sequence data. In our formulation, an optimized scaling factor transforms the time distance to match the genetic distance and is independent of any specific genome-region. However, the scaling factor assumes a constant rate over the course of an individual's infection and does not account for the possibility that multiple mutations have occurred at an individual site. Thus a branch in the genetic distance matrix would appear shorter than if the branch was represented in time.

The calculation of a ranked distance matrix score depends on a one-to-one match with the empirical data and therefore we generate participant-specific ARG simulations. Despite considering Θs from the same prior distributions, we find significant differences between the topological and distance features of reconstructed phylogenies (Fig. 4.4). This suggests that the sampling scheme constrains the evolutionary histories in an observable way and is an important feature to consider in phylodynamic inference and can impact parameter estimation.

While our model advances the biological realism of within-host HIV-1 evolution, it could be further modified to include population structure. Population structure might be an important feature for reducing the range of evolutionary histories for a given set of parameter by limiting the lineages that could coalesce and recombine to lineages in a shared subpopulation. In effect, this would increase the number of lineages that survive through time and maintain the ladder-like structure of simultaneously evolving lineages as seen in Participant 7.

Appendices

Appendix A

Assessing real-time Zika risk in the United States

A.1 Extended Methods

A.1.1 Fitting the Generation Time

To capture the correct outbreak timing, we fit the generation time of our SEIR model to estimates for the ZIKV exposure and infectious periods in humans. The generation time measures the average duration from initial symptom onset to the subsequent exposure of a secondary case, and is estimated to range from 10 to 23 days for ZIKV [96]. In our model, the generation time corresponds to the sum of the exposure period and 1/2 the infectious period. We therefore fit the infectious period in our model to human ZIKV estimates for duration of viral shedding, and then fit the exposure period so that the sum of the two classes match the estimated ZIKV serial interval.

According to our modeling framework: with one infectious compartment, the distribution of waiting times in the compartment would follow a geometric distribution, with the most common waiting time equal to one day regardless of the transition rate. As this is a biologically unrealistic waiting time distribution, we use Boxcar implementations to yield a more realistic distribution [92]. In such a framework one splits a compartment into multiple separate compartments (boxes), has individuals transition through these compartments, and alters the transition rate for each compartment so the average waiting time spent in all compartments equals that of the original desired average. For example, if a 10 day infectious period were desired, one could model the infectious period as 1 compartment with a daily transition rate of 1/10, or 5 compartments with a daily transition rate of 5/10. The number of infectious individuals is either the number of individuals in the single compartment, or the total number of individuals in all five boxes. Both scenarios would have an average waiting time of 10 days to move through the infectious period, but the 5 boxes would necessitate individuals being infectious for at least 5 days giving a more realistic waiting time distribution that follows a negative binomial distribution (sum of multiple independent geometric distributions).

First, we solved for transition rates and compartments of a Boxcar Model infectious period that yielded an infectious period with 3 compartments and mean duration of 9.88 days and 95% CI of (3-22) [90]. Then, we fit the exposure period so that the combined duration of the infectious and exposure periods matched the empirical ZIKV generation time range [96], yielding 6 compartments and a mean exposure period of 10.4 days (95% CI 6-17) and finally a mean generation time of 15.3 days (95% CI 9.5-23.5). Given that the exposure period includes human and mosquito incubation periods and mosquito biting rates, this range is consistent with the estimated 5.9 day human ZIKV incubation period [90].

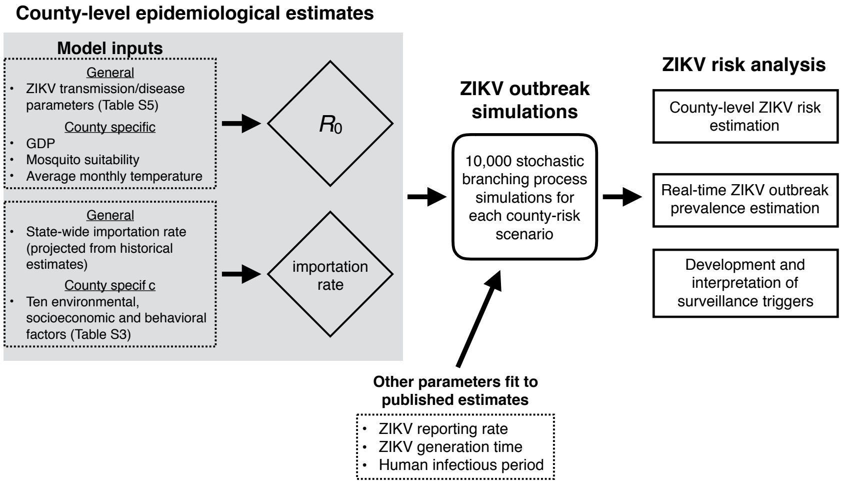
County-level epidemiological estimates
Model inputs
General
ZIKV transmission/disease parameters (Table S5)
County specific
GDP
Mosquito suitability
Average monthly temperature
R_0
General
State-wide importation rate (projected from historical estimates)
County specific
Ten environmental, socioeconomic and behavioral factors (Table S3)
importation rate
ZIKV outbreak simulations
10,000 stochastic branching process simulations for each county-risk scenario
Other parameters fit to published estimates
ZIKV reporting rate
ZIKV generation time
Human infectious period
ZIKV risk analysis
County-level ZIKV risk estimation
Real-time ZIKV outbreak prevalence estimation
Development and interpretation of surveillance triggers

Figure A.1: **ZIKV Risk Assessment Framework.** The method consists of three steps. First, we use data-derived models to estimate county-level ZIKV introduction rates and ZIKV transmission rates. Each estimate is based on a combination of general and county-specific factors. Second, for every county-risk combination, we simulate 10,000 ZIKV outbreaks using a stochastic branching process ZIKV transmission model parameterized by the county-level importation and transmission rate estimates along with several other recently published disease progression estimates. The simulations track the numbers of autochthonous and imported cases (unreported and reported) and, based on the total size and maximum daily prevalence, classifies each outbreak as self-limiting or epidemic. Third, we analyze the simulations to determine (1) robust relationships between the number of reported cases in a county and the current and future ZIKV threat and (2) surveillance triggers (number of reported cases) indicative of imminent epidemic expansion.

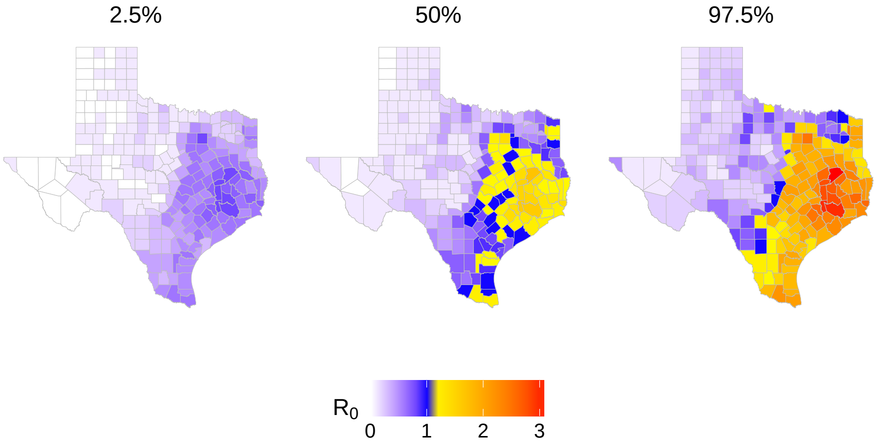

Figure A.2: **The 95% CI of R_0 Distributions for August.** From left to right, the 2.5%, 50% and 97.5% quantile R_0 values for August. The range of absolute values spans 0.02-6.90. Given the considerable uncertainty in socioeconomic and environmental drivers of ZIKV, we analyzed relative rather than absolute transmission risks.

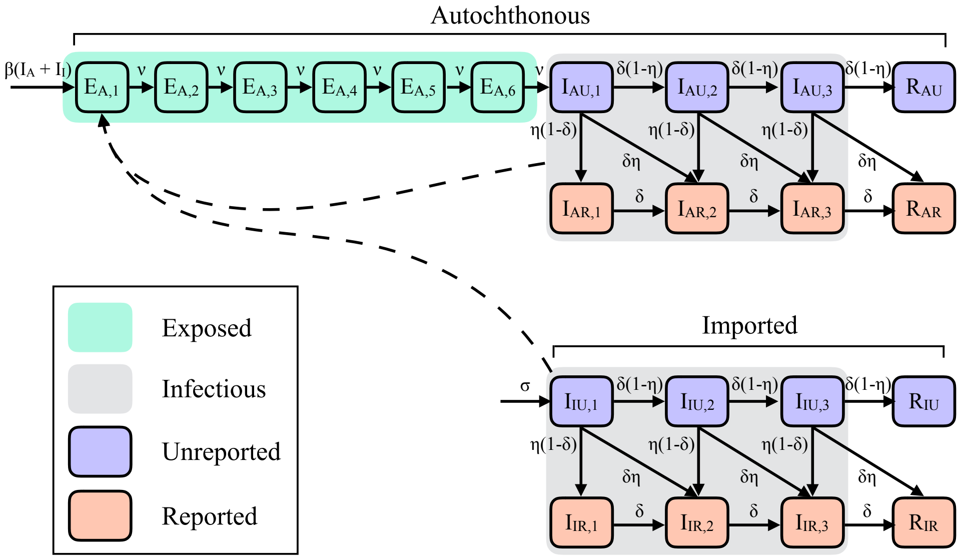

Figure A.3: **Diagram of ZIKV outbreak model.** The model tracks disease progression, transmission, and reporting of both imported and autochthonous ZIKV cases. Individuals progress through compartments via a daily Markovian process, according to the solid arrows in the diagram. The *Exposed* and *Infectious* periods consist of several (boxcar) compartments to simulate realistic outbreak timing. Unreported infected individuals have a daily probability of being reported. Imported cases are assumed to arrive daily according to a Poisson distribution (with mean σ) at the beginning of their infectious period, and otherwise follow the same infectious process as autochthonous cases. Autochthonous transmission occurs at rate $\beta(I_A + I_I)$, where I_A and I_I are the total number of infectious autochthonous and imported cases, respectively (dashed arrows).

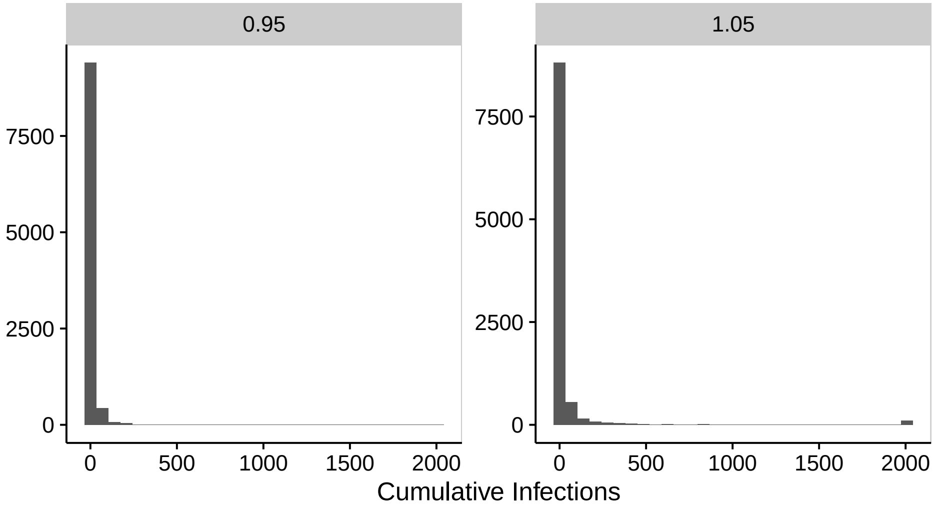

Figure A.4: **Determining outbreak simulation length.** If outbreak simulations are too short, self-limiting outbreaks may reach the maximum number of infections due to stochasticity. We chose to run our simulations to 2,000 cumulative infections as it conservatively differentiated the large outbreaks of simulations with R_0 just below 1 (0.95) from the epidemics of those with R_0 just above 1 (1.05). We therefore chose to run our simulations until a maximum number of 2,000 infections.

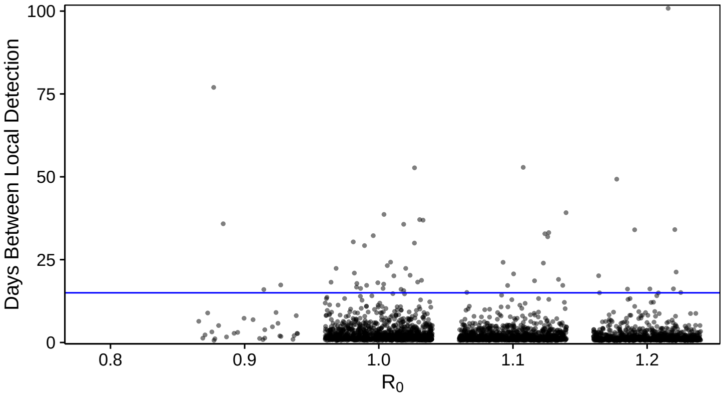

Figure A.5: **Time between detection of locally transmitted cases during epidemics.** Across a range of R_0 values with an importation rate 0.1 cases/day, we plot the time between detection events of autochthonous cases for simulations out of the 10,000 trials in which epidemics occurred (black dots). The blue line indicates a two-week threshold as recommended by the CDC for follow-up of local transmission. Even under a high importation rate of 0.1 cases/day, epidemics do not occur when $R_0 = 0.8$, and rarely occur when $R_0 = 0.9$. As R_0 increases, a greater proportion of simulations have fewer days in between detection events as the number of infections rapidly increase.

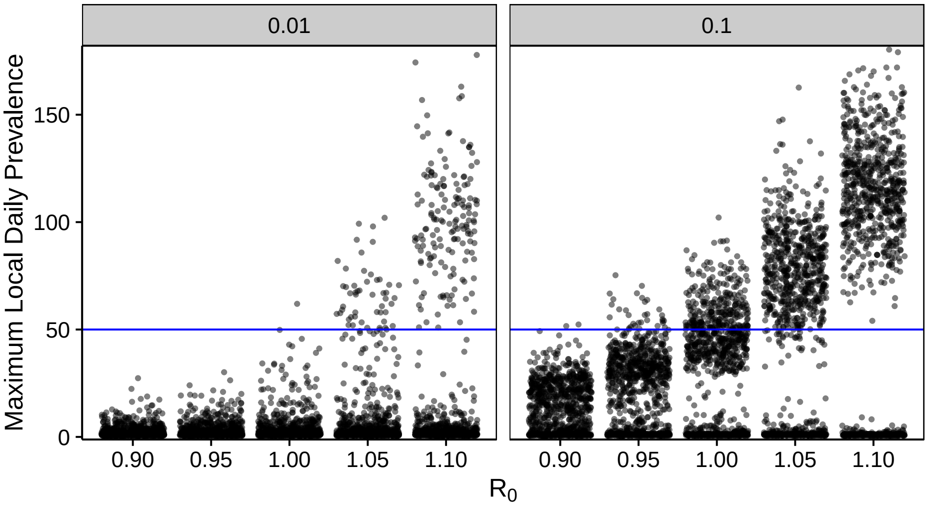

Figure A.6: **Selecting daily prevalence threshold for distinguishing self-limiting outbreaks from epidemics.** Across a range of R_0 values, we plot the maximum daily total autochthonous infectious individuals for 1,000 of our 10,000 trials (black dots). The blue line indicates the threshold (50) selected to differentiate epidemics with $R_0 > 1$ from outbreaks with $R_0 \leq 1$. At a low importation rate (0.01), the majority of simulations with $R_0 \leq 1$ are self-limiting and rarely progress into large sustained outbreaks. As R_0 increases, a greater proportion of simulations exceed the threshold. As the importation rate increases (panels from left to right) the separation between self-limiting outbreaks and epidemics becomes more pronounced.

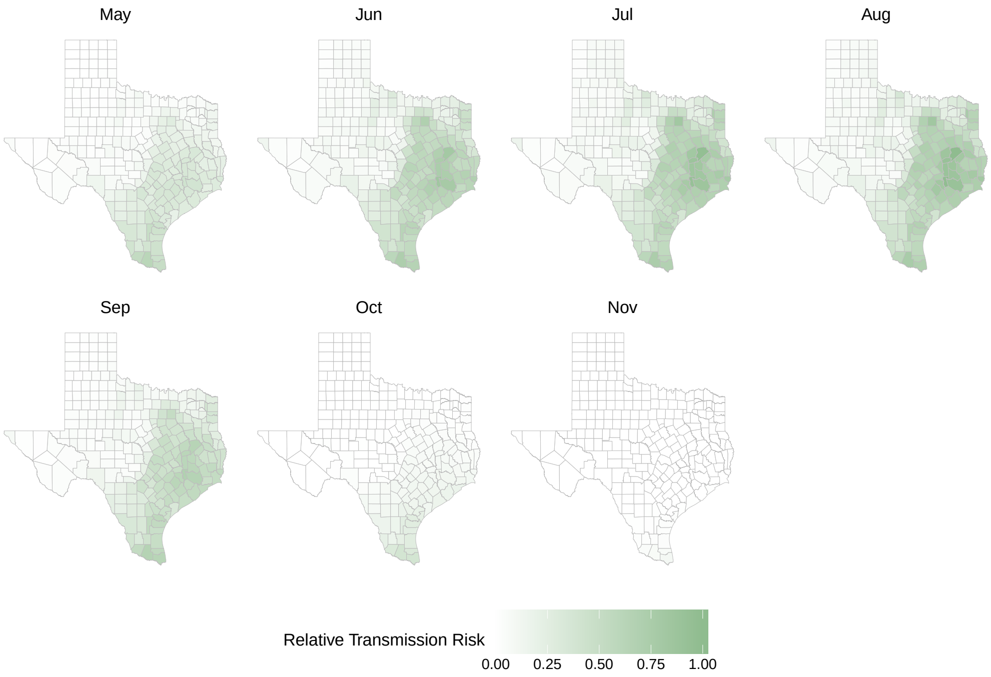

Figure A.7: **Monthly R_0 estimates based on seasonal changes in the temperature-dependent extrinsic incubation period of ZIKV in *Ae. aegypti* and the mosquito mortality rate of *Ae. aegypti***
.

Appendix B

Early Prediction of Antigenic Transitions for Influenza A H3N2

B.1 Extended Methods

B.1.1 Influenza Phylodynamic Simulations

Deriving the criteria for cluster establishment in the population

To separate the clusters that are eventually successful from those that only transiently circulate, we derived a two-criteria threshold of establishment based on reaching a minimum frequency in the population, and circulating above the minimum frequency for a specified duration of time. We choose the most stringent criteria that, when only accounting for clusters that reached the criteria, still maintained cyclical influenza dynamics. In addition, we compared the proportion of the total infected population attributable to established clusters under different criteria (Fig. B.1).

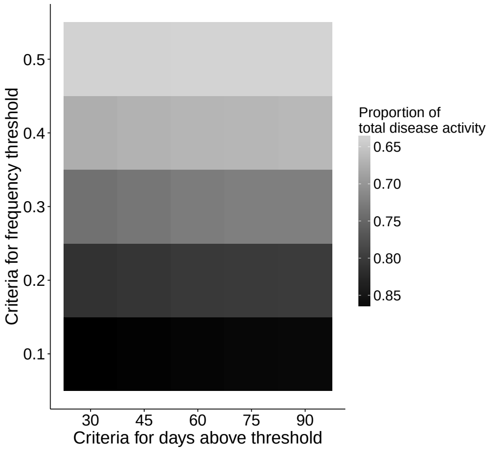

Figure B.1: **The proportion of total infections caused by established clusters is more sensitive to a frequency criteria than the duration of time**. Established antigenic clusters account for the majority of the disease activity at any point in time. In our analysis, established clusters are those that circulate above 20% relative frequency for at least 45 days. Using this criteria, infections caused by future successful clusters account for a median of 81% of the disease activity at any point in time.

B.1.2 Candidate Predictors

Candidate predictors are those population genetic and epidemiological indicators that were tracked during the phylodynamic simulations or calculated from the output. The full list is in Table B.1.

Candidate Predictor	Formula	Population	Cluster	Relative
Number of Infected Individuals	I	X		
Number of Uninfected Individuals	S	X		
Proportion of Individuals Infected	I/N	X		
Number of Circulating Antigenic Clusters	N_c	X		
Frequency of Current Dominant Cluster	$f_c = \max\left[I_c/I\right]$	X		
Entropy (Shannon's Diversity Index)	$H = \frac{1}{N_C}\sum_{j=1}^{N_C} f_j \ln\frac{1}{f_j}$	X		
Serial Interval of Infection*	$\mathrm{SI} = \frac{1}{I}\sum_{a,b\in\mathrm{infecteds}}\left(t_{a_0} - t_{b_0}\right)$	X		
The most recent common ancestor*	$\mathrm{TMRCA} = \max\left[\frac{1}{2}\left(\left(t_{TMRCA_0} - t_{a_0}\right) + \left(t_{TMRCA_0} - t_{b_0}\right)\right)\right]$	X		
Genetic Diversity*	$\omega = \frac{1}{I}\sum_{a,b\in\mathrm{infecteds}}\frac{1}{2}\left(\left(t_{TMRCA_0} - t_{a_0}\right) + \left(t_{TMRCA_0} - t_{b_0}\right)\right)$	X		
Antigenic Diversity*	$\lambda = \frac{1}{I}\sum_{a,b\in\mathrm{infecteds}}\lambda_{ab}$	X		
Deleterious Mutational Load	$\mathrm{k} = \frac{1}{I}\sum_{i=1}^{I} k(v_i)$	Mean, Var	Mean, Var	Mean, Var
Transmissibility	$\beta = \frac{1}{I}\sum_{i=1}^{I}\beta_0\left(1 - s_d\right)^{k(v_i)}$	Mean, Var	Mean, Var	Mean, Var
Effective Susceptibility*	$S_{\mathrm{eff}}(v) = \frac{S}{N}\sum_{h=1}^{N}\left(\sigma_{v(h)}\right)$	Mean, Var		
Covariance in transmissibility and effective susceptibility	$\mathrm{cov} = \frac{1}{I}\sum_{i=1}^{I}\left((\beta_i - \bar{\beta}) * (\sigma_v(i) - \bar{\sigma})\right)$	X		
Cluster Susceptibility*	$\sigma(v) = \sum_{h=1}^{N}\min(1, \sigma_{v,c(h,v)})$		Mean, Var	
Reproductive Growth Rate	$R(v) = \frac{\beta_0\left(1 - s_d\right)^{k(v)}}{\mu+\nu}\left(\frac{S_{\mathrm{eff}}(v)}{N}\right)$	Mean, Var	Mean, Var	Mean, Var

Table B.1: **Full set of candidate predictors considered**. Values were taken at the moment a focal antigenic cluster reached a specified surveillance threshold. The columns *Population, Cluster, Relative* indicate the scale and measure (e.g. mean and/or variance) that a predictor was considered in the model. Depending on the scale of the predictor, the *formula* could refer to all strains in the population, i.e. the strains of infected hosts, or the subset of strains in a specific cluster. *For computational simplicity, these quantities were calculated using strains from a random sample of 10,000 infected individuals. N = number of hosts; t_{a_0} = the time of birth of virus a; λ = antigenic distance between two strains. The antigenic distance is the pairwise degree of cross-immunity between two strains determined by the size of antigenic mutations and parent-offspring relationships; $k(v_i)$ = the number of deleterious mutations on a virus v of infected host i; s_d = the fitness effect of a deleterious mutation; σ_v = the average individual population susceptibility to cluster c; $\sigma_{v,c(h,v)}$ = the probability of infection of a host with historical infection i by a strain of cluster v

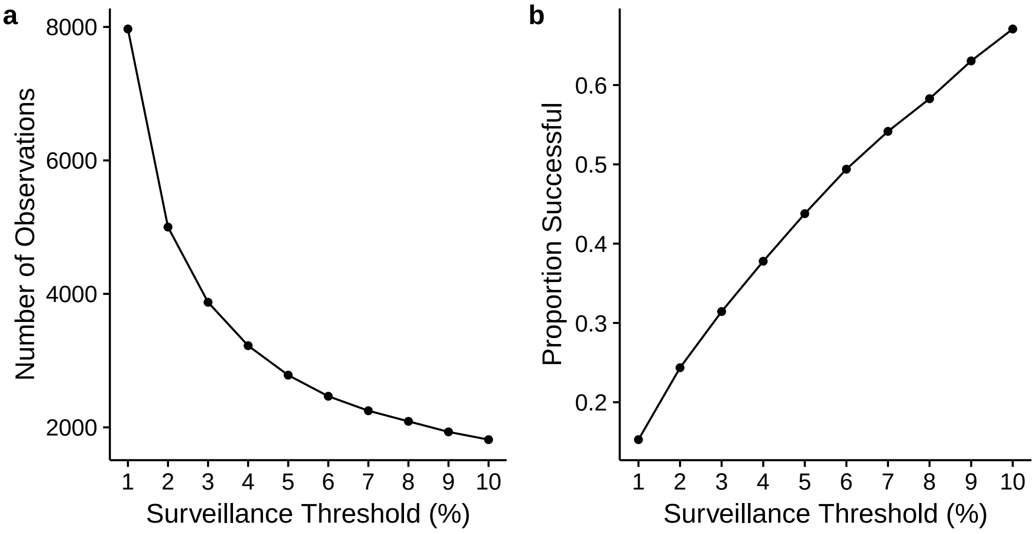

Figure B.2: **The fate of novel antigenic clusters** (A) Each point represents the number of antigenic clusters in our simulations that reach each increasing surveillance threshold (i.e., relative frequency in the population). As the surveillance threshold increases from 1 to 10%, the number of candidate clusters decreases from 7969 clusters at the 1% threshold to 1816 clusters at the 10% surveillance threshold. (B) Given a cluster has reached a surveillance threshold, the proportion of antigenic clusters that will establish (i.e. reaches $> 20\%$ for 45 days) increases with higher surveillance thresholds.

B.2 Extended Results

B.2.1 Incorporation of historical data

We tested an additional strategy for building classifier models: one that incorporated previously sampled data for each cluster.

To predict a cluster's evolutionary outcome for a specific surveillance threshold, we used data from three timepoints: 1) when it reached 1%, 2) when it reached the half-way frequency level between 1% and the surveillance threshold of interest, and 3) when it reached the surveillance threshold. The candidate predictors included all variables listed in Table B.1, as well as the difference in predictor values between 1% and the halfway point, and the halfway point and the maximum surveillance threshold. In addition, the difference of these differences was an additional predictor. Models were fit using the same approach described in the manuscript.

To compare the performance of models that incorporated data from past time points to models based only on current data, we compared the sensitivity and positive predictive value across surveillance thresholds from $1-10\%$. Both model types performed similarly across the range of surveillance thresholds (Fig. B.3). We focus on strategy that only incorporates current data because of the simplicity in the methodology and reduction of candidate predictors.

B.2.2 Surveillance Threshold Results

The best-fit logistic regression models for surveillance thresholds from $1 - 10\%$ are listed in Table B.2.

Surveillance Threshold (%)	**Predictor Variable**	**Coefficient Estimate**	**Std. Error**
1	$R_c/\langle R\rangle$	[2.61, 2.77]	[0.09,0.09]
	$\text{var}(R)$	[-0.54, -0.60)]	[0.06,0.06]
	$\langle R\rangle$	[0.30, 0.40]	[0.05-0.05]
	$k_c/\langle k\rangle$	[-0.20, -0.28)]	[0.06, 0.06]
	$\text{var}(\sigma_c)/\text{var}(S_{\text{eff}})$	[0.12, 0.16]	[0.05-0.06]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.17, 0.24]	[0.05-0.05]
	$\text{var}(\sigma_c)$	[0.11, 0.19]	[0.05-0.06]
2	$R_c/\langle R\rangle$	[2.60, 2.71]	[0.09-0.10]
	$\text{var}(R)$	[-0.52, -0.57]	[0.06-0.07]
	$\langle R\rangle$	[0.28, 0.40]	[0.05-0.05]
	$k_c/\langle k\rangle$	[-0.27, -0.34]	[0.06-0.06]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.32, 0.36]	[0.05-0.06]
	I/N	[-0.16, 0.18]	[0.06-0.06]
	$\text{var}(\sigma_c)$	[0.13, 0.19]	[0.06-0.06]
3	$R_c/\langle R\rangle$	[2.36, 2.43]	[0.09-0.10]
	$\text{var}(R)$	[-0.46, -0.57]	[0.07-0.07]
	$\langle R\rangle$	[0.39, 0.47]	[0.06-0.06]
	$k_c/\langle k\rangle$	[-0.32, -0.37]	[0.06-0.06]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.26, 0.29]	[0.06-0.06]
	I	[-0.14, 0.21]	[0.06-0.06]
	$\max\left[[I_c//I_t\right]$	[0.11, 0.19]	[0.06-0.06]
4	$R_c/\langle R\rangle$	[2.53-2.75]	[0.10-0.11]
	$\text{var}(R)$	[-0.47, -0.63]	[0.07-0.08]
	$\langle R\rangle$	[0.30, 0.42]	[0.06-0.06]
	$k_c/\langle k\rangle$	[-0.22, -0.28]	[0.06-0.07]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.30, 0.33]	[0.06-0.06]
	$\text{var}(\sigma_c)$	[0.15, 0.22]	[0.06-0.06]
	I	[-0.15, 0.17]	[0.06-0.06]
5	$R_c/\langle R\rangle$	[2.35, 2.58]	[0.11-0.12]
	$\text{var}(R)$	[-0.46, -0.59]	[0.08-0.08]
	$\langle R\rangle$	[0.31, 0.44]	[0.06-0.06]
	$k_c/\langle k\rangle$	[-0.25, -0.32]	[0.06-0.07]
	$\text{var}(\sigma_c)$	[0.17, 0.23]	[0.06-0.06]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.14, 0.20]	[0.06-0.06]
	$\max\left[I_c/I_t\right]$	[0.11, 0.17]	[0.06-0.06]
6	$R_c/\langle R\rangle$	[2.36, 2.60]	[0.11-0.12]
	$\text{var}(R)$	[-0.60, -0.71]	[0.08-0.08]
	$\langle R\rangle$	[0.32, 0.43]	[0.06-0.06]
	$k_c/\langle k\rangle$	[-0.26, -0.17]	[0.06-0.06]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.19, 0.23]	[0.06-0.06]
	$\text{var}(\sigma_c)$	[0.09, 0.21]	[0.06-0.07]
7	$R_c/\langle R\rangle$	[2.36, 2.59]	[0.12-0.13]
	$\text{var}(R)$	[-0.69, -0.78]	[0.08-0.08]
	$\langle R\rangle$	[0.32, 0.40]	[0.07-0.07]
	$k_c/\langle k\rangle$	[-0.25, -0.32]	[0.07-0.07]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.17, 0.25]	[0.06-0.07]
8	$R_c/\langle R\rangle$	[2.22, 2.40]	[0.11-0.12]
	$\text{var}(R)$	[-0.51, -0.62]	[0.08-0.09]
	$\langle R\rangle$	[0.32, 0.42]	[0.07-0.07]
	$k_c/\langle k\rangle$	[-0.27, 0.34]	[0.07-0.07]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.17, 0.25]	[0.07-0.07]
	$\max\left[I_c/I_t\right]$	[0.15, 0.29]	[0.07-0.07]
9	$R_c/\langle R\rangle$	[2.24, 2.45]	[0.12-0.13]
	$\text{var}(R)$	[-0.48, -0.63]	[0.09-0.10]
	$\langle R\rangle$	[0.30, 0.38]	[0.07.-0.07]
	$k_c/\langle k\rangle$	[-0.22, 0.31]	[0.07-0.07]
	$\max\left[I_c/I_t\right]$	[0.13, 0.25]	[0.07-0.07]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.09, 0.24]	[0.07-0.07]
10	$R_c/\langle R\rangle$	[2.38, 2.55]	[0.13-0.14]
	$\text{var}(R)$	[-0.63, -0.74]	[0.09-0.09]
	$\langle R\rangle$	[0.27, 0.35]	[0.07-0.07]
	$k_c/\langle k\rangle$	[-0.18, -0.23]	[0.07-0.08]
	$\text{var}(\beta_c)/\text{var}(\beta)$	[0.14, 0.22]	[0.08-0.08]
	TMRCA	[-0.10, -0.21]	[0.08-0.08]

Table B.2: **Best-fit model results for surveillance thresholds** $1-10\%$**.** The predictor variables are listed in the order by which they were selected using a forward selection algorithm. The coefficient estimate is the maximum and minimum coefficient (log-odds) from the five-fold cross validation of the final full-term model with the corresponding std. error.

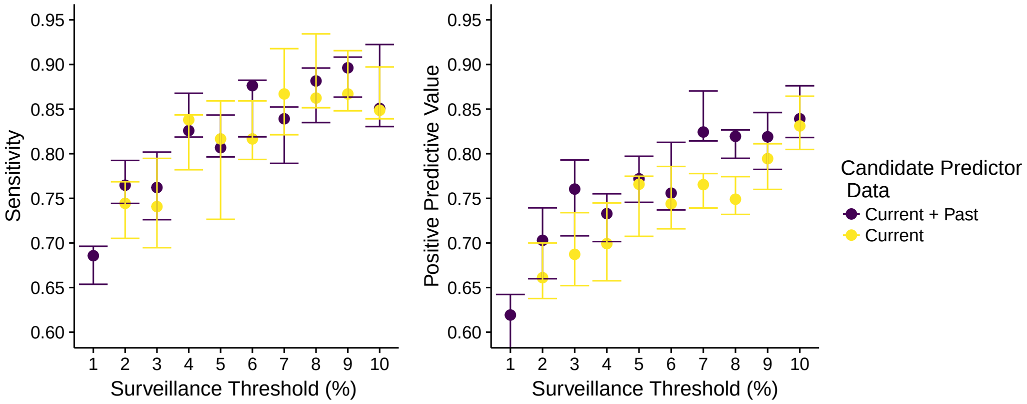

Figure B.3: **Predictive models that rely on data from a single sample of time perform similarly to those that include data from multiple time points**. Each point represents the combination of candidate predictors that best predict the evolutionary fate of antigenic clusters at varying surveillance thresholds, whether the combination includes data from a single time point (yellow) or multiple time points (purple). Model performance is measured in terms of sensitivity and positive predictive value. Dots represent the median, and error bars span the range of performance values across the five folds of cross-validation of the best-fit model.

B.2.3 Proxy Model Results

Because the top selected predictors across all models cannot easily be estimated using readily available surveillance data, we evaluated several proxy measures of viral growth rates and viral competition. The results of the models and how they compare to the performance of a model using only the true relative epidemiological growth rate and the background variance in growth rate can be seen in Table B.3. In addition to the terms included in the table, we tested the fold change of the dominant cluster from t_1 and t_2 as a predictor, but did not find that this term was a significant proxy in any model.

Surveillance Thresholds	**Model**	Type	**Balanced Accuracy**	**AUC**	**PPV**	**Sensitivity**
5%	1. $R_c/\langle R\rangle + \mathrm{var}(R)$	Actual	0.78	0.88	0.81	0.89
1-5%	2. $\delta_c(t_1, t_2) + N_{\Delta_j(t_1,t_2)>1}$	Proxy	0.57	0.71	0.65	0.93
1-5%	3. $\chi_c(t_1, t_2) + \mathrm{var}(\Delta_j(t_1, t_2))$	Proxy	0.50	0.58	0.61	0.99
5%	1. $R_c/\langle R\rangle + \mathrm{var}(R)$	Actual	0.78	0.88	0.81	0.89
3-5%	2. $\delta_c(t_1, t_2) + N_{\Delta_j(t_1,t_2)>1}$	Proxy	0.56	0.67	0.64	0.95
3-5%	3. $\chi_c(t_1, t_2) + \mathrm{var}(\Delta_j(t_1, t_2))$	Proxy	0.50	0.58	0.61	0.99
10%	1.$R_c/\langle R\rangle + \mathrm{var}(R)$	Actual	0.78	0.88	0.81	0.89
6-10%	2.$\delta_c(t_1, t_2) + N_{\Delta_j(t_1,t_2)>1}$	Proxy	0.70	0.78	0.74	0.87
6-10%	3.$\chi_c(t_1, t_2) + \mathrm{var}(\Delta_j(t_1, t_2))$	Proxy	0.59	0.67	0.66	0.95
10%	1. $R_c/\langle R\rangle + \sigma_R$	Actual	0.78	0.88	0.81	0.89
8-10%	2. $\delta_c(t_1, t_2) + N_{\Delta_j(t_1,t_2)>1}$	Proxy	0.63	0.72	0.68	0.95
8-10%	3. $\chi_c(t_1, t_2) + \mathrm{var}(\Delta_j(t_1, t_2))$	Proxy	0.58	0.70	0.65	0.97

Table B.3: **Evaluating proxy measures for different phases of a novel antigenic cluster's early expansion**. Model 1 shows the performance of the best-fit model using the actual values of relative fitness (relative growth rate) and competition (variance in the population growth rate) for clusters that reached the 5% surveillance thresholds (top two sections) and the 10% surveillance threshold (bottom two sections). Within each section, Model 2 substituted a time proxy for the fitness term and the absolute number of clusters that were growing for the competition term. Model 3 substituted a relative fold change for the fitness term and the population-wide variance in fold change for the competition term. t_1 is when a focal cluster reaches the lower surveillance threshold $(1\%, 3\%, 6\%, 8\%)$; t_2 is when the same cluster reaches the higher surveillance threshold $(5\%, 10\%)$ Performance metric values are the median across the five folds in cross-validation. Balanced accuracy measures the accuracy of the model, accounting for the imbalance in outcomes (i.e. number of transient versus established clusters) in the data set.

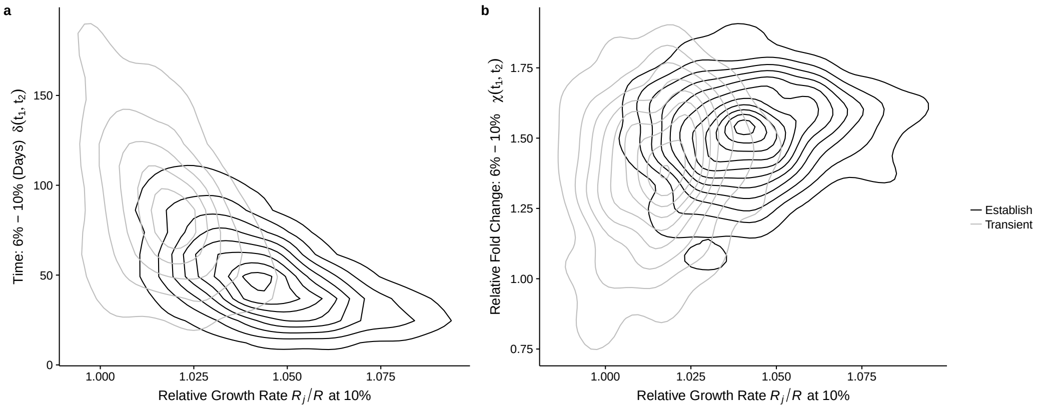

Figure B.4: **The rate of change (a) and relative fold change (b) as proxy measures for the relative growth rate** $R/\langle R\rangle$. Contour lines indicate the density of cluster values for clusters that establish (black) and those that transiently circulate (grey). Values along the x-axis indicate the measured relative growth rate of a cluster the moment it reaches the 10% surveillance threshold. Values along the y-axis indicate the proxy measure (rate of change in (a) and relative fold change in (b) for the cluster, approximated for the time between the 6% and 10% thresholds. The rate of change, measured in the number of days between the two thresholds, is a better proxy measure than relative fold change.

B.3 Alternate Surveillance Paradigms

Instead of a making antigenic transition predictions based on specific surveillance thresholds, one may opportunistically make predictions on cluster fate when samples become available. To compare the robustness of important predictors and model performance under this type of surveillance strategy, we fit two other model types, focused on 1) predicting the evolutionary fate of a cluster and 2) predicting the frequency up to a year out in 3-month increments.

Our data set consisted of all the antigenic clusters present in 10 random time points over a 25 year time period for each of the 62 independent simulations (N = 2846 clusters). We collected candidate predictors for all antigenic clusters that were present above 1% frequency and that had not already surpassed the successful criteria. In addition to the candidate predictors listed in Table B.1, the present frequency of each antigenic cluster, f_c, at the time of sampling was also included as a predictor. To directly compare the two types of surveillance strategies, we found the best fitting model that predicted the antigenic cluster's evolutionary fate using the same cross-validation model fitting process previously described (Fig. B.6.)

Second, rather than a binary transient-successful classification, we predicted frequency levels of present transient clusters at 3 month intervals up to 12 months into the future. Out of the 2846 unique clusters in this dataset, 2279 cluster had $f_c' \geq 0.01\%$ at 3 months; 1921 clusters at 6 months, 1624 clusters at 9 months, and 1378 clusters at 12 months. For each 3 month increment, we went through the following model fitting process. First, classification models were built to assess whether an antigenic cluster would be present or absent in X months time. Next, regression models were fit to predict the frequency for any antigenic cluster that was present above 1% in X months time. At each step, candidate predictor values of the eligible clusters were centered and scaled. To improve model fit, the target frequency, f_c' in X months was log-transformed. We tested the performance of the best-fit model for each 3-month increment on a new data set consisting of 5 random time points over a 25 year period, corresponding to 310 time points over all 62 simulations (Tables B.4, B.5, Fig. B.7). In addition, we tried fitting the model the frequency fold in X months time, i.e. $f_c(t+X)/f_c(t)$; however the model's goodness-of-fit, as measured by the adjusted R^2 was consistently lower than that of the models predicting the log-transformed frequency.

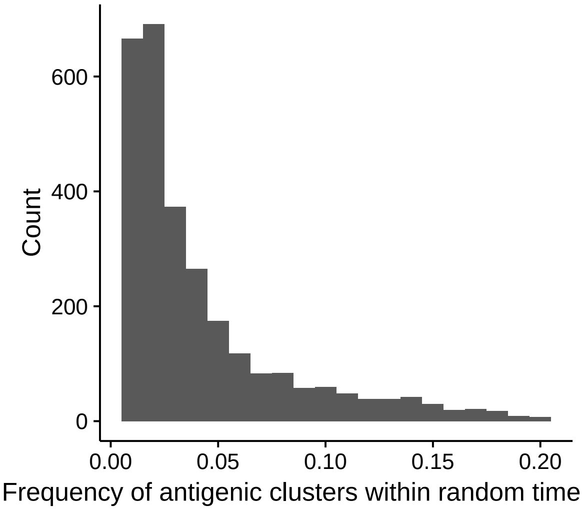

Figure B.5: **The distribution of the 2846 cluster frequencies from 620 random time samples across the 1500 years of influenza evolution.** Clusters that were below 1% relative frequency in the population or those that had already reached our establish criteria were excluded.

Months Ahead	Predictors	AUC	PPV	Sensitivity
3	f_c R_c $\langle R \rangle$ $\beta_c/\langle \beta \rangle$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$	0.93	0.90	0.96
6	f_c R_c $\langle R \rangle$ $\beta_c/\langle \beta \rangle$ $\mathrm{var}(R)$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$ I	0.93	0.89	0.89
9	R_c f_c $\langle R \rangle$ $\beta_c/\langle \beta \rangle$ $\mathrm{var}(R)$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$ tMRCA	0.93	0.84	0.91
12	R_j f_c $\langle R \rangle$ $\mathrm{var}(R)$ $\beta_c/\langle \beta \rangle$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$	0.92	0.81	0.87

Table B.4: **Best-fit logistic regression results for predicting presence-absence of a cluster in X months time into the future.** Terms are listed in the order they were added to the model through forward-selection.

Months Ahead	**Predictors**	R^2**Adjusted**	**RMSE (log)**
3	f_c R_c $\langle R \rangle$ $\beta_c / \langle \beta \rangle$	0.84	0.36
6	f_c R_c $\langle R \rangle$ $\beta_c / \langle \beta \rangle$ $\mathrm{var}(R)$	0.74	0.51
9	R_c f_c $\langle R \rangle$ $\mathrm{var}(R)$ $\beta_c / \langle \beta \rangle$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$	0.66	0.65
12	R_j f_c $\langle R \rangle$ $\mathrm{var}(R)$ $\mathrm{var}(\beta_c)/\mathrm{var}(\beta)$ $\beta_c / \langle \beta \rangle$	0.57	0.77

Table B.5: **Best-fit linear regression models for predicting frequency of a cluster in X months time into the future.** Terms are listed in the order they were added to the model through forward-selection. The R^2Adjusted and Root Mean Squared Error (RMSE) were measured on a testing data set of 5 random time samples over a 25-year period.

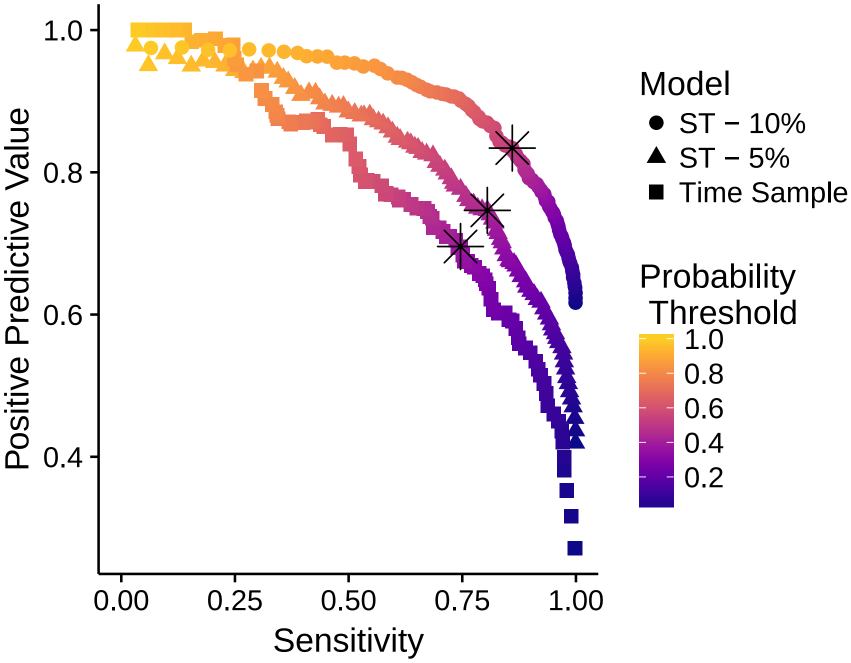

Figure B.6: **Comparing the sensitivity and positive predictive value trade-off of two surveillance strategies: 1) surveillance threshold (circles, triangles) and 2) random sampling through time (squares).** Each line highlights the trade off between sensitivity and positive predictive value at different probability thresholds for what constitutes a positive prediction, i.e. a future successful cluster. All models converge in areas with low sensitivity and high positive predictive value, where the model has to predict with probability 0.90 that the cluster will establish in order to classify it as a positive prediction. However, in areas of greater sensitivity, the time sample model consistently under performs surveillance threshold models based on a 5% or higher surveillance threshold. The black stars represent the probability threshold that maximizes the F1 value.

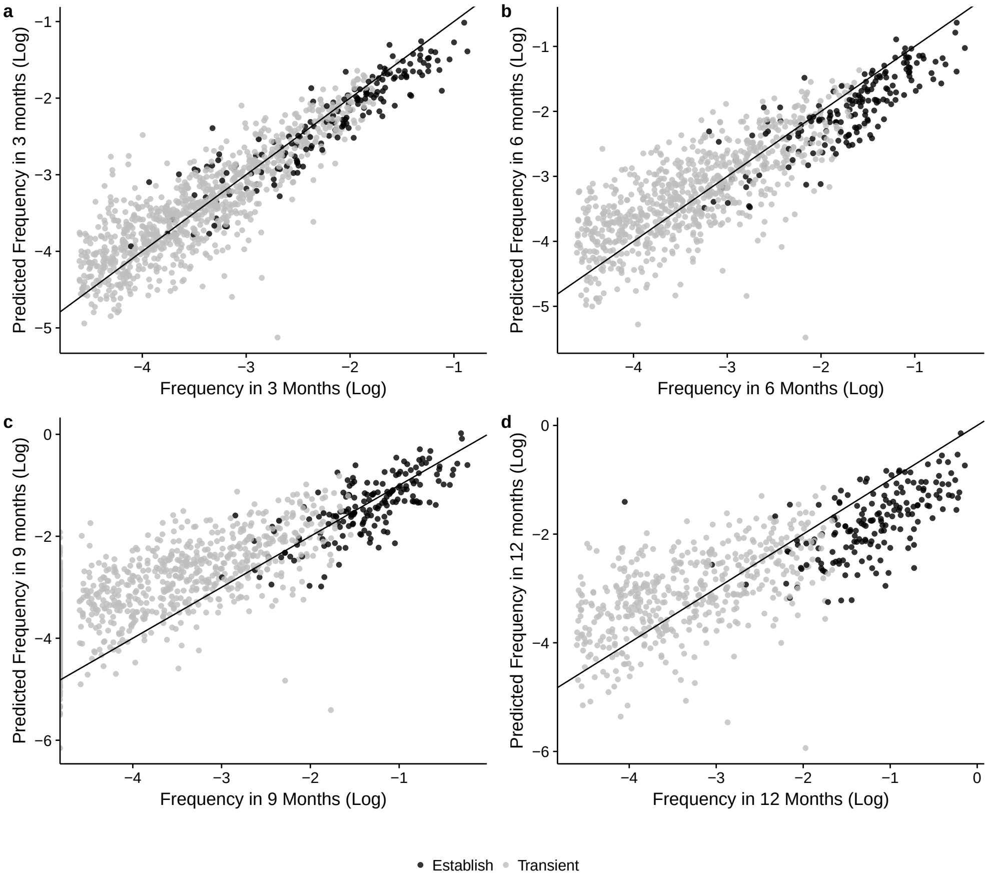

Figure B.7: **Model predictions of cluster frequency up to a year in advance in three-month increments.** Black dots represent clusters that will eventually establish, and grey dots are clusters that will transiently circulate. The black line represents perfect agreement between the actual and predicted log frequencies. The number of clusters present at the time of sampling, but expected to persist in the future, decreases with increasing month-ahead predictions. By 12 months, fewer than half of the starting clusters will still be circulating. As you predict further into the future, the model underestimates future-high frequency circulating clusters, which are usually clusters that will establish.

Appendix C

Within-host HIV-1 biological processes reflected in deconstructed Ancestral Recombination Graphs

C.1 Extended Figures

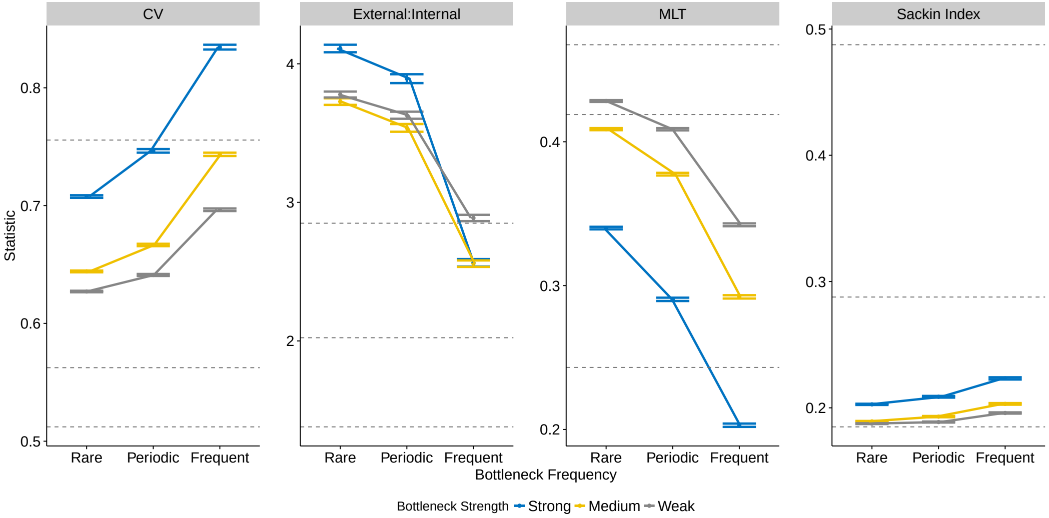

Figure C.1: **The impact of bottleneck strength and frequency on distance and tree statistics.** Bottleneck frequency is measured in days *Rare* [200-365], *Periodic* [100-200) *Frequent* [14-100). Bottleneck strength is the size of N_e during a bottleneck. As N_e grows linearly with time, the bottleneck accounts for progressively greater proportions of the effective population size from the time of infection. Bottleneck Strength Categories: *Strong* [1-30), *Medium* [30-60), *Weak* [60-100]. The grey lines represent the empirical minimum, median, and maximum value of each statistic.

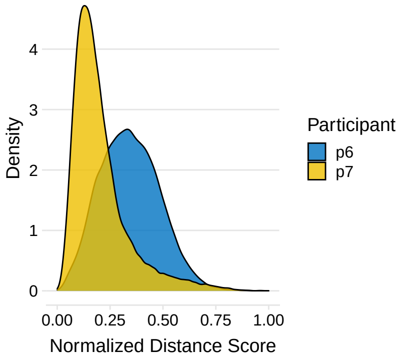

Figure C.2: **Distributions of distance score, d, for each participant (P6: N= 30,918, P7: N = 30,853).** Within each participant, we normalize the distance scores across simulations so that a score $d = 0$ represents the closest-matching simulation and $d = 1$ represents the worst-matching simulation. The median of Participant's 7 score distribution is lower than that of Participant 6 (Wilcoxon Rank Sum, p ¡ 2.2e-16) because simulations with a $\rho = 0.01$ had the lowest scores compared to Participant 6, where $\rho = 0.05$ and $\rho = 0.1$ had lower scores. Because of computational cost and the reduced variability in simulation outcome at higher recombination rates, we sampled 10-100 times less at these recombination rates than at $\rho = 0.01$. Distance summary statistics: Participant 6 (median: 0.34, IQR: 0.24-0.44); Participant 7 (median: 0.17, IQR: 0.12-0.25).

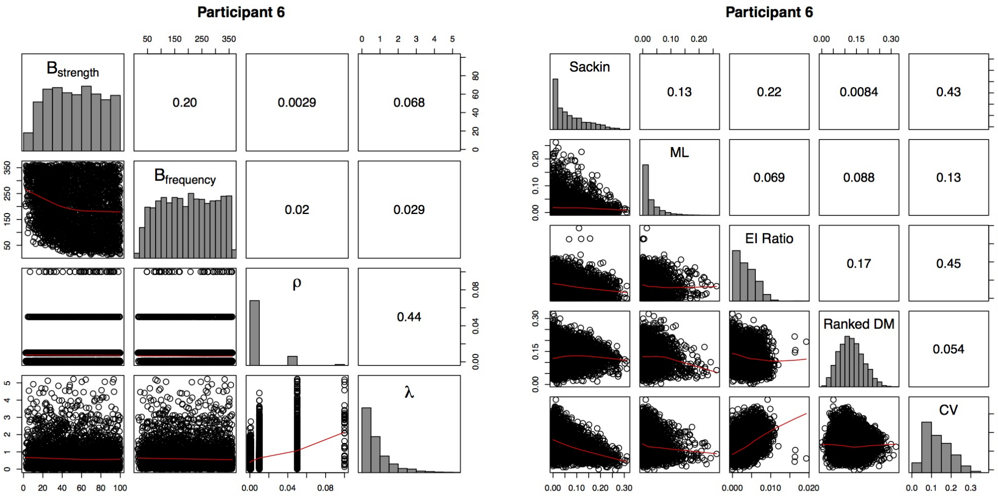

Figure C.3: **Sampling diagnostics for Participant 6**. (Left) Pairwise relationships between parameters in the top 5% matching simulations for Participant 6. (Right) Pairwise relationships between statistical probes in the top 5% matching simulations for Participant 6.

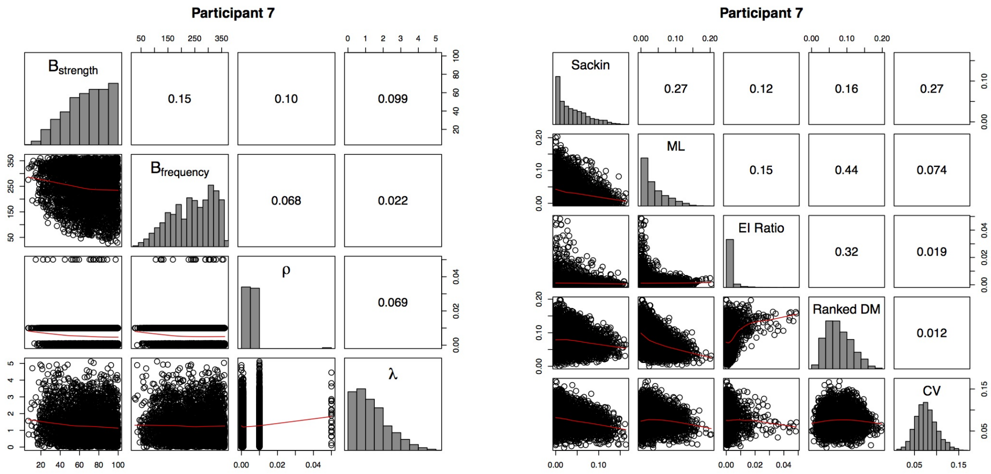

Figure C.4: **Sampling diagnostics for Participant 7**. Pairwise relationships between parameters in the top 5% matching simulations for Participant 7. (Right) Pairwise relationships between statistical probes in the top 5% matching simulations for Participant 7

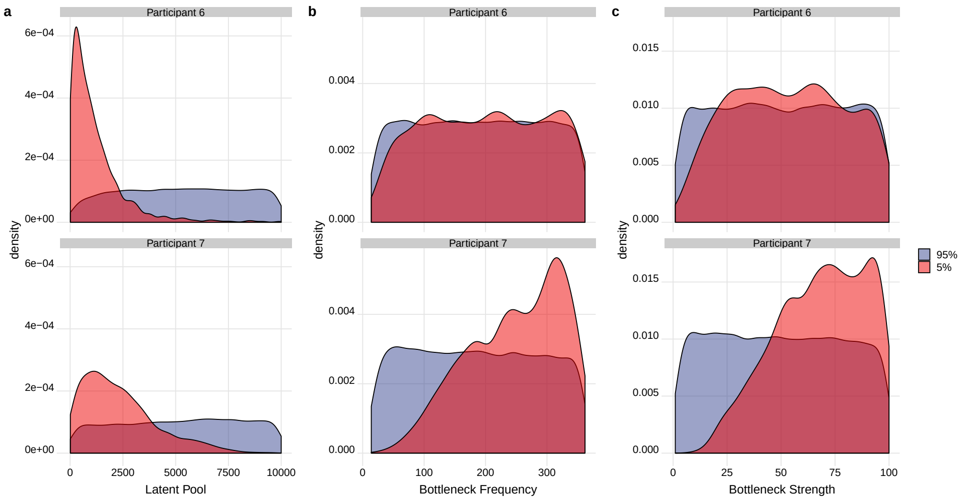

Figure C.5: **Marginal distributions of N_L, bottleneck frequency, and bottleneck strength.** Colors distinguish the density of parameters between simulations with the lowest 5% of distance scores (yellow) and highest 95% of distance scores (blue). Two-sided Kolmogorov-Smirnov tests indicated the distributions for the latent reservoir size and for bottleneck strength are significantly different between parameters in the lowest 5% of scores and those in the top 95% for Participant 6 (Latent Reservoir: $D = 0.735$, $p < 2.2e^{-16}$; Strength: $D = 0.110$, $p = 5.651^{-05}$; Frequency: $D = 0.063$, $p = 0.060$). Two-sided Kolmogorov-Smirnov tests indicated the distributions for all parameters are significantly different for Participant 7 (Latent Reservoir: $D = 0.270$, $p < 2.2e^{-16}$; Strength: $D = 0.308$, $p < 2.2e^{-16}$; Frequency: $D = 0.271$, $p < 2.2e{-}16$). The lowest 5% distributions are significantly different between Participant 6 and Participant 7 for each parameter (Kolmogorov-Smirno two-sided tests, $p < 2.2e^{-16}$). Summary statistics for the lowest 5% (yellow) in the form of Q1-Median-Q3 Patient 6: Latent Pool Size (0.12-0.30-0.63); Strength (32.44-55.43-78.99); Frequency (114.34-202.03-287.80). Patient 7: Latent Reservoir (0.58-1.14-1.85); Strength: (52.46-70.05-86.09); Frequency (187.18-253.63-312.11).

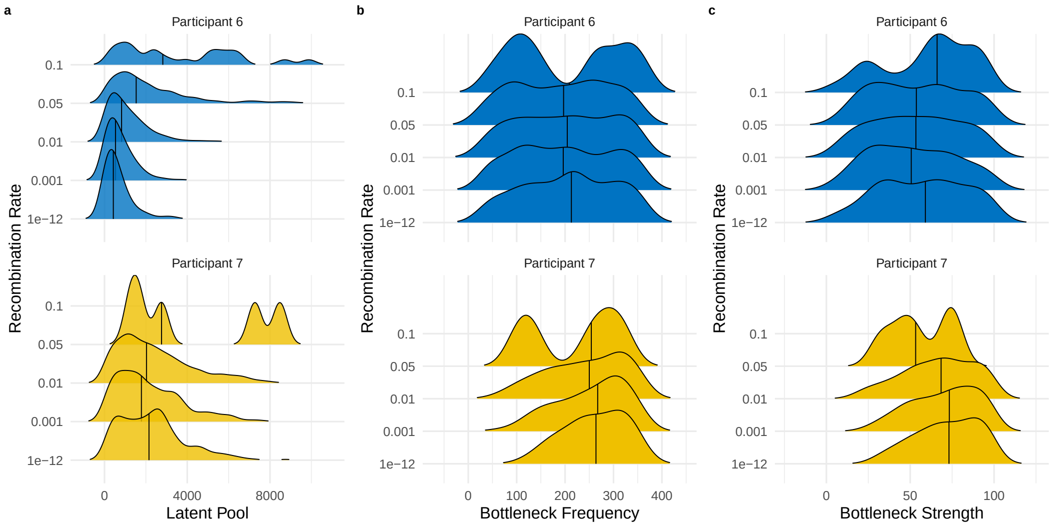

Figure C.6: **Conditional distributions of N_L, bottleneck frequency, and bottleneck strength by recombination rate for the lowest 5% of distance scores.** The line through each ridge represents the 50th quantile. Within each participant, the median of the Latent Reservoir increases with the recombination rate.

C.2 Extended Tables

Recombination Rate ρ	**Latent Reservoir Category N_L**		
	Low	Medium	High
1e-12	343	1035	3336
0.001	342	1091	3477
0.01	358	1018	3389
0.05	38	93	339
0.1	2	10	38

Table C.1: **The numbers of simulations used to calculate the mean and standard error for Figure 4.3**

.

Participant				ρ		
		1e-12	0.001	0.01	0.05	0.1
6	Complete	0.37 (0.13)	0.36 (0.13)	0.31 (0.13)	0.20 (0.10)	0.20 (0.09)
	Top 5%	0.11 (0.02)	0.10 (0.02)	0.09 (0.03)	0.08(0.03)	0.09 (0.02)
7	Complete	0.22(0.14)	0.21 (0.13)	0.18 (0.13)	0.20 (0.11)	0.24 (0.09)
	Top 5%	0.05 (0.01)	0.05 (0.01)	0.04 (0.01)	0.06 (0.00)	

Table C.2: **Mean distance score, $\bar{d}$, and standard deviation by recombination rate ρ for Participant 6 and Participant 7.** Participant 6 Top 5%: $\rho = 1e^{-12}, N = 163$; $\rho = 0.001, N = 242$; $\rho = 0.01, N = 884$; $\rho = 0.05, N = 235$; $\rho = 0.1, N = 22$. Participant 7 Top 5%: $\rho = 1e^{-12}, N = 263$; $\rho = 0.001, N = 421$; $\rho = 0.01, N = 855$; $\rho = 0.05, N = 5$.

Metric	Testing Parameter	Conditional Group	χ^2	d.f.	p-value
Sackin index	N_L	$\rho = 1e^{-12}$	693.24	2	$< 2.2e^{-16}$
		$\rho = 0.001$	694.87	2	$< 2.2e^{-16}$
		$\rho = 0.01$	966.07	2	$< 2.2e^{-16}$
		$\rho = 0.05$	183.58	2	$< 2.2e^{-16}$
		$\rho = 0.1$	17.96	2	$< 2.2e^{-16}$
CV	N_L	$\rho = 1e^{-12}$	36.62	2	$< 2.2e^{-16}$
		$\rho = 0.001$	190.80	2	$< 2.2e^{-16}$
		$\rho = 0.01$	666.07	2	$< 2.2e^{-16}$
		$\rho = 0.05$	26.69	2	$< 2.2e^{-16}$
		$\rho = 0.1$	6.75	2	0.03
EI Ratio	N_L	$\rho = 1e^{-12}$	2590.74	2	$< 2.2e^{-16}$
		$\rho = 0.001$	2568.72	2	$< 2.2e^{-16}$
		$\rho = 0.01$	2107.92	2	$< 2.2e^{-16}$
		$\rho = 0.05$	104.81	2	$< 2.2e^{-16}$
		$\rho = 0.001$	2568.72	2	$< 2.2e^{-16}$
		$\rho = 0.1$	3.08	2	0.21
MLT	N_L	$\rho = 1e^{-12}$	1481.35	2	$< 2.2e^{-16}$
		$\rho = 0.001$	1329.80	2	$< 2.2e^{-16}$
		$\rho = 0.01$	974.41	2	$< 2.2e^{-16}$
		$\rho = 0.05$	103.40	2	$< 2.2e^{-16}$
		$\rho = 0.1$	4.29	2	0.12
Sackin index	ρ	$N_L < 500$	40.00	4	$< 2.2e^{-16}$
		$N_L < 2500$	105.18	4	$< 2.2e^{-16}$
		$N_L < 10000$	544.34	4	$< 2.2e^{-16}$
CV	ρ	$N_L < 500$	546.85	4	$< 2.2e^{-16}$
		$N_L < 2500$	1505.05	4	$< 2.2e^{-16}$
		$N_L < 10000$	3565.02	4	$< 2.2e^{-16}$
EI Ratio	ρ	$N_L < 500$	382.92	4	$< 2.2e^{-16}$
		$N_L < 2500$	170.39	4	$< 2.2e^{-16}$
		$N_L < 10000$	1909.82	4	$< 2.2e^{-16}$
MLT	ρ	$N_L < 500$	71.08	4	$< 2.2e^{-16}$
		$N_L < 2500$	181.28	4	$< 2.2e^{-16}$
		$N_L < 10000$	1261.14	4	$< 2.2e^{-16}$

Table C.3: **Kruskal-Wallis χ^2 ranked sum tests associated with Figure 4.3.** For a given *Metric*, and *Conditional Group*, the Kruskal-Wallis test is testing for significant differences in the means of the *Testing Parameter*. N_L refers to the latent reservoir size and ρ refers to the recombination rate of lineages per day. For example, there is a significant difference in the mean statistic of the Sackin index across possible values of N_L when $\rho = 0.1$. However, there is not a significant difference in the MLT across groups of N_L when $\rho = 0.1$.

Metric	Conditional Parameter	Testing Group 1	Testing Group 2	Z-Statistic	P-value
Sackin	$\rho = 0.1$	$N_L < 2500$	$N_L < 10000$	0.823	0.205
CV	$\rho = 1e^{-12}$	$N_L < 2500$	$N_L < 10000$	-1.52	0.063
	$\rho = 0.001$	$N_L < 2500$	$N_L < 10000$	-1.357	0.087
	$\rho = 0.01$	$N_L < 2500$	$N_L < 10000$	1.302	0.193
	$\rho = .05$	$N_L < 2500$	$N_L < 10000$	0.631	0.0264
	$\rho = 0.1$	$N_L < 500$	$N_L < 10000$	-0.132	0.448
EI Ratio	$\rho = 0.05$	$N_L < 2500$	$N_L < 10000$	-1.83	0.034
	$\rho = 0.1$	$N_L < 500$	$N_L < 2500$	1.314	0.2833
	$\rho = 0.1$	$N_L < 500$	$N_L < 10000$	1.29	0.196
	$\rho = 0.1$	$N_L < 2500$	$N_L < 10000$	-0.638	0.262
MLT	$\rho = 0.1$	$N_L < 500$	$N_L < 2500$	1.92	0.082
	$\rho = 0.1$	$N_L < 500$	$N_L < 10000$	0.97	0.1665
	$\rho = 0.1$	$N_L < 2500$	$N_L < 10000$	-1.36	0.175
Sackin	$N_L < 500$	$\rho = 1e^{-12}$	$\rho = 0.001$	-2.36	0.055
	$N_L < 500$	$\rho = 1e^{-12}$	$\rho = 0.01$	-2.55	0.037
	$N_L < 500$	$\rho = 1e^{-12}$	$\rho = 0.1$	1.93	0.081
	$N_L < 500$	$\rho = 0.001$	$\rho = 0.01$	1.67	0.434
	$N_L < 500$	$\rho = 0.001$	$\rho = 0.1$	-2.184	0.058
	$N_L < 500$	$\rho = .01$	$\rho = 0.1$	-2.20	0.069
	$N_L < 500$	$\rho = 0.05$	$\rho = 0.1$	-0.086	0.389
	$N_L < 2500$	$\rho = 1e^{-12}$	$\rho = 0.01$	2.125	0.050
	$N_L < 2500$	$\rho = 0.001$	$\rho = 0.01$	1.270	0.204
	$N_L < 2500$	$\rho = 0.05$	$\rho = 0.1$	-0.617	0.269
	$N_L < 10000$	$\rho = 1e^{-12}$	$\rho = 0.1$	-1.816	0.173
	$N_L < 10000$	$\rho = 0.001$	$\rho = 0.1$	1.00	0.460
	$N_L < 10000$	$\rho = 0.001$	$\rho = 0.05$	1.298	0.243
	$N_L < 10000$	$\rho = 0.01$	$\rho = 0.1$	-1.666	0.192
	$N_L < 10000$	$\rho = 0.05$	$\rho = 0.1$	-0.370	0.712
CV	$N_L < 500$	$\rho = 0.01$	$\rho = 0.1$	1.139	0.255
	$N_L < 500$	$\rho = 0.05$	$\rho = 0.1$	0.255	0.400
	$N_L < 2500$	$\rho = 0.05$	$\rho = 0.1$	0.548	0.292
	$N_L < 10000$	$\rho = 0.05$	$\rho = 0.1$	0.183	0.427
MLT	$N_L < 500$	$\rho = 1e^{-16}$	$\rho = 0.001$	0.507	0.612
	$N_L < 500$	$\rho = 1e^{-16}$	$\rho = 0.01$	-2.119	0.068
	$N_L < 500$	$\rho = 1e^{-16}$	$\rho = 0.1$	-2.242	0.0623
	$N_L < 500$	$\rho = 0.001$	$\rho = 0.1$	2.300	0.065
	$N_L < 500$	$\rho = 0.01$	$\rho = 0.1$	2.017	0.066
	$N_L < 500$	$\rho = 0.05$	$\rho = 0.1$	0.390	0.348
	$N_L < 2500$	$\rho = 0.05$	$\rho = 0.1$	0.541	0.589
	$N_L < 10000$	$\rho = 0.05$	$\rho = 0.1$	1.805	0.036
EI Ratio	$N_L < 500$	$\rho = 1e^{-12}$	$\rho = 0.001$	1.920	0.082
	$N_L < 500$	$\rho = 0.01$	$\rho = 0.1$	-1.525	0.127
	$N_L < 500$	$\rho = 0.05$	$\rho = 0.1$	-0.163	0.435
	$N_L < 2500$	$\rho = 1e^{-12}$	$\rho = 0.001$	-0.987	0.324
	$N_L < 2500$	$\rho = 1e^{-12}$	$\rho = 0.01$	0.087	0.465
	$N_L < 2500$	$\rho = 0.001$	$\rho = 0.01$	-1.071	0.426
	$N_L < 10000$	$\rho = 1e^{-12}$	$\rho = 0.1$	-2.171	0.045
	$N_L < 10000$	$\rho = 0.001$	$\rho = 0.01$	0.871	0.192
	$N_L < 10000$	$\rho = 0.01$	$\rho = 0.05$	-2.049	0.041

Table C.4: **Insignificant pairwise comparisons between *Testing Group 1* and *Testing Group 2* for a given *Conditional Parameter* using Dunn's z-test-statistic.** The pairwise The p-value is adjusted using the Holm's adjusted method for multiple comparison correction. The null hypothesis for each pairwise comparison is rejected if the p-value is less than 0.5α, where $\alpha = 0.05$. All other pairwise comparisons are significant.

Bibliography

[1] Ben Adams and Alice Carolyn McHardy. The impact of seasonal and year-round transmission regimes on the evolution of influenza A virus. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1716):2249–56, aug 2011.

[2] T. Alex Perkins, Amir S. Siraj, Corrine W. Ruktanonchai, Moritz U. G. Kraemer, Andrew J. Tatem, J. Mlakar, P. Brasil, J. P. Messina, S. Cauchemez, M. A. Johansson, M. U. G. Kraemer, D. R. Lucey, L. O. Gostin, J. Elith, J. Leathwick, W. Kermack, A. McKendrick, A. Sorichetta, K. Sergon, D. L. Smith, O. J. Brady, M. Chan, M. A. Johansson, S. Funk, P. Reiter, V. Andreasen, K. A. Liebman, S. Bhatt, L. Feldstein, J. Brownstein, O. J. Brady, S. I. Hay, M. A. Johansson, A. J. Tatem, R. Hijmans, W. Nordhaus, B. P. Taylor, C. J. E. Metcalf, A. J. Tatem, F. R. Stevens, D. Weiss, P. Gerardin, D. Wright, N. Pya, S. Wood, O. J. Brady, L. Muir, B. Kay, H. Nishiura, S. Halstead, T. W. Scott, R. C. Reiner, J. Lessler, J. Ma, D. J. Earn, M. U. G. Kraemer, M. Duffy, N. Schwarz, D. Sissoko, B. Dwibedi, N. Gay, M. Moro, and A. Balmaseda. Model-based projections of Zika virus infections in childbearing women in the Americas. *Nature Microbiology*, 1(9):16126, 7 2016.

[3] Jorge A. Alfaro-Murillo, Alyssa S. Parpia, Meagan C. Fitzpatrick, Jules A.

Tamagnan, Jan Medlock, Martial L. Ndeffo-Mbah, Durland Fish, María L Ávila-Agüero, Rodrigo Marín, Albert I. Ko, and Alison P. Galvani. A Cost-Effectiveness Tool for Informing Policies on Zika Virus Control. *PLOS Negl*, 10(5):e0004743, 5 2016.

[4] Benjamin M. Althouse, Yih Yng Ng, and Derek A. T. Cummings. Prediction of Dengue Incidence Using Search Query Surveillance. *PLoS Neglected Tropical Diseases*, 5(8):e1258, aug 2011.

[5] Benjamin M Althouse, Samuel V. Scarpino, Lauren Ancel Meyers, John W. Ayers, Marisa Bargsten, Joan Baumbach, John S Brownstein, Lauren Castro, Hannah Clapham, Derek Cummings, Sara Del Valle, Stephen Eubank, Geoffrey Fairchild, Lyn Finelli, Nicholas Generous, Dylan George, David R Harper, Laurent Hébert-Dufresne, Michael A Johansson, Kevin Konty, Marc Lipsitch, Gabriel Milinovich, Joseph D. Miller, Elaine O. Nsoesie, Donald R. Olson, Michael Paul, Philip M. Polgreen, Reid Priedhorsky, Jonathan M. Read, Isabel Rodríguez-Barraquer, Derek J. Smith, Christian Stefansen, David L. Swerdlow, Deborah Thompson, Alessandro Vespignani, and Amy Wesolowski. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*, 4(1):17, dec 2015.

[6] Miguel Arenas and David Posada. The Effect of Recombination on the Reconstruction of Ancestral Sequences. *Genetics*, 184:1133–1139, 2010.

[7] Nimalan Arinaminpathy, Oliver Ratmann, Katia Koelle, Suzanne L Epstein, Graeme E. Price, Cecile Viboud, Mark A. Miller, and Bryan T. Grenfell. Impact of cross-protective vaccines on epidemiological and evolutionary dynamics of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 109(8):3173–7, feb 2012.

[8] Prasith Baccam, Catherine Beauchemin, Catherine A Macken, Frederick G Hayden, and Alan S Perelson. Kinetics of Influenza A Virus Infection in Humans. *Journal of Virology*, 80(15):7590–7599, 2006.

[9] Rebecca Batorsky, Mary F Kearney, Sarah E Palmer, Frank Maldarelli, Igor M Rouzine, and John M Coffin. Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. *Proceedings of the National Academy of Sciences of the United States of America*, 108(14):5661–6, 4 2011.

[10] Trevor Bedford, Sarah Cobey, Peter Beerli, and Mercedes Pascual. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLoS Pathogens*, 6(5):e1000918, may 2010.

[11] Trevor Bedford and Richard A Neher. Seasonal influenza circulation patterns and projections for Feb 2018 to Feb 2019. *bioRxiv*, 2018.

[12] Trevor Bedford, Andrew Rambaut, and Mercedes Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biology*, 10(38), 2012.

[13] Trevor Bedford, Steven Riley, Ian G. Barr, Shobha Broor, Mandeep Chadha, Nancy J. Cox, Rodney S. Daniels, C. Palani Gunasekaran, Aeron C. Hurt, Anne Kelso, Alexander Klimov, Nicola S. Lewis, Xiyan Li, John W. McCauley, Takato Odagiri, Varsha Potdar, Andrew Rambaut, Yuelong Shu, Eugene Skepner, Derek J. Smith, Marc A. Suchard, Masato Tashiro, Dayan Wang, Xiyan Xu, Philippe Lemey, and Colin A. Russell. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, 2015.

[14] Trevor Bedford, Marc A Suchard, Philippe Lemey, Gytis Dudas, Victoria Gregory, Alan J Hay, John W Mccauley, Colin A Russell, Derek J Smith, and Andrew Rambaut. Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 3:1914, 2014.

[15] Samir Bhatt, Edward C Holmes, and Oliver G Pybus. The genomic rate of molecular adaptation of the human influenza A virus. *Molecular Biology and Evolution*, 28(9):2443–2451, 2011.

[16] Matthew Biggerstaff, Simon Cauchemez, Carrie Reed, Manoj Gambhir, Lyn Finelli, R Mikolajczyk, M Massari, S Salmaso, GS Tomba, J Wallinga, J Heijne, M Sadkowska-Todys, M Rosinska, WJ Edmunds, LA Kamimoto, TL Merlin, M Nowell, SC Redd, C Reed, A Schuchat, MI Meltzer, RC Brunham, DM Guevara, F Checchi, E Garcia, S Hugonnet, and C Roth. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature.

BMC Infectious Diseases, 14(1):480, dec 2014.

[17] William C. Black, Kristine E. Bennett, Norma Gorrochótegui-Escalante, Carolina V. Barillas-Mury, Ildefonso Fernández-Salas, Marıa de Lourdes Muñoz, José A. Farfán-Alé, Ken E. Olson, and Barry J. Beaty. Flavivirus Susceptibility in Aedes aegypti. *Archives of Medical Research*, 33(4):379–388, 7 2002.

[18] Jesse D Bloom and Matthew J Glassman. Inferring Stabilizing Mutations from Protein Phylogenies: Application to Influenza Hemagglutinin. *PLoS Computational Biology*, 5(4):e1000349, apr 2009.

[19] Seth Blumberg and James O. Lloyd-Smith. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics*, 5(3):131–145, sep 2013.

[20] Peter Bogner, Ilaria Capua, David J Lipman, Nancy J Cox, and Others. A global initiative on sharing avian flu data. *Nature*, 442:981, aug 2006.

[21] Oliver J Brady, Michael A Johansson, Carlos A Guerra, Samir Bhatt, Nick Golding, David M Pigott, Hélène Delatte, Marta G Grech, Paul T Leisnham, Rafael Maciel-de Freitas, Linda M Styer, David L Smith, Thomas W Scott, Peter W Gething, and Simon I Hay. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors*, 6(1):351, 2013.

[22] Evelien M Bunnik, Linaida Pisas, Ad C van Nuenen, and Hanneke Schuitemaker. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. *Journal of virology*, 82(16):7932–41, aug 2008.

[23] Robin M Bush, Walter M Fitch, Catherine A Bender, and Nancy J Cox. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol*, 16, 1999.

[24] Fabrice Carrat and Antoine Flahault. Influenza vaccine: the challenge of antigenic drift. *Vaccine*, 25(39-40):6852–62, sep 2007.

[25] Fabrice Carrat, Elisabeta Vergu, Neil M Ferguson, Magali Lemaitre, Simon Cauchemez, Steve Leach, and Alain-Jacques Valleron. Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am J Epidemiol*, 167, 2008.

[26] Lauren A Castro, Trevor Bedford, and Lauren A Meyers. Early Prediction of Antigenic Transitions for Influenza A H3N2. *bioRxiv*, 2019.

[27] Lauren A. Castro, Spencer J. Fox, Xi Chen, Kai Liu, Steven E. Bellan, Nedialko B. Dimitrov, Alison P. Galvani, and Lauren Ancel Meyers. Assessing real-time Zika risk in the United States. *BMC Infectious Diseases*, 17(1):284, 12 2017.

[28] Simon Cauchemez, Scott Epperson, Matthew Biggerstaff, David Swerdlow, Lyn Finelli, and Neil M. Ferguson. Using Routine Surveillance Data to Estimate the Epidemic Potential of Emerging Zoonoses: Application to the Emergence of US Swine Origin Influenza A H3N2v Virus. *PLoS Medicine*, 10(3):e1001399, mar 2013.

[29] Centers for Disease Control and Prevention. Zika Virus Risk-Based Preparedness and Response Guidance for States, 2016.

[30] Centers for Disease Control and Prevention. Summary of the 2017-2018 Influenza Season, 2018.

[31] National Center for Immunization Centers for Disease Control and Prevention and Respiratory Diseases (NCIRD). Flu Activity and Surveillance, 2019.

[32] Centers for Disease Control and Prevention (Organization). *Principles of Epidemiology: Home—Self-Study Course SS1978—CDC*. 2012.

[33] Miranda Chan and Michael A. Johansson. The Incubation Periods of Dengue Viruses. *PLoS ONE*, 7(11):e50972, 11 2012.

[34] Tae-Wook Chun, Delphine Engel, M Michelle Berrey, Theresa Shea, Lawrence Corey, and Anthony S Fauci. Early establishment of a pool of latently infected, resting CD4+ T cells during primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, 95(15):8869–8873, 1998.

[35] Tae-Wook Chun, Lieven Stuyver, Stephanie B Mizell, Linda A Ehler, Jo Ann M Mican, Michael Baseler, Alun L Lloyd, Martin A Nowak, and Anthony S Fauci. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13193–7, nov 1997.

[36] Arthur Chun-Chieh Shih, Tzu-Chang Hsiao, Mei-Shang Ho, and Wen-Hsiung Li. Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 2007.

[37] James F. Crow and Motoo Kimura. Evolution in sexual and asexual population. *The American Naturalist*, 99(909):439–450, 1965.

[38] Gabor Csardi. Package 'igraph' Title Network Analysis and Visualization, 2019.

[39] Adel Dayarian and Boris I Shraiman. How to infer relative fitness from a sample of genomic sequences. *Genetics*, 197(3):913–23, jul 2014.

[40] EJ Dechant and JG Rigau-Perez. Hospitalizations for suspected dengue in Puerto Rico, 1991-1995: estimation by capture-recapture methods. The Puerto Rico Association of Epidemiologists. *Am J Trop Med Hyg*, 61(4):574–578, 10 1999.

[41] Richard Desper and Olivier Gascuel. Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle. pages 357–374. Springer, Berlin, Heidelberg, 2002.

[42] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*, 34(4):997–1007, jan 2017.

[43] Hilje M. Doekes, Christophe Fraser, and Katrina A. Lythgoe. Effect of the Latent Reservoir on the Evolution of HIV at the Within- and Between-Host Levels. *PLoS Computational Biology*, 2017.

[44] Mark R. Duffy, Tai-Ho Chen, W. Thane Hancock, Ann M. Powers, Jacob L. Kool, Robert S. Lanciotti, Moses Pretrick, Maria Marfel, Stacey Holzbauer, Christine Dubray, Laurent Guillaumot, Anne Griggs, Martin Bel, Amy J. Lambert, Janeen Laven, Olga Kosoy, Amanda Panella, Brad J. Biggerstaff, Marc Fischer, and Edward B. Hayes. Zika virus outbreak on Yap Island, Federated States of Micronesia. *The New England journal of medicine*, 360(24):2536–2543, 6 2009.

[45] Jonathan Dushoff, Joshua B Plotkin, Simon A Levin, and David J D Earn. Dynamical resonance can account for seasonality of influenza epidemics. *Proceedings of the National Academies of Science*, 30:16915–16916, 2004.

[46] Joseph Felsenstein. The evolutionary advantage of recombination. *Genetics*, 78(2), 1974.

[47] Diana Finzi, Monika Hermankova, Theodore Pierson, Lucy M Carruth, Christopher Buck, Richard E Chaisson, Thomas C Quinn, Karen Chadwick, Joseph Margolick, Ronald Brookmeyer, Joel Gallant, Martin Markowitz, David D Ho, Douglas D Richman, and Robert F Siliciano. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science*, 278(5341):1295–1300, 1997.

[48] Ronald A Fisher. *The genetical theory of natural selection.* Clarendon Press, Oxford, 1930.

[49] Florida Department of Health, Office of Communications. Department of Health Daily Zika Update. Technical report, Florida Department of Health, 2017.

[50] J M Fonville, S H Wilks, S L James, A Fox, M Ventresca, M Aban, L Xue, T C Jones, N M H Le, Q T Pham, N D Tran, Y Wong, A Mosterin, L C Katzelnick, D Labonte, T T Le, G van der Net, E Skepner, C A Russell, T D Kaplan, G F Rimmelzwaan, N Masurel, J C de Jong, A Palache, W E P Beyer, Q M Le, T H Nguyen, H F L Wertheim, A C Hurt, A D M E Osterhaus, I G Barr, R A M Fouchier, P W Horby, and D J Smith. Antibody landscapes after influenza virus infection or vaccination. *Science (New York, N.Y.)*, 346(6212):996–1000, nov 2014.

[51] Centers for Disease Control and Prevention. Estimated range of Aedes aegypti and Aedes albopictus in the United States, 2016.

[52] Centers for Disease Control and Prevention. Interim CDC recommendations for Zika vector control in the continental United States, 2016.

[53] Sylvain Gandon, Troy Day, C. Jessica E. Metcalf, and Bryan T. Grenfell. Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases. *Trends in Ecology and Evolution*, 31(10):776–788, 2016.

[54] Feng Gao, Yalu Chen, David N Levy, Joan A Conway, Thomas B Kepler, and Huxiong Hui. Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. *Journal of virology*, 78(5):2426–33, mar 2004.

[55] Federica Giardina, Ethan Obie Romero-Severson, Jan Albert, Tom Britton, and Thomas Leitner. Inference of Transmission Network Structure from HIV Phylogenetic Trees. *PLOS Computational Biology*, 13(1):e1005316, jan 2017.

[56] Paolo Giorgi Rossi, Flavia Riccardo, Annamaria Pezzarossi, Paola Ballotari, Maria Grazia Dente, Christian Napoli, Antonio Chiarenza, Cesar Velasco Munoz, Teymur Noori, and Silvia Declich. Factors Influencing the Accuracy of Infectious Disease Reporting in Migrants: A Scoping Review. *International journal of environmental research and public health*, 14(7), 2017.

[57] S. K. Gire, A. Goba, K. G. Andersen, R. S. G. Sealfon, D. J. Park, L. Kanneh, S. Jalloh, M. Momoh, M. Fullah, G. Dudas, S. Wohl, L. M. Moses, N. L. Yozwiak, S. Winnicki, C. B. Matranga, C. M. Malboeuf, J. Qu, A. D. Gladden, S. F. Schaffner, X. Yang, P.-P. Jiang, M. Nekoui, A. Colubri, M. R. Coomber, M. Fonnie, A. Moigboi, M. Gbakie, F. K. Kamara, V. Tucker, E. Konuwa, S. Saffa, J. Sellu, A. A. Jalloh, A. Kovoma, J. Koninga, I. Mustapha, K. Kargbo, M. Foday, M. Yillah, F. Kanneh, W. Robert, J. L. B. Massally, S. B. Chapman, J. Bochicchio, C. Murphy, C. Nusbaum, S. Young, B. W. Birren, D. S. Grant, J. S. Scheiffelin, E. S. Lander, C. Happi, S. M. Gevao, A. Gnirke, A. Rambaut, R. F. Garry, S. H. Khan, and P. C. Sabeti. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–72, aug 2014.

[58] Rebecca Tave Gluskin, Michael A. Johansson, Mauricio Santillana, and John S. Brownstein. Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends. *PLoS Neglected Tropical Diseases*, 8(2):e2713, feb 2014.

[59] Benjamin H Good, Igor M Rouzine, Daniel J Balick, Oskar Hallatschek, Michael M Desai, and Richard E Lenski. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. *PNAS*, 109(13):4950–4955, 2012.

[60] Rebecca R Gray, Joe Parker, Philippe Lemey, Marco Salemi, Aris Kat-

zourakis, and Oliver G Pybus. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC evolutionary biology*, 11(1):131, jan 2011.

[61] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James L N Wood, Janet M Daly, Jenny a Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, 303(5656):327–332, 2004.

[62] R.C. Griffiths and P. Marjoram. Ancestral Inference from Samples of DNA Sequences with Recombination. *Journal of Computational Biology*, 3(4):479–502, 1 1996.

[63] Zhengxian Gu, Qing Gao, Emmanuel A Faust, and Mark A Wainberg. Possible involvement of cell fusion and viral recombination in generation of human immunodeficiency virus variants that display dual resistance to AZT and 3TC. *Journal of General Virology*, pages 2601–2605, 1995.

[64] Anne Gulland. Zika virus is a global public health emergency, declares WHO. *BMJ*, 352, 2016.

[65] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, may 2018.

[66] William T. Harvey, Donald J. Benton, Victoria Gregory, James P. J. Hall, Rodney S. Daniels, Trevor Bedford, Daniel T. Haydon, Alan J. Hay, John W. McCauley, and Richard Reeve. Identification of Low- and High-Impact Hemagglutinin Amino Acid Substitutions That Drive Antigenic Drift of Influenza A(H1N1) Viruses. *PLOS Pathogens*, 12(4):e1005526, apr 2016.

[67] David L Heyman. *Control of communicable diseases manual, 18th ed.* American Public Health Association, 2004.

[68] Alison L Hill. Mathematical Models of HIV Latency. *Current Topics in Microbiology and Immunology*, 417:131–156, 2017.

[69] Ya-Chi Ho, Liang Shan, Nina N. Hosmane, Jeffrey Wang, Sarah B. Laskey, Daniel I.S. Rosenbloom, Jun Lai, Joel N. Blankson, Janet D. Siliciano, and Robert F. Siliciano. Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure. *Cell*, 155(3):540–551, 10 2013.

[70] Sean Hoban, Giorgio Bertorelle, and Oscar E. Gaggiotti. Computer simulations: tools for population and evolutionary genetics. *Nature Reviews Genetics*, 13(2):110–122, feb 2012.

[71] Taina T. Immonen, Jessica M. Conway, Ethan O. Romero-Severson, Alan S. Perelson, and Thomas Leitner. Recombination Enhances HIV-1

Envelope Diversity by Facilitating the Survival of Latent Genomic Fragments in the Plasma Virus Population. *PLoS Computational Biology*, 2015.

[72] Thijs Janzen, Sebastian Hohna, and Rampal S Etienne. Approximate Bayesian Computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT. *Methods in Ecology and Evolution*, 6:566–575, 2015.

[73] E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, 5 1957.

[74] Amanda E. Jetzt, Hong Yu, George J. Klarmann, Yacov Ron, Bradley D. Preston, and Joseph P. Dougherty. High Rate of Recombination throughout the Human Immunodeficiency Virus Type 1 Genome. *Journal of Virology*, 74(3):1234–1240, feb 2000.

[75] Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1):e1003457, jan 2014.

[76] Lina Josefsson, Sarah Palmer, Nuno R. Faria, Philippe Lemey, Joseph Casazza, David Ambrozak, Mary Kearney, Wei Shao, Shyamasundaran Kottilil, Michael Sneller, John Mellors, John M. Coffin, and Frank Maldarelli. Single Cell Analysis of Lymph Node Tissue from HIV-1 Infected

Patients Reveals that the Majority of CD4+ T-cells Contain One HIV-1 DNA Molecule. *PLoS Pathogens*, 9(6):e1003432, jun 2013.

[77] Rowland R Kao. The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *Trends in Microbiology*, 10(6), 2002.

[78] Paul Kellam and Brendan A Larder. Retroviral recombination can lead to linkage of reverse transcriptase mutations that confer increased zidovudine resistance. *Journal of Virology*, 69(2):669–74, feb 1995.

[79] Eben Kenah, Tom Britton, M. Elizabeth Halloran, and Ira M. Longini. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. *PLOS Computational Biology*, 12(4):e1004869, apr 2016.

[80] Katia Koelle, Meredith Kamradt, and Mercedes Pascual. Understanding the dynamics of rapidly evolving pathogens through modeling the tempo of antigenic change: Influenza as a case study. *Epidemics*, 1(2):129–137, 2009.

[81] Katia Koelle and David A. Rasmussen. Prediction is worth a shot. *Nature*, 507(7490):47–48, mar 2014.

[82] Katia Koelle and David A. Rasmussen. The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *eLife*, 4(September):1–31, 2015.

[83] Fyodor A Kondrashov and Alexey S Kondrashov. Multidimensional epistasis and the disadvantage of sex. *Proceedings of the National Academy of Sciences*, 98(21):12089–12092, 2001.

[84] Moritz U G Kraemer, Marianne E Sinka, Kirsten A Duda, Adrian Mylne, Freya M Shearer, Christopher M Barker, Chester G Moore, Roberta G Carvalho, Giovanini E Coelho, Wim Van Bortel, Guy Hendrickx, Francis Schaffner, Iqbal RF Elyazar, Hwa-Jen Teng, Oliver J Brady, Jane P Messina, David M Pigott, Thomas W Scott, David L Smith, GR W Wint, Nick Golding, and Simon I Hay. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *eLife*, 4:e08347, 6 2015.

[85] Adam Kucharski and Julia R Gog. Influenza emergence in the face of evolutionary constraints. *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1729):645–52, feb 2012.

[86] Adam J. Kucharski, Sebastian Funk, Rosalind M. Eggo, Henri-Pierre Mallet, W. John Edmunds, Eric J. Nilles, EB Hayes, MR Duffy, TH Chen, WT Hancock, AM Powers, JL Kool, RS Lanciotti, VM Cao-Lormeau, C Roche, A Teissier, E Robin, AL Berry, HP Mallet, D Musso, VM Cao-Lormeau, DJ Gubler, A Roth, A Mercier, C Lepers, D Hoy, S Duituturaga, E Benyon, GS Campos, AC Bandeira, SI Sardi, F Corsica, E Camacho, M Paternina-Gomez, PJ Blanco, JE Osorio, MT Aliota, D Musso, C Roche, E Robin, T Nhan, A Teissier, VM Cao-Lormeau,

D Musso, EJ Nilles, VM Cao-Lormeau, II Bogoch, OJ Brady, MU Kraemer, M German, MI Creatore, MA Kulkarni, E Oehler, L Watrin, P Larre, I Leparc-Goffart, S Lastere, F Valour, E Oehler, E Fournier, I Leparc-Goffart, P Larre, S Cubizolle, C Sookhareea, L Schuler-Faccini, VM Cao-Lormeau, A Blake, S Mons, S Lastère, C Roche, J Vanhomwegen, S Cauchemez, M Besnard, P Bompard, T Dub, GA P, EG D, M Keeling, B Grenfell, VM Cao-Lormeau, C Roche, D Musso, HP Mallet, T Dalipanda, A Dofai, A Camacho, S Ballesteros, AL Graham, F Carrat, O Ratmann, B Cazelles, CA Manore, KS Hickmann, S Xu, HJ Wearing, JM Hyman, A Pandey, A Mubayi, J Medlock, V Duong, L Lambrechts, RE Paul, S Ly, RS Lay, KC Long, F Lardeux, J Cheffort, OJ Brady, MA Johansson, CA Guerra, S Bhatt, N Golding, DM Pigott, F Rivière, J Boorman, J Porterfield, M Aubry, J Finke, A Teissier, C Roche, J Broult, S Paulous, C Bretó, D He, EL Ionides, AA King, K Soetaert, T Petzoldt, R Woodrow, H Nishiura, SB Halstead, M Chan, MA Johansson, Ds Li, W Liu, A Guigon, C Mostyn, R Grant, J Aaskov, G Chowell, C Torre, C Munayco-Escate, L Suarez-Ognio, R Lopez-Cruz, J Hyman, LC Harrington, A Fleisher, D Ruiz-Moreno, F Vermeylen, CV Wa, RL Poulson, M Aubry, A Teissier, C Roche, V Richard, AS Yan, K Zisou, and FL Black. Transmission Dynamics of Zika Virus in Island Populations: A Modelling Analysis of the 2013–14 French Polynesia Outbreak. *PLOS Neglected Tropical Diseases*, 10(5):e0004726, 5 2016.

[87] Michael Lässig, Ville Mustonen, and Aleksandra M. Walczak. Predict-

ing evolution. *Nature Ecology & Evolution*, 1(3):0077, feb 2017.

[88] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205, 2014.

[89] Philippe Lemey, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, and Marc A Suchard. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS pathogens*, 10(2), 2014.

[90] Justin T Lessler, Cassandara T Ott, Andrea C Carcelen, Jacob M Konikoff, Joe Williamson, Qifang Bi, Lauren M Kucirka, Derek AT Cummings, Nicholas G Reich, and Lelia H Chaisson. Times to key events in the course of Zika infection and their implications: a systematic review and pooled analysis. *Bulletin of the World Health Organization*, 4 2016.

[91] Joseph Lewnard and Sarah Cobey. Immune History and Influenza Vaccine Effectiveness. *Vaccines*, 6(2):28, may 2018.

[92] Alun L. Lloyd. Realistic Distributions of Infectious Periods in Epidemic Models: Changing Patterns of Persistence and Dynamics. *Theoretical Population Biology*, 60(1):59–71, 2001.

[93] James O. Lloyd-Smith, Sebastian Funk, Angela R. McLean, Steven Riley, and James L N Wood. Nine challenges in modelling the emergence

of novel pathogens. *Epidemics*, 10:35–39, 2014.

[94] Eric T Lofgren, M Elizabeth Halloran, Caitlin M Rivers, John M Drake, Travis C Porco, Bryan Lewis, Wan Yang, Alessandro Vespignani, Jeffrey Shaman, Joseph N S Eisenberg, Marisa C Eisenberg, Madhav Marathe, Samuel V Scarpino, Kathleen A Alexander, Rafael Meza, Matthew J Ferrari, James M Hyman, Lauren A Meyers, and Stephen Eubank. Opinion: Mathematical models: a key tool for outbreak response. *Proceedings of the National Academy of Sciences of the United States of America*, 111(51):18095–6, dec 2014.

[95] Marta Luksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507:57–61, 2014.

[96] Maimuna S Majumder, Emily Cohn, Durland Fish, and John S Brownstein. Estimating a feasible serial interval range for Zika fever. *Bulletin of the World Health Organization*, 2016.

[97] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–94, aug 1995.

[98] Darren P. Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), mar 2015.

[99] Jochen Maydt and Thomas Lengauer. Recco: recombination analysis using cost optimization. *Bioinformatics*, 22(9):1064–1071, may 2006.

[100] Cory Merow, Matthew J. Smith, and John A. Silander. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069, 10 2013.

[101] Richard E. Michod, Harris Bernstein, and Aurora M. Nedelcu. Adaptive value of sex in microbial pathogens. *Infection, Genetics and Evolution*, 8(3):267–285, may 2008.

[102] Noelle-Angelique M. Molinari, Ismael R. Ortega-Sanchez, Mark L. Messonnier, William W. Thompson, Pascale M. Wortley, Eric Weintraub, and Carolyn B. Bridges. The annual impact of seasonal influenza in the US: Measuring disease burden and costs. *Vaccine*, 25(27):5086–5096, jun 2007.

[103] Danesh Moradigaravand, Roger Kouyos, Trevor Hinkley, Mojgan Haddad, Christos J. Petropoulos, Jan Engelstädter, and Sebastian Bonhoeffer. Recombination Accelerates Adaptation on a Large-Scale Empirical Fitness Landscape in HIV-1. *PLoS Genetics*, 10(6):e1004439, jun 2014.

[104] Dylan H Morris, Katelyn M Gostic, Simone Pompei, Trevor Bedford, Marta Łuksza, Richard A Neher, Bryan T Grenfell, Michael Lässig, and John W Mccauley. Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. *Trends in Microbiology*, 2017.

[105] L Moutouh, J Corbeil, and D D Richman. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. *Proceedings of the National Academy of Sciences of the United States of America*, 93(12):6106–11, jun 1996.

[106] L E Muir and B H Kay. Aedes aegypti survival and dispersal estimated by mark-release-recapture in northern Australia. *The American journal of tropical medicine and hygiene*, 58(3):277–82, 3 1998.

[107] H J Muller. Some Genetic Aspects of Sex. *The American Naturalist*, 66(703):118–138, 1932.

[108] John M Murray, John Zaunders, Sean Emery, David A Cooper, William J Hey-Nguyen, Kersten K Koelsch, and Anthony D Kelleher. HIV dynamics linked to memory CD4+ T cell homeostasis. *PLoS one*, 12(10), 2017.

[109] Didier Musso, Van Mai Cao-Lormeau, and Duane J Gubler. Zika virus: following the path of dengue and chikungunya? *Lancet (London, England)*, 386(9990):243–4, 7 2015.

[110] National Foundation for Infectious Diseases. 2018 NFID Influenza/Pneumococcal News Conference, 2018.

[111] Richard A. Neher and Trevor Bedford. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, nov 2015.

[112] Richard A Neher, Trevor Bedford, Rodney S Daniels, Colin A Russell, and Boris I Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 113(12):E1701–9, mar 2016.

[113] Richard A. Neher and Thomas Leitner. Recombination Rate and Selection Strength in HIV Intra-patient Evolution. *PLoS Computational Biology*, 6(1):e1000660, 1 2010.

[114] Richard A Neher, Colin A Russell, and Boris I Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3:e03568, nov 2014.

[115] Martha I Nelson and Edward C Holmes. The evolution of epidemic influenza. *Nature Reviews Genetics*, 8, 2007.

[116] Hiroshi Nishiura and Scott B Halstead. Natural history of dengue virus (DENV)-1 and DENV-4 infections: reanalysis of classic studies. *The Journal of infectious diseases*, 195(7):1007–13, 4 2007.

[117] Eamon B O'Dea and Claus O Wilke. Contact heterogeneity and phylodynamics: how contact networks shape parasite evolutionary trees. *Interdisciplinary perspectives on infectious diseases*, 2011:238743, 2011.

[118] Office of Travel & Tourism Industries. US Monthly Arrivals Trend Line, 2014.

[119] World Health Organization. Zika situation report. Technical report, World Health Organization.

[120] World Health Organization. Influenza (Seasonal). Fact sheet no. 211, 2014.

[121] Dave Osthus, Ashlynn R Daughton, and Reid Priedhorsky. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS Computational Biology*, 15(2), 2019.

[122] Sarah P Otto and Michael C Whitlock. The Probability of Fixation in Populations of Changing Size. *Genetics 14*, 146:723–733, 1997.

[123] Pan American Health Organization and World Health Organization. Cumulative Zika suspected and confirmed cases reported by countries and territories in the Americas, 2015-2016, 2016.

[124] Emmanuel Paradis. Package 'ape' Title Analyses of Phylogenetics and Evolution Depends R ¿= 3.2.0), 2018.

[125] Andrew W Park, Janet M Daly, Nicola S Lewis, Derek J Smith, James LN Wood, and Bryan T Grenfell. Quantifying the impact of immune escape on transmission dynamics of influenza. *Science*, 326, 2009.

[126] David Posada and Keith A Crandall. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. *Proceedings of the National Academy of Sciences*, 98(24):13757–13762, 2001.

[127] David Posada and Keith A Crandall. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal of Molecular Evolution*, 54:396–402, 2002.

[128] David A Rasmussen, Erik M Volz, and Katia Koelle. Phylodynamic inference for structured epidemiological models. *PLoS computational biology*, 10(4):e1003570, apr 2014.

[129] Nicholas G Reich, Logan C Brooks, Spencer J Fox, Sasikiran Kandula, Craig J McGowan, Evan Moore, Dave Osthus, Evan L Ray, Abhinav Tushar, Teresa K Yamana, Matthew Biggerstaff, Michael A Johansson, Roni Rosenfeld, and Jeffrey Shaman. A collaborative multiyear, multi-model assessment of seasonal influenza forecasting in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 116(8):3146–3154, feb 2019.

[130] Douglas D Richman, Terri Wrin, Susan J Little, and Christos J Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4144–9, apr 2003.

[131] Diana Patricia Rojas, Natalie E Dean, Yang Yang, Eben Kenah, Juliana Quintero, Simon Tomasi, Erika Lorena Ramirez, Yendi Kelly, Carolina Castro, Gabriel Carrasquilla, M Elizabeth Halloran, and Ira M Longini. The epidemiology and transmissibility of Zika virus in Girardot and San

Andres island, Colombia, September 2015 to January 2016. *Eurosurveillance*, 21(28):30283, 7 2016.

[132] Ethan Romero-Severson, Helena Skar, Ingo Bulla, Jan Albert, and Thomas Leitner. Timing and order of transmission events is not directly reflected in a pathogen phylogeny. *Molecular Biology and Evolution*, 2014.

[133] Ethan O. Romero-Severson, Ingo Bulla, Nick Hengartner, Inês Bártolo, Ana Abecasis, José M. Azevedo-Pereira, Nuno Taveira, and Thomas Leitner. Donor-recipient identification in para- and poly-phyletic trees under alternative HIV-1 transmission hypotheses using approximate bayesian computation. *Genetics*, 2017.

[134] Ethan O. Romero-Severson, Ingo Bulla, and Thomas Leitner. Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences*, 2016.

[135] Debbie S Ruelas and Warner C Greene. Leading Edge Review An Integrated Overview of HIV-1 Latency. *Cell*, 155, 2013.

[136] Christian T Ruff, Stuart C Ray, Patricia Kwon, Rebekah Zinn, Amanda Pendleton, Nancy Hutton, Roxann Ashworth, Stephen Gange, Thomas C Quinn, Robert F Siliciano, and Deborah Persaud. Persistence of Wild-Type Virus and Lack of Temporal Structure in the Latent Reservoir for Human Immunodeficiency Virus Type 1 in Pediatric Patients with Extensive Antiretroviral Exposure. *Journal of Virology*, 76(18):9481–9492, 2002.

[137] Colin A. Russell and Menno D. de Jong. Infectious disease management must be evolutionary. *Nature Ecology & Evolution*, 1(8):1053–1055, aug 2017.

[138] M. J. Sackin. "Good"and "Bad" Phenograms. *Systematic Biology*, 21(2):225–226, 7 1972.

[139] Keri B. Sanborn, Mohan Somasundaran, Katherine Luzuriaga, and Thomas Leitner. Recombination elevates the effective evolutionary rate and facilitates the establishment of HIV-1 infection in infants after mother-to-child transmission. *Retrovirology*, 2015.

[140] Rafael Sanjuán. Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1548):1975–82, jun 2010.

[141] Mauricio Santillana, André T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology*, 11(10):e1004513, oct 2015.

[142] Samuel V Scarpino, Atila Iamarino, Chad Wells, Dan Yamin, Martial Ndeffo-Mbah, Natasha S Wenzel, Spencer J Fox, Tolbert Nyenswah, Frederick L Altice, Alison P Galvani, Lauren Ancel Meyers, and Jeffrey P Townsend. Epidemiological and viral genomic sequence analysis of the

2014 ebola outbreak reveals clustered transmission. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 60(7):1079–82, 4 2015.

[143] Gergely J Schiffels, Stephan and Szöll} Osi, Ville Mustonen, and Michael Lässig. Emergent Neutrality in Adaptive Asexual Evolution. *Genetics*, 189:1361–1375, 2011.

[144] Jeffrey Shaman and Melvin Kohn. Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academies of Science*, 106(9):3243–3248, 2009.

[145] Raj Shankarappa, Joseph B Margolick, Stephen J Gange, Allen G Rodrigo, David Upchurch, Homayoon Farzadegan, Phalguni Gupta, Charles R Rinaldo, Gerald H Learn, X I He, Xiao-Li Huang, and James I Mullins. Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection Downloaded from. *Journal of Virology*, 73(12):10489–10502, 1999.

[146] Yuelong Shu and John McCauley. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13):30494, mar 2017.

[147] Tom J Sidwa. Mosquito Surveillance/Control in Texas. Technical report, Texas Department of State Health Services, 2016.

[148] Janet D. Siliciano, Joleen Kajdas, Diana Finzi, Thomas C. Quinn, Karen Chadwick, Joseph B. Margolick, Colin Kovacs, Stephen J. Gange, and Robert F. Siliciano. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+T cells. *Nature Medicine*, 2003.

[149] Henrik Singmann, Ben Bolker, Jake Westfall, and Frederik Aust. afex: Analysis of Factorial Experiments, 2018.

[150] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, jul 2009.

[151] Hongshuo Song, Elena E. Giorgi, Vitaly V. Ganusov, Fangping Cai, Gayathri Athreya, Hyejin Yoon, Oana Carja, Bhavna Hora, Peter Hraber, Ethan Romero-Severson, Chunlai Jiang, Xiaojun Li, Shuyi Wang, Hui Li, Jesus F. Salazar-Gonzalez, Maria G. Salazar, Nilu Goonetilleke, Brandon F. Keele, David C. Montefiori, Myron S. Cohen, George M. Shaw, Beatrice H. Hahn, Andrew J. McMichael, Barton F. Haynes, Bette Korber, Tanmoy Bhattacharya, and Feng Gao. Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature Communications*, 9(1):1928, dec 2018.

[152] J Conrad Stack, J David Welch, Matt J Ferrari, Beth U Shapiro, and Bryan T Grenfell. Protocols for sampling viral sequences to study epi-

demic dynamics. *Journal of the Royal Society, Interface / the Royal Society*, 7(48):1119–27, jul 2010.

[153] L. Steinbrück, T. R. Klingen, and A. C. McHardy. Computational Prediction of Vaccine Strains for Human Influenza A (H3N2) Viruses. *Journal of Virology*, 88(20):12123–12132, oct 2014.

[154] Natalja Strelkowa and Michael Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–82, oct 2012.

[155] Rahul Subramanian, Andrea L. Graham, Bryan T. Grenfell, and Nimalan Arinaminpathy. Universal or Specific? A Modeling-Based Comparison of Broad-Spectrum Influenza Vaccines against Conventional, Strain-Matched Vaccines. *PLOS Computational Biology*, 12(12):e1005204, dec 2016.

[156] James D. Tamerius, Jeffrey Shaman, Wladmir J. Alonso, Kimberly Bloom-Feshbach, Christopher K. Uejio, Andrew Comrie, and Cécile Viboud. Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. *PLoS Pathogens*, 9(3):e1003194, mar 2013.

[157] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10(3):512–526, may 1993.

[158] R Core Team. R: A Language and Environment for Statistical Computing, 2016.

[159] Texas Department of State Health and Human Services. Arbovirus Activity in Texas 2013 Surveillance Report, 2014.

[160] Texas Department of State Health and Human Services. Arbovirus Activity in Texas 2014 Surveillance Report, 2014.

[161] Texas Department of State Health and Human Services. Texas Announces Local Zika Virus Case in Rio Grande Valley, 2016.

[162] "Texas Department of State Health and Human Services". Zika in Texas, 2016.

[163] Xiping Wei, Julie M. Decker, Shuyi Wang, Huxiong Hui, John C. Kappes, Xiaoyun Wu, Jesus F. Salazar-Gonzalez, Maria G. Salazar, J. Michael Kilby, Michael S. Saag, Natalia L. Komarova, Martin A. Nowak, Beatrice H. Hahn, Peter D. Kwong, and George M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312, mar 2003.

[164] Frank Wen, Trevor Bedford, and Sarah Cobey. Explaining the geographical origins of seasonal influenza A (H3N2). *Proceedings of the Royal Society of London B: Biological Sciences*, 283(1838), 2016.

[165] Frank T. Wen, Sidney M. Bell, Trevor Bedford, and Sarah Cobey. Estimating vaccine-driven selection in seasonal influenza. *Viruses*, 10(9):1–37, 2018.

[166] James B. Whitney, Alison L. Hill, Srisowmya Sanisetty, Pablo Penaloza-MacMaster, Jinyan Liu, Mayuri Shetty, Lily Parenteau, Crystal Cabral, Jennifer Shields, Stephen Blackmore, Jeffrey Y. Smith, Amanda L. Brinkman, Lauren E. Peter, Sheeba I. Mathew, Kaitlin M. Smith, Erica N. Borducchi, Daniel I. S. Rosenbloom, Mark G. Lewis, Jillian Hattersley, Bei Li, Joseph Hesselgesser, Romas Geleziunas, Merlin L. Robb, Jerome H. Kim, Nelson L. Michael, and Dan H. Barouch. Rapid seeding of the viral reservoir prior to SIV viraemia in rhesus monkeys. *Nature*, 512(7512):74–77, aug 2014.

[167] Laurence A. Wolsey. *Integer programming.* Wiley New York, New York, vol. 42. edition, 1998.

[168] Joseph K Wong, Marjan Hezareh, Huldrych F Günthard, Diane V Havlir, Caroline C Ignacio, Celsa A Spina, and Douglas D Richman. Recovery of Replication-Competent HIV Despite Prolonged Suppression of Plasma Viremia. *Science*, 278(5341):1291–1295, 1997.

[169] Michael Worobey, Marlea Gemmel, Dirk E. Teuwen, Tamara Haselkorn, Kevin Kunstman, Michael Bunce, Jean-Jacques Muyembe, Jean-Marie M. Kabongo, Raphaël M. Kalengayi, Eric Van Marck, M. Thomas P. Gilbert, and Steven M. Wolinsky. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661–664, oct 2008.

[170] Qian Zhang, Kaiyuan Sun, Matteo Chinazzi, Ana Pastore-Piontti, Natalie E Dean, Diana P Rojas, Stefano Merler, Dina Mistry, Piero Poletti,

Luca Rossi, Margaret Bray, M. Elizabeth Halloran, Ira M Longini, and Alessandro Vespignani. Projected spread of Zika virus in the Americas. *bioRxiv*, 2016.

[171] Jianling Zhuang, Amanda E Jetzt, Guoli Sun, Hong Yu, George Klarmann, Yacov Ron, Bradley D Preston, and Joseph P Dougherty. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *Journal of virology*, 76(22):11273–82, nov 2002.

Vita

Lauren Ann Castro grew up in Los Alamos, New Mexico, where exposure to the thriving scientific community surrounding the national laboratory inspired her to pursue a career in science. After graduating from Los Alamos High School in 2009, she left the mountains of New Mexico for Princeton, New Jersey. In 2013 she received a Bachelor of Arts with honors degree in Ecology and Evolutionary Biology from Princeton University. Following graduation, she worked at Los Alamos National Laboratory as a Post-Baccalaureate Student in the Defense Systems and Analysis Division. In 2014 she enrolled in the Ecology, Evolution, and Behavior doctoral graduate program at The University of Texas at Austin.

Permanent address: 2823 East Martin Luther King Jr. Blvd
Austin, Texas 78702

This dissertation was typeset with LaTeX[†] by Lauren A Castro.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.