Learning RNA Viral Disease Dynamics from Molecular Sequence Data

by

Matteo Vaiente

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2020 by the
Graduate Supervisory Committee:

Matthew Scotch, Chair
Anuj Mubayi
Li Liu

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

The severity of the health and economic devastation resulting from outbreaks of viruses such as Zika, Ebola, SARS-CoV-1 and, most recently, SARS-CoV-2 underscores the need for tools which aim to delineate critical disease dynamical features underlying observed patterns of infectious disease spread. The growing emphasis placed on genome sequencing to support pathogen outbreak response highlights the need to adapt traditional epidemiological metrics to leverage this increasingly rich data stream. Further, the rapidity with which pathogen molecular sequence data is now generated, coupled with advent of sophisticated, Bayesian statistical techniques for pathogen molecular sequence analysis, creates an unprecedented opportunity to disrupt and innovate public health surveillance using 21st century tools. Bayesian phylogeography is a modeling framework which assumes discrete traits — such as age, location of sampling, or species — evolve according to a continuous-time Markov chain process along a phylogenetic tree topology inferred from molecular sequence data.

While myriad studies exist which reconstruct patterns of discrete trait evolution along an inferred phylogeny, attempts to translate the results of phylogeographic analyses into actionable metrics that can be used by public health agencies to direct the development of interventions aimed at reducing pathogen spread are conspicuously absent from the literature. In this dissertation, I focus on developing an intuitive metric, the phylogenetic risk ratio (PRR), which I use to translate the results of Bayesian phylogeographic modeling studies into a form actionable by public health agencies. I apply the PRR to two case studies: i) age-associated diffusion of influenza A/H3N2 during the 2016-17 US epidemic and ii) host associated diffusion of West Nile virus in the US. I discuss the limitations of this (and Bayesian phylogeographic) approaches when studying non-geographic traits for which limited metadata is available in public molecular sequence databases and statistically principled approaches to this missing metadata problem. Then, I perform a simulation study to evaluate the statistical performance of the missing metadata solution. Finally, I provide a solution for researchers interested in using the PRR and phylogenetic UTMs in their own genomic epidemiological studies yet are deterred by the idiosyncratic, error-prone processes required to implement these models using popular Bayesian phylogenetic inference software packages. My solution, Build-A-BEAST, is a publicly available, object-oriented system written in python which aims to reduce the complexity and idiosyncrasy of creating XML files necessary to perform the aforementioned analyses. This dissertation extends the conceptual framework of Bayesian phylogeographic methods, develops a summary statistic for translating the output of these models into an actionable form, and evaluates the use of priors for missing metadata. In doing so, I lay the foundation for future work in disseminating and implementing Bayesian phylogeographic methods for routine public health surveillance.

DEDICATION

“Dreams turn too easily into disillusionment without the power of truth and grace behind them” - Dr. William L. Herzfeld

This work is dedicated to my family. To my grandfather, Dr. William L. Herzfeld, whose untimely passing ignited my passion for infectious disease epidemiology. To my grandmother, Thressa Alston, who taught me the peace afforded by serving a purpose greater than one’s self. To my aunt, Dr. Ciby Kimbrough, who instilled in me a love of learning (and that chocolate is indeed, a vegetable). To my mother, Katherine Vaiente, who taught me the power which resides in a well-penned sentence. To my father, Joe Vaiente, who taught me that character is built in the face of adversity. To my brother, Gianni Vaiente, for whom I strive to be a role model. To my partner, Max Epstein-Lee, for being an unending source of encouragement and support. You are the source of my strength and inspiration.

ACKNOWLEDGMENTS

I would like to thank my Graduate Supervisory Committee, Drs. Matthew Scotch, Anuj Mubayi, and Li Liu for their guidance in the completion of my dissertation. I cannot express my gratitude for the many hours spent in meetings, reviewing written work, and providing mentorship critical to my development as a scientist. I would like to thank the members of the ZooPhy lab, past and present, who provided companionship and were always willing to discuss research ideas; Arjun Magge, Demetrius Shargani, Daniel Magee, Davy Weissenbacher, Tasnia Tahsin and Rachel Beard. Similarly, I extend my gratitude to members of the Center for Environmental Health Engineering for their collegiality, insightful feedback, and for their patience (especially when discussing Bayesian statistical methods). I would also like to thank individuals who provided academic and administrative support to me, including Drs. Dana Epstein, Rolf Halden, Anita Murcko, Rob McCulloch and Richard Hahn; Marin Hanlin, Lauren Madjidi, Marcia Spurlock and Rachel LaVine. I thank the various agencies and programs that have provided funding and other opportunities which supported the completion of this work and enriched my doctoral experience: EcoHealth Alliance, the National Institutes of Health, the PLus Alliance, the ASU Graduate College and the Dissemination and Implemetation affinity network of the College of Health Solutions. Finally, I would like to thank those individuals who gave me an opportunity to become involved in research; Stephanie Adamson, Drs. Anne Wertheimer, Kelly Reynolds, Donna Wolk and Gayatri Vedantam.

TABLE OF CONTENTS

Page

CHAPTER

CHAPTER Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

VIRUS SEQUENCES IMPLICATE IMMUNE IMPRINTING AS A KEY DRIVER OF AGE ASSOCIATED DIFFUSION DURING THE 2016-17 US INFLUENZA H3N2 EPIDEMIC

Author Summary

We employ Bayesian phylogenetic generalized linear models to identify the drivers of age associated H3N2 influenza diffusion during the 2016-2017 US epidemic across each of the ten Health and Human Services regions. We show that age associated diffusion for the 2016-2017 H3N2 epidemic was primarily driven by immunologic imprinting and other age-based effects. However, our results show that this is not true across all Health and Human services regions in the US. We use “phylogenetic relative risks” to demonstrate regional heterogeneity in age associated transmission risks, suggesting that strategies for controlling influenza spread should be tailored to specific regional risk patterns. To our knowledge, our study is the first to utilize genomic epidemiology to analyze the factors underlying age associated influenza H3N2 diffusion during annual outbreaks.

1.1 Abstract

The growing emphasis placed on pathogen genome sequencing to support outbreak response highlights the need to adapt traditional epidemiological metrics to leverage this increasingly rich data stream. In this paper, we employ Bayesian phylogenetic generalized linear models (GLM) to identify the drivers of age associated H3N2 influenza diffusion during the 2016-2017 US epidemic across each of the ten Health and Human Services (HHS) regions. We show that age associated diffusion for the 2016-2017 H3N2 epidemic was primarily driven by immunologic imprinting and other age-based effects. However, our results show that this is not true across all regions. We validate our phylogenetic results by converting our sequence records to an equivalent line list representation and fitting a multinomial Bayesian GLM utilizing the same set of immunologic, demographic, social contact and age-based predictors. We then use “phylogenetic relative risks” which further demonstrate regional heterogeneity in age associated transmission risks, suggesting that strategies for controlling influenza spread should be tailored to specific regional risk patterns. To our knowledge, our study is the first to utilize genomic epidemiology to shed light on the factors underlying age associated influenza H3N2 diffusion during annual outbreaks. Future work should aim to quantify the effectiveness of genomic epidemiology informed control programs in real-world contexts to facilitate translation of these methods into routine public health practice.

1.2 Introduction

During the 2016-2017 season it is estimated that influenza killed approximately 51,000 people in the US (CDC 2019) alone. Putri et al (Putri et al. 2018) recently estimated influenza direct health care costs to be approximately $11.2 billion dollars per year, which further underscores the magnitude of these epidemics. Given recent global pandemics and increasing seasonal severity of influenza outbreaks, there is a growing need for targeted control strategies to curb the spread of these epidemics. Often, epidemiologists would benefit by knowing "whom acquires infection from whom". In terms of transmission dynamics, what are the sources, and complementary sinks, of an outbreak. These may be geographic locations, consumer products, or can be based on host characteristics; such as age, species or vaccination status. An understanding of such transmission dynamics among various discrete groups provides evidence necessary for designing and implementing effective disease control strategies. For example, during influenza outbreaks, identifying key age groups driving transmission would provide public health agencies with actionable information necessary to inform targeted control measures.

Recent work in quantifying the importance of various age groups during influenza (or more broadly, respiratory viral) outbreaks can be roughly categorized into two groups: methods based on analysis of surveillance count data (Schanzer, Vachon, and Pelletier 2011; Worby et al. 2015; Goldstein et al. 2017; Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019; Ranjeva et al. 2019; Arevalo et al. 2019) and methods based on compartmental mathematical models (Wallinga, Teunis, and Kretzschmar 2006; Basta et al. 2009; Apolloni, Poletto, and Colizza 2013; De Luca et al. 2018). While informative, these approaches either obfuscate details of or rely on assumptions about transmission processes which ultimately give rise to observed epidemic patterns. For example, statistical models often rely on observing the magnitude of events in a particular group without distinguishing the *origin* of those events. Alternatively, compartmental models often solely rely on assumptions about between group contact parameters that are difficult to obtain directly in the population of interest (Mossong et al. 2008a; Prem, Cook, and Jit 2017; Arregui et al. 2018). Inference of epidemiological patterns via virus sequences complements these approaches by offering a rich analytical framework with which one can integrate sequence data with travel, economic, social and other predictor data (Lemey et al. 2014; Dudas et al. 2017a; Grubaugh et al. 2017) while simultaneously reconstructing the relatedness of observed cases via phylogeny: an approximation of 'whom acquired infection from whom'.

The frequency with which molecular data is now generated motivates the use of virus sequences to shed light on transmission processes and patterns underlying disease outbreaks. For the last decade researchers have used the nascent field of Bayesian phylogeography (Lemey et al. 2009) to study the epidemiology of viral disease outbreaks including Ebola (Dudas et al. 2017a), Mers-CoV (Dudas et al. 2018), and Zika (Grubaugh et al. 2017) viruses. These methods combine data on discrete characteristics, such as geographic location or host species, with molecular sequence data, to infer historical characteristics of infectious disease outbreaks. The advent of the phylogeographic generalized linear model (GLM) (Lemey et al. 2014) extends this approach by parameterizing the rates of pairwise discrete trait change as a log-linear combination of predictors of interest, permitting statistical assessment of predictor support for reconstructing patterns of trait change across a phylogeny (Magee, Suchard, and Scotch 2017; Dudas et al. 2017a). Though discrete traits are often taken to be geographic location of the virus, analyses using host species as a discrete trait for pathogens such as SARS-CoV-2 (Fauver et al. 2020), WNV (Swetnam et al. 2018), HIV (Oster et al. 2015), MERS-CoV (Dudas et al. 2018), and avian influenza (Trovão et al. 2015; Bahl et al. 2016) have revealed rich between-host transmission dynamics underlying these respective outbreaks. By reconciling these two approaches, phylogeographic models can be used to assess the role of specific host groups in propagating new infections while simultaneously evaluating statistical support for host-associated factors driving underlying differences in reconstructed transmission risks.

In this paper, we aim to elucidate the contributions of immunologic, social contact, demographic and age-based factors to age associated influenza H3N2 diffusion within a phylogenetic framework. We synthesize available evidence on social mixing patterns (Mossong et al. 2008a; Prem, Cook, and Jit 2017; Arregui et al. 2018) and immunological imprinting (Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019; Arevalo et al. 2019), while additionally accounting for demographic structure and age-based factors, together with virus sequences to reveal the mechanisms driving age associated influenza spread. In the case of identifying factors driving observed influenza case age distributions Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019; Arevalo et al. 2019; Ranjeva et al. 2019, we believe that signals of host immune pressures exerted on influenza sequences, which are shaped by early life influenza exposures, should be detectable within phylogenetic models. We validate our phylogenetic models by fitting an analogous GLM to line list data derived from the same set of virus sequence records. To do so, we emulate previous models (Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019; Arevalo et al. 2019) and perform parameter inference using the Bayesian paradigm. Finally, we translate a traditional epidemiological association measure into a phylogenetic context: the “phylogenetic relative risk” and apply it to identify age group(s) driving national and regional transmission during the 2016-17 US influenza A/H3N2 epidemic. Reconstruction of the magnitude of transmission events between specific groups during phylogenetic analysis permits computation of the relative risk for specific between group transmission events using standard formulas. Our goal in framing a common measure of epidemiological association in a phylogenetic context is to advocate for wider use of phylogenetic methods in public health risk assessment during viral infectious disease outbreaks. We believe our study takes a necessary step in bridging the gap between genomic epidemiological research methods and public health practice by synthesizing classical and genomic epidemiological methods and applying these principles to the study of age-associated influenza diffusion.

1.3 Methods

1.3.1 Sequence Data Collection and Evolutionary Model Selection

We downloaded all available influenza A/H3N2 hemagglutinin (HA) sequence data from GenBank (Dennis A. Benson et al. 2012) and discarded isolates associated with non-human hosts. We applied the following inclusion criteria to generate the final analysis data set: a) sequences were from isolates with reported sampling dates between January 1, 2016 and December 31, 2017; b) sequences had a known sampling location (to state level); c) sequences had a host age reported in their GenBank record and d) sequences represented unique influenza H3N2 isolates. The final data set included 2812 unique HA sequences annotated with sampling time, sampling location and host age information. We summarize the results of our data collection method in

Figure 1. In Table S1, we provide sequence data, including sampling dates, locations and GenBank accessions. We linked each virus to its respective Health and Human Services (HHS) region, resulting in ten total analyses. We aligned sequences using MAFFT v. 7.407 (Katoh et al. 2002) with the default settings and inspected the results manually in Seqotron (Fourment and Holmes 2016). We used jModelTest2 (Darriba et al. 2012) to perform nucleotide substitution model selection prior to Bayesian analysis. We compared the fit of the 11 named substitution models by using the Bayesian Information Criterion as the objection function and found the GTR + Γ model to be most appropriate for our data. We used the maximum likelihood (ML) phylogenies generated during the jModelTest2 analyses to test the fit of strict vs relaxed molecular clocks. We identified the most appropriate molecular clock model by calculating likelihood scores of the ML phylogeny under both strict and relaxed molecular clock models and performing a likelihood ratio test (LR test).

1.3.2 Bayesian Phylogenetic Analysis

We modeled molecular evolution using a GTR + Γ nucleotide substitution model with 4 rate categories and uncorrelated log-normal molecular clock (Drummond et al. 2006). We specified the non-parametric generalized Bayesian Skyride (Minin, Bloomquist, and Suchard 2008) model as a coalescent tree prior for our analysis to account for the *a priori* expectation that population of infected individuals giving rise to influenza phylogenies follow non-linear dynamics. To estimate divergence times, we fixed tip dates as the dates of sampling reported in GenBank for each sequence, respectively. We used Bayesian MCMC to perform inference for our phylogeographic model as implemented in BEAST v1.10 (Suchard et al. 2018). For each HHS region, we ran the MCMC for 200 million iterations, sampling every 20,000 steps and removed the first 20% as burn-in. We diagnosed convergence of the MCMC procedure using Tracer v1.7.1 (Rambaut et al. 2018a) checking that all model parameters had Effective Sample Sizes (ESS) of 200 or greater.

1.3.3 Measuring Phylogeny-Age Association

The relationship between a discrete state of interest (e.g. location, infected host, clinical endpoint, etc.) and the genetic evolution of virus can be a powerful analytical tool in genomic epidemiology. If closely related sequences tend to share discrete traits more often than would be expected by chance (i.e. that is under tip label permutations), then, the evolution of these traits are tightly coupled to the phylogenetic tree topology. Specifically, the null hypothesis is that discrete traits are uncorrelated with genetic distances among individuals (Parker, Rambaut, and Pybus 2008). We test this null hypothesis using BaTS software (Parker, Rambaut, and Pybus 2008), and focus on two metrics, the Parsimony Score (PS) and Association Index (AI) for which null distributions are obtained via trait label permutations. We performed BaTS analysis using the last 1,000 posterior trees from our Bayesian phylogenetic analysis. We calculated null distributions for PS and AI using 100 trait permutations on each tree topology.

1.3.4 Identifying Social, Demographic and Immunological Correlates of Influenza Diffusion via Virus Sequences

Since ecological and evolutionary processes occur on the same time scale for RNA viral pathogens (Drummond et al. 2003), we can use molecular phylogenies to infer epidemiological processes giving rise to the estimated trees (Holmes and Grenfell 2009). Common applications of the technique are to reconstruct the geographic migration (Lemey et al. 2009; Lemey et al. 2014; Magee, Suchard, and Scotch 2017; Dudas et al. 2017a; Grubaugh et al. 2017) and host transmission (Trovão et al. 2015; Bahl et al. 2016; Swetnam et al. 2018; Dudas et al. 2018) histories of RNA viral disease outbreaks; processes which occurs on faster timescales than viral molecular evolution. Here, we propose treating host age as a discrete trait upon which phylogeographic inference is performed. Since geographic migration and host transition processes to take place on similar timescales for RNA viruses, we expect this to be a valid inferential target. For example, imagine the transmission of influenza on an airplane where migration and host diffusion processes occur on the same timescale. Here, there is influenza transmission between individuals of several age classes while these viral lineages are, simultaneously, being moved between geographic locations. A popular approach for identifying factors that are associated with phylogenetic inference of discrete trait diffusion involves parameterizing the rate parameters r_{ij} of the Markov matrix (Λ_{ij}) describing the rate of character change between each pair states i and j as a log-linear function of specific covariates of interest (Lemey et al. 2014). Employing this approach we parameterize the rate parameters describing transmission between age groups as a log linear function of social contact, demographic, and immunologic factors. Particularly, we emulate the examples of (Katelyn M Gostic et al. 2016;

Ranjeva et al. 2019; Arevalo et al. 2019) which have suggested that these factors drive age-specific dynamics observed during influenza epidemics. We incorporated social contact data for the US (Prem, Cook, and Jit 2017) by adjusting the US wide contact matrix by region specific population densities using the method of (Arregui et al. 2018). Similar to (Ranjeva et al. 2019; Arevalo et al. 2019) we include the population proportion of each age group, per HHS region, using population estimates from the US Census Bureau for 2017. Finally, we incorporate the probability of immune imprinting to H3-type hemagglutinin via the methods presented in (Arevalo et al. 2019). Briefly, these models assume that the attack rate on naive: unexposed individuals is approximately 28%. Then, the probability of infection by a specific age is assumed to be geometrically distributed (Arevalo et al. 2019; Katelyn M. Gostic et al. 2019). These age of first infection probabilities are then scaled by seasonal influenza intensities and the percent of specimens testing positive for each season from 2017 to 1918. We take the average probability of first infection for birth years in each of the aforementioned age groups our cohorts We additionally include an indicator variable in our models to describe increases in transmission risk for toddlers and adults aged 65+ years. We discuss details of the model in the Supplementary Material.

1.3.5 Discrete Trait Analysis and Phylogenetic Relative Risk

We used Markov jumps (MJ) to measure the relative magnitude of transmissions between age groups within each of the ten HHS region. Briefly, these counting processes enumerate labeled character changes when using continuous-time Markov chains to model character evolution along a phylogeny (Minin and Suchard 2008). We defined and annotated sequences with sixteen age categories for our discrete trait analysis. We defined non-overlapping 5 year age intervals from ages 0 to 75+ (ages 0-4, ages 5-9 ... ages 75+) as the groups for our analysis. This decision was driven by the resolution with which demographic data are typically recorded by the US Census for each state. Utilizing an asymmetric model of character change (Lemey et al. 2009) allows us to determine whether each age group acted as a source or sink of influenza transmissions within distinct regions during our study period. We manually edited BEAUti generated XML files to count all Markov jumps describing the relative magnitude of each specific character changes. We quantified the risk of each group to act as a source or sink by computing the relative risk of viral exchange between each group using Markov jumps following the example of (Bahl et al. 2016). We obtained total MJ counts for the last 2000 posterior trees of the Bayesian phylogenetic analysis in order to incorporate phylogenetic uncertainty into the analysis. We then created 2×2 contingency tables for each pairwise combination of age groups within a given HHS region and calculated relative risks as shown in Figure 2. Similar to the classical statistic, phylogenetic relative risk values below one indicate the group acts as a sink and tends to not transmit to its partner group. Alternatively, relative risk values above one indicate a group acts as a source and tends to transmit infection to a group. We performed this analysis for each of the ten HHS regions separately and then combined the results of each individual region to perform our national scale analysis.

1.3.6 Identifying Correlates of Age Associated Diffusion using Sequence Databases

In previous works (Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019) describing the role of immunologic imprinting in shaping influenza A case age distributions, correlates of age associated influenza risk were identified by fitting suites of multinomial generalized linear models (GLMs) to case count data binned by birth year. We emulate this approach by noting that molecular data reported in sequence databases may be converted to a line list representation given a set of sufficient metadata is present with the sequence record. Concretely, a virus sequence record containing both i) age of infected host and ii) date of isolate collection can be converted to an equivalent line list representation by calculating the birth year. We converted the 2812 sequence records meeting the aforementioned inclusion criteria to a line list representation. For the purposes of statistical analysis, we assumed sequenced isolates are a random sample of all isolates tested for influenza at public health laboratories. We take care to justify this assumption in the context of influenza surveillance in the US. Guidance on seasonal influenza surveillance is administered through the Association for Public Health Laboratories (APHL) via the "Flu Right Size Roadmap" (APHL 2013) which establishes several key components of a successful surveillance system. Specifically, it recommends that surveillance systems: i) establish a representative network of specimen submitters from ILINet providers, clinical and commercial labs, ii) utilize a statistical, systematic approach to collect and appropriate, adequate number of specimens for testing, iii) utilize sampling approaches that ensure submitted specimens are clinically, temporally, geographically and virologically representative of the population and iv) send representative clinical specimens and/or virus isolates to CDC or a CDC-designated laboratory for national surveillance purposes (APHL 2013). We expect that this serves to reduce judgement and convenience sampling which ultimately comprise the external validity the sampled isolates. In the absence of information on the specific sampling strategy employed by each submitting public health laboratory in the US, it is reasonable to assume that most follow a statistically principle sampling approach as prescribed above.

We fit our multinomial models using the Bayesian paradigm to all 2812 sequences available from the US to determine national level effects and drivers of case age distributions. Some advantages of using a Bayesian approach are parameter uncertainty is available in the form of posterior distributions and best-subset model selection is available using well-established techniques (George and McCulloch 1997). We discuss model fitting and selection specifics in the Supplementary Material. We utilized this analysis as a baseline to which we compare our phylogenetic risk models results.

1.4 Results

1.4.1 Phylogenetic Analysis within US HHS regions

The final data set included 2812 influenza A/H3N2 HA sequences annotated with state-level sampling location, isolation date and host age information. We provide a graphical summary of collected sequences in Figure 1. Sequence counts for each region ranged between 147 (HHS region 7) and 422 (HHS region 4) as we show in Figure 3. We observed large sequence counts for the 75+ age group in all regions. We additionally observe increased counts individuals aged 24 years and younger for some regions. Taking these distributions to be representative of the full case age distributions in each region, there is clearly heterogeneity among regions. In Figure 1, we show an overview of the sequence inclusion algorithm used to derive our analysis data set. We provide the accession numbers for the sequences included in this study in the Supplementary Material.

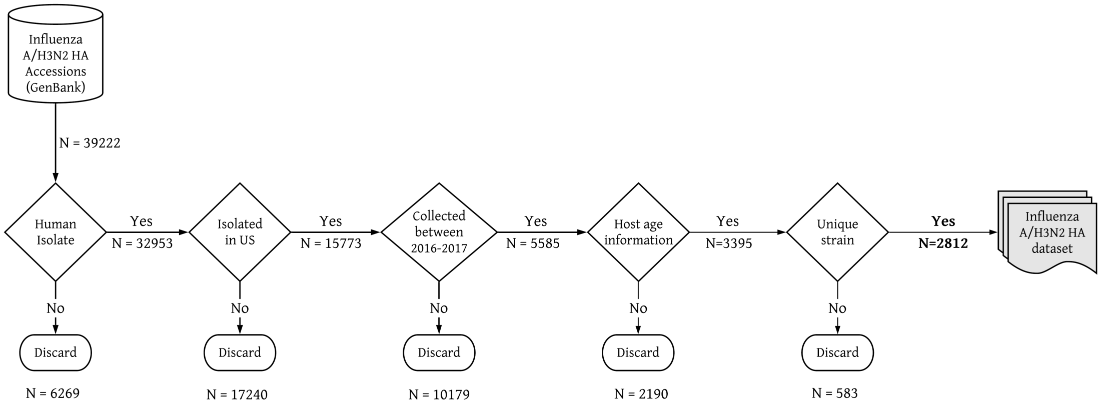
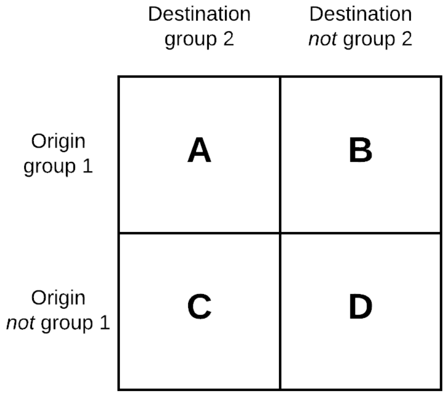


Figure 1. Overview of sequence inclusion criteria. Total number of sequences remaining after each step is indicated by the number below each arrow. The number of sequences discarded during each filtering step is shown below the corresponding discard node.

In Table 1, we provide the estimated mean substitution rate and 95% highest posterior densities (HPD) for the substitution rate and tree root age, stratified by HHS region. Our phylogenetic analyses estimated posterior hierarchichal mean substitution rates to be between 3.53×10^{-3} to 5.41×10^{-3} substitutions/site/year for all regions. These posterior means are commensurate with other studies of influenza A/H3N2 HA sequences (Bedford et al. 2010; Bahl et al. 2011). The estimated tree root ages ranged between late May 2015 (HHS region 2) and December 2015 (HHS region 10) across all regions.

1.4.2 Measuring the Strength of Phylogeny-Age Associations

We measured the strength of association between age groups and phylogenetic relatedness using PS and AI statistics as implemented in the program BaTS (Parker, Rambaut, and Pybus 2008) separately for each HHS region included in our study.



	Destination group 2	Destination *not* group 2
Origin group 1	**A**	**B**
Origin *not* group 1	**C**	**D**

Figure 2. In the classical epidemiological method, relative risk (RR) is computed as the ratio of the probability of an event of interest (disease progression) in one group versus the probability of the same event in another group. These groups are often defined by the presence/absence of an exposure of interest. The computation of relative risk from observational count data is then straightforward using the formula (A/A+C)/(B/B+D). In the phylogenetic context, we are similarly interested in estimating the relative probability of transmission from one group to another. We populate (for each pair of discrete traits in a model) a 2X2 contingency table. Cell A contains the number of Markov Jumps (MJ) from group 1 to group 2. Cell C contains the MJs from all other groups to group 2. The probability that a MJ to group 2 originates in group 1 is A/(A+C). We now need to compute the probability that an introduction to other regions (not including group 2) originates in group 1. We populate Cell B with the MJs originating in group 1 and ending in groups not including group 2. Cell D contains the remaining MJs between all other groups excluding groups 1 and 2. We compute the required probability as B/(B+D). The relative risk of an introduction from group 1 to group 2 is then the ratio of these probabilities, (A/A+C)/(B/B+D).

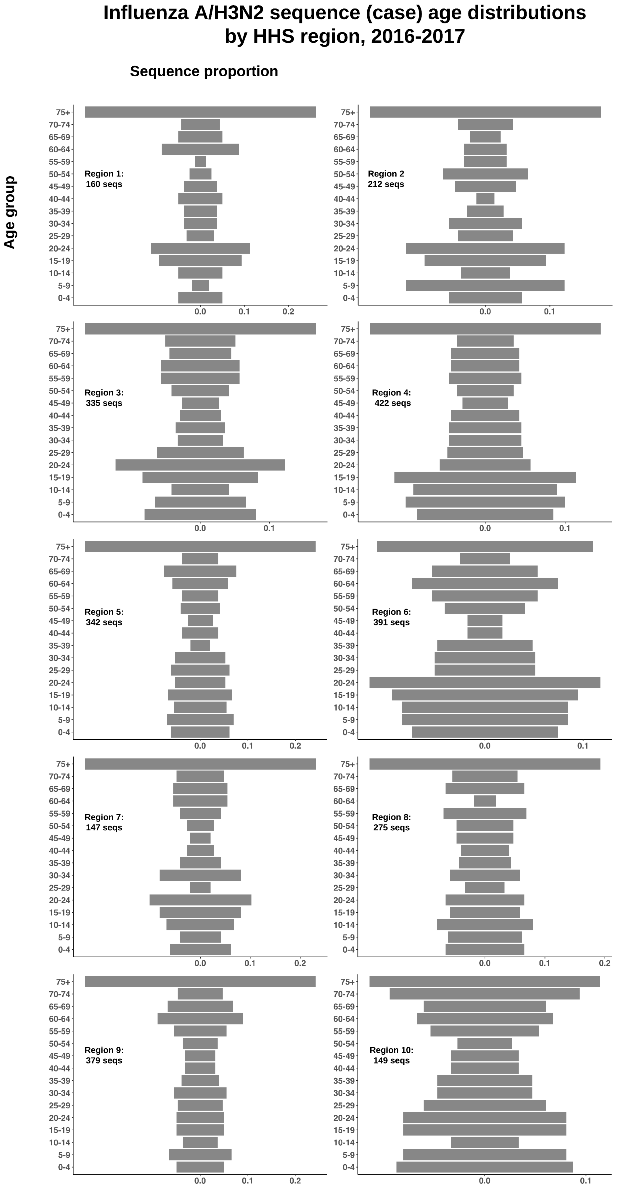


Figure 3. Age distribution of H3N2 sequences by HHS region. We observe a large proportion of cases in the 75+ year age group in all HHS regions. Some regions show a larger proportion of sequences among children ages 0-9 as well as young adults between ages 20-24. We assume that random isolates are selected for sequencing and reporting to GenBank and these distributions are, in part, reflective of underlying case age distributions in each respective HHS region.

	Posterior mean (95% HPD)		
	TMRCA	Subs. rate	MJs
Region 1	2015.6 (2015.2-2016.0)	4.58e-03 (3.32e-03-6.05e-03)	107 (103-112)
Region 2	2015.3 (2015.1-2015.6)	3.54e-03 (2.71e-03-4.39e-03)	135 (130-139)
Region 3	2015.6 (2015.5-2015.8)	4.11e-03 (3.40e-03-4.95e-03)	236 (247-260)
Region 4	2015.7 (2015.6-2015.9)	4.00e-03 (3.39e-03-4.64e-03)	315 (305-325)
Region 5	2015.7 (2015.5-2015.8)	4.92e-03 (4.06e-03-5.82e-03)	232 (226-238)
Region 6	2015.7 (2015.5-2015.9)	3.86e-03 (3.17e-03-4.68e-03)	303 (284-324)
Region 7	2015.7 (2015.3-2016.0)	4.11e-03 (3.04e-03-5.28e-03)	108 (105-114)
Region 8	2015.4 (2015.2-2015.7)	3.83e-03 (3.11e-03-4.51e-03)	205 (198-211)
Region 9	2015.5 (2015.4-2015.7)	4.56e-03 (3.77e-03-5.44e-03)	250 (242-258)
Region 10	2015.9 (2015.6-2016.2)	5.41e-03 (4.08e-03-6.71e-03)	134 (123-145)

Table 1. Posterior summary of Bayesian phylogenetic analysis. We show the mean and 95% highest posterior density regions (95% HPD) for the root age, substitution rate, and total Markov jump (MJ) transitions for each HHS region. Time to most recent common ancestor is denoted as TMRCA.

We present the mean, 95% HPD of the AI statistics for the observed tip distribution along with the mean and 95% HPD from the null distributions for each HHS region in Table 2. Our AI analysis results show that the pattern of age group changes is tightly correlated with genetic distance between isolates for each HHS regions except 3, 7 & 10 (Tables 2) We suspect that increased sampling of the 75+ age group (perhaps due to higher rates of medically attended ILI or case ascertainmentm) for regions 3 & 7 likely contributed to the lack of observable phylogeny-trait association signals in these regions. This would be due to smaller distances between the permutation (null) and observed distributions. For the PS statistics, we show the results in Table 3. We uncovered a similar pattern as that of the AI analysis, however, this analysis additionally suggests a lack of detectable phylogeny-trait correlation signal for region 5 & 8. These regions seem to similarly have over-representation of the 75+ age groups in their sequence distributions (Figure 4) and we suspect this leads invariant distributions for the test statistics, as mentioned above. Nonetheless, for the remaining regions, our results are interpreted as the phylogeny being informative about the pattern of evolution for a specific discrete trait; the pattern of age group changes. In our application, this suggests that genetically similar isolates are more likely to share an age group. Compared with our *a priori* expectation that most transmissions occur within (rather than between) age groups, driven by the assortativity observed in the inferred contact patterns (Mossong et al. 2008a; Prem, Cook, and Jit 2017), this seems quite reasonable. Motivated by these results, we continued the analysis of age associated diffusion via phylogeographic GLMs, noting reduced confidence in the results of these models for region 3, 7, & 10.

1.4.3 Determining Drivers of Influenza Case Age Distributions via Sequence Databases

Modeling the observed case counts in sequence databases offers a complementary approach to the analysis of line list data when the latter is unavailable in the population of interest. Given sufficient metadata is reported with the sequence record, it is straightforward to convert sequence data to a line list representation. We fit Bayesian multinomial GLMs to case counts binned by 5-year age intervals to mirror the granularity with which social contact and US population data were available. We show the fit of our final model to the case age distribution of all 2,812 records included in this study in Figure 4. Similar to previous studies, we find strong support for modest protection via hemagglutinin (HA) imprinting (0.44, 95% HPD 0.23-0.92, BF: 75.89, Table4) as a predictor of the case distribution among age groups. Additionally, we found that increased increased influenza risk for the elderly (ages 75+) was strongly

	Posterior mean (95% HPD)		
	AI (95% HPD)	null AI (95% HPD)	p-value
Region 1	**14.527 (13.482–15.459)**	**16.179 (15.187–17.029)**	$< 10^{-2}$
Region 2	**19.013 (17.676–20.313)**	**22.376 (21.486–23.153)**	$< 10^{-2}$
Region 3	34.918 (33.255–36.477)	35.395 (34.263–36.351)	0.21
Region 4	**41.220 (39.339–42.999)**	44.317 **(43.098–45.327)**	$< 10^{-2}$
Region 5	**34.162 (32.637–35.727)**	**35.481 (34.208–36.560)**	**0.04**
Region 6	**39.467 (37.711–41.165)**	**42.763 (41.586–43.948)**	$< 10^{-2}$
Region 7	14.879 (13.885–15.799)	15.569 (14.831–16.213)	0.1
Region 8	**27.852 (26.512–29.220)**	**29.026 (28.211–29.814)**	$< 10^{-2}$
Region 9	**36.275 (34.408–38.100)**	**38.029 (36.713–38.952)**	**0.01**
Region 10	15.637 (14.636–16.606)	15.372 (14.706–16.013)	0.77

Table 2. Posterior summary of BaTS analysis. We show the mean and 95% highest posterior density regions (95% HPD) for the Association Index (AI) for each HHS region, along with the mean and 95% HPD values from the null distributions. We find that AI scores are lower than would be expected by chance except in HHS regions 3, 7 & 10. Broadly, this indicates the presence of phylogeny-trait correlation and trait structure in the data. We suspect that increased sampling of the 75+ age group (due to higher rates of medically attended ILI) for regions 3 & 7 likely contributed to the lack of phylogeny-trait association signals in these regions. For HHS region 10, we believe particularly sparse sequence sampling (149 sequences) contributed to the observed lack of phylogeny-trait association. We display p-values < 0.05 in bold.

supported (5.47, 95% HPD 1.92-3.99, BF: 24999, Table 4). We observed reduced risks for cases between ages 0-4 with a strong protective effect inferred by our models (0.44, 95% HPD 0.02-0.69, BF: 6.90, Table 4). Similar to other H3N2 outbreak years, there was a larger proportion of cases in elderly birth cohorts (aged 75+ years) which is reflected in the empirical age distributions (Figure 3). As previously suggested, these patterns are congruent with the notion of birth year specific imprinting effects (Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019). Social contact showed limited posterior support for inclusion in our models (BF < 1, Table 4) which is expected since social contact rates are incorporated into the null model expectation

	Posterior mean (95% HPD)		
	PS (95% HPD)	null PS (95% HPD)	p-value
HHS region 1	**103.271 (101–105)**	**110.895 (107.282–114.206**)	$< 10^{-2}$
HHS region 2	**136.714 (134–139)**	**156.555 (152.878–161.405**)	$< 10^{-2}$
HHS region 3	246.786 (243–250)	251.606 (246.287–255.531)	0.09
HHS region 4	**304.473 (300–309)**	**320.486 (313.662–325.940)**	$< 10^{-2}$
HHS region 5	237.506 (234–241)	243.202 (238.002–247.193)	0.05
HHS region 6	**282.337 (278–287)**	**299.310 (292.610–304.612)**	$< 10^{-2}$
HHS region 7	105.215 (103–107)	105.709 (102.259–109.055)	0.35
HHS region 8	201.846 (199–205.)	205.091 (201.301–210.088)	0.08
HHS region 9	**255.902 (252–260)**	**269.758 (264.275–274.473)**	$< 10^{-2}$
HHS region 10	114.506 (112–117)	115.926 (111.511–119.898)	0.33

Table 3. Posterior summary of BaTS analysis. We show the mean and 95% highest posterior density regions (95% HPD) for the Parsimony Score (PS) for each HHS region, along with the mean and 95% HPD values from the null distributions. We find that the PS scores are lower than would be expected by chance in all regions except HHS regions 3, 5, 7, 8 &10 indicating the presence of phylogeny-trait correlation. Similar to the results reported for the AI statistic, dense sampling of the 75+ age group in HHS regions 3, 5, 7 & 8 likely contributed to the lack of observable signal. Sparse sequence sampling in HHS region 10 likely contributed to the lack of an observable phylogeny-trait correlation signal. We display p-values < 0.05 in bold.

(derived from the numerical solution of an age structured SIR model) as described in the Supplementary Material.

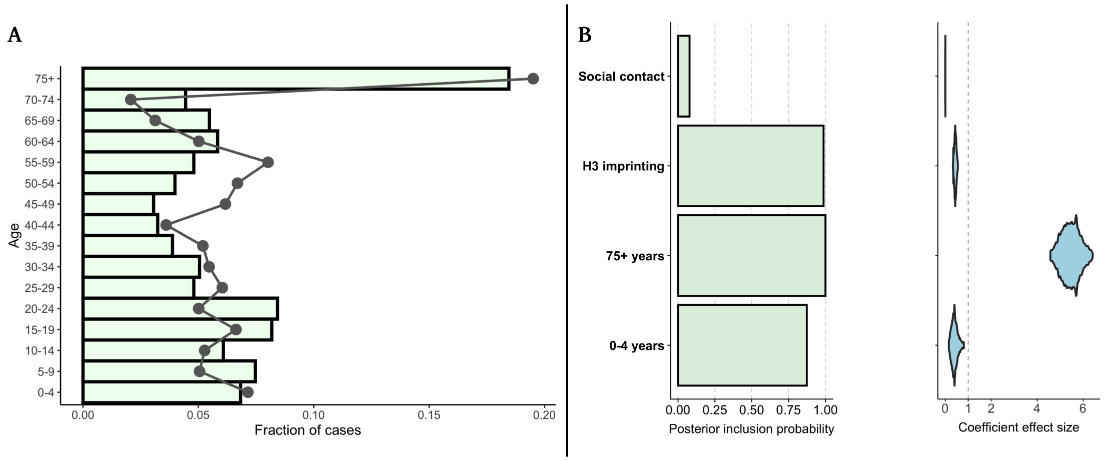


Figure 4. Observed and fitted influenza H3N2 case age distributions, US, 2016-17. We show the predicted age distribution using posterior mean parameter estimates over the 2^p possible models. These predictions are primarily driven by imprinting and age (elderly) risk parameters since other parameters receive little support for posterior inclusion.

	Posterior mean (95% HPD)		
Predictor	Effect size (β)	Inclusion prob.	Bayes' Factor
HA imprinting	0.44 (0.231-0.924)	0.986	75.89
0-4 age-based risk	0.42 (0.020-0.691)	0.87	6.90
75+ age-based risk	5.47 (1.92-3.99)	0.99	24999
Total daily contacts	1.7e^{-3} (3e-6-0.055)	0.078	8.48e^{-2}

Table 4. Posterior summary of Bayesian multinomial GLM analysis. We show the mean and 95% highest posterior density regions (95% HPD) for the coefficient effect sizes and posterior inclusion probability from BSSVS. We compute Bayes' factors for coefficient inclusion and see support for imprinting and age-based risk as strong determinants of sequence (case) age distributions.

1.4.4 Virus Sequences Strongly Support Hemagglutinin Imprinting as the Key Driver of Age Associated Influenza H3N2 Diffusion in the US

We examined the roles of social contact, immunologic, age and demographic factors on the rates of age associated diffusion (i.e. the rates of a Markov transition matrix) using a phylogeographic GLM, emulating the Bayesian multinomial model fit to the H3N2 sequence age distribution. However, since we are now modelling the between group transition rates directly, we used appropriate modifications to the included predictor variables as discussed in the supplementary material. Since we used CTMCs to model the transmission of viral lineages between age groups, positive coefficient values are associated with supportive effects such that increases in predictor vales increase transmission rates between age groups. Conversely, negative coefficient values are associated with protective effects such that increases in predictor values results in reduced transmission rates between groups. Our models show strong support for HA imprinting (in the group of origin, i.e. the transmitting group) in HHS regions 1, 3, 5, 8, 9 and 10 as demonstrated by high posterior inclusion probabilities in these models (Figure 11). Across these regions, we found a protective effect of homotypic immune imprinting on viral transmission rates as demonstrated by strictly negative 95% HPD regions for model coefficients. In HHS regions 2 and 7, we see a similar trend toward support for protective HA imprinting, however, the 95% HPD regions both cross zero and have relatively low posterior inclusion probabilities. The notion that HA imprinting reduces between group transmission rates is consistent with both total and partial protection hypotheses. In the case of the former, total protection conferred by imprinting from a matched HA type would eliminate transmissions from this group entirely since they would not transition to an infectious state. If HA imprinting were to confer partial protection such that case severity is reduced, it is easy to imagine that reductions in viral shedding rates, length of infectious period or myriad other mechanisms could lead to reduced transmission rates from these groups.

Across numerous regions (2-6 and 8, Figure 11) we found a supportive effect of population density on transmission rates, as demonstrated by positive coefficient values and high posterior inclusion probabilities. We observed similar trends in regions 1, 7, 9 and 10 though the 95% HPD regions for model coefficients often included zero, or, in the case of region 10, was associated with a small posterior inclusion probability. This could be due to several reasons: small sample sizes for these regions (Figure 3) could limit the power of the phylogenetic GLM to detect statistical associations, (Lemey et al. 2014; Magee and Scotch 2018) or, other predictors in these regions better explain the reconstructed transmission dynamics between groups. Notwithstanding potential limitations, a supportive effect of population density on groupwise transmission rates is consistent with predictions from structured compartmental models where the ratio of group populations appears in the next generation matrix which defines the growth of infectious individuals in each compartment (Driessche 2017). We similarly observed a positive, supportive effect of elderly age on between group transmission rates across regions 2-6 and 8. While this may partially be driven by the increased rates of medically attended respiratory illness, it is also possible that these groups transmit to other at increased rates. Lee et al. show that elderly influenza patients sustain prolonged viremia accompanied by viral shedding (Lee et al. 2009), which may increase the probability of transmission to an elderly individual's contacts.

Overall, we see broad posterior support for HA imprinting, population density and age-based risk in our phylogeographic models of age-assocaited influenza A/H3N2 diffusion. These results are consistent with associations identified during our standard Bayesian GLM analysis which used a line list representation of the sequence data. We believe the concordance between phylogenetic and standard GLM analysis results bolster the results obtained in each independent analysis and that the inferences drawn by the phylogenetic GLM are due to true signal in the data. Together our results indicate homotypic immune imprinting was a key driver of age related age diffusion at national and regional scales (Table 4). Given the confidence that our models are detecting true signal in the data, we proceed to analyze reconstructed patterns of age associated diffusion afforded by casting this problem in a phylogenetic context.

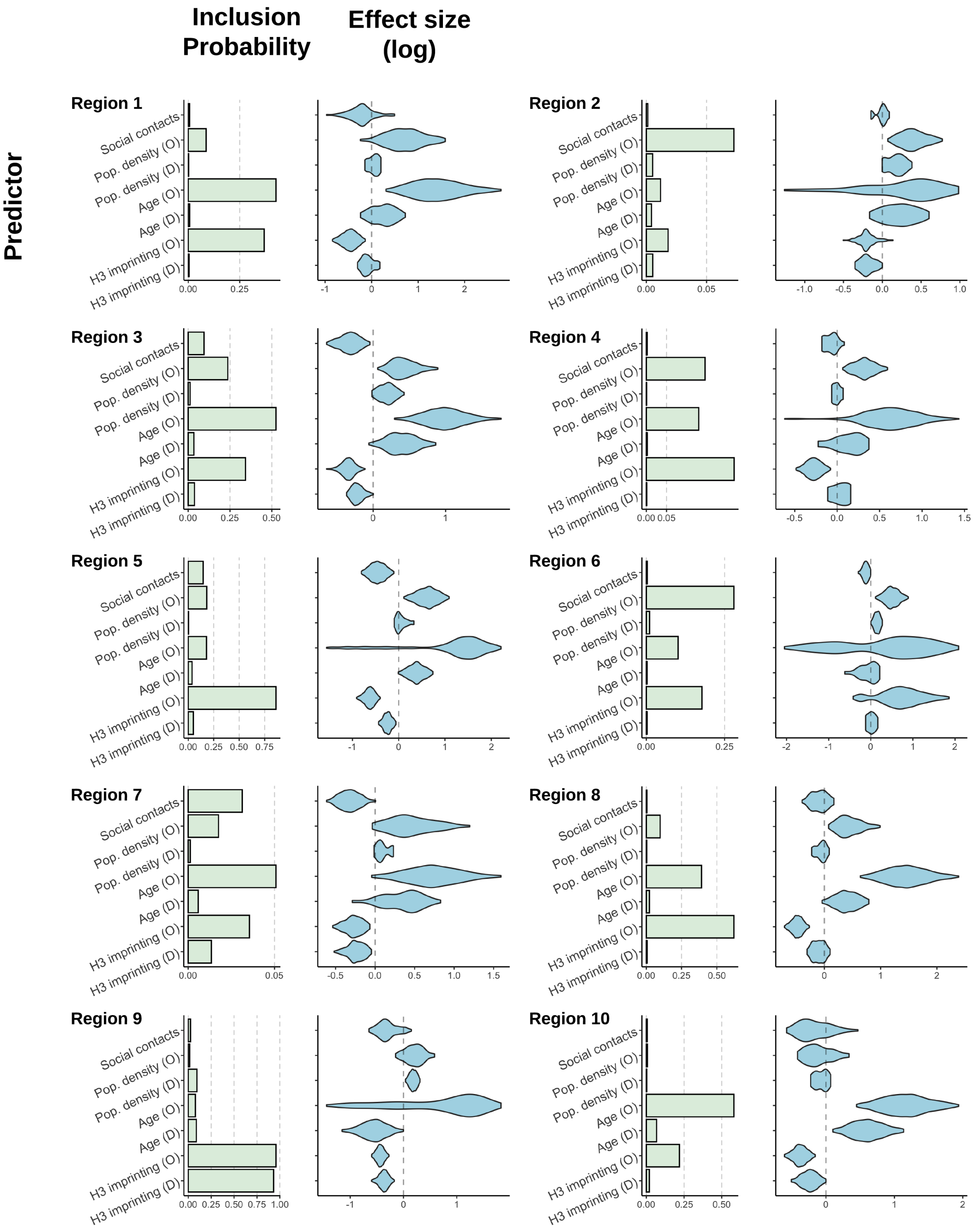


Figure 5. Phylogenetic GLM results for each of 10 HHS regions. We modeled the age associated diffusion of influenza as a log-linear combination of social contact, immunologic, age and demographic factors. We found strong support for hemagglutinin imprinting as the key driver in a majority of regions. We also found considerable support for age and population density predictors across several regions

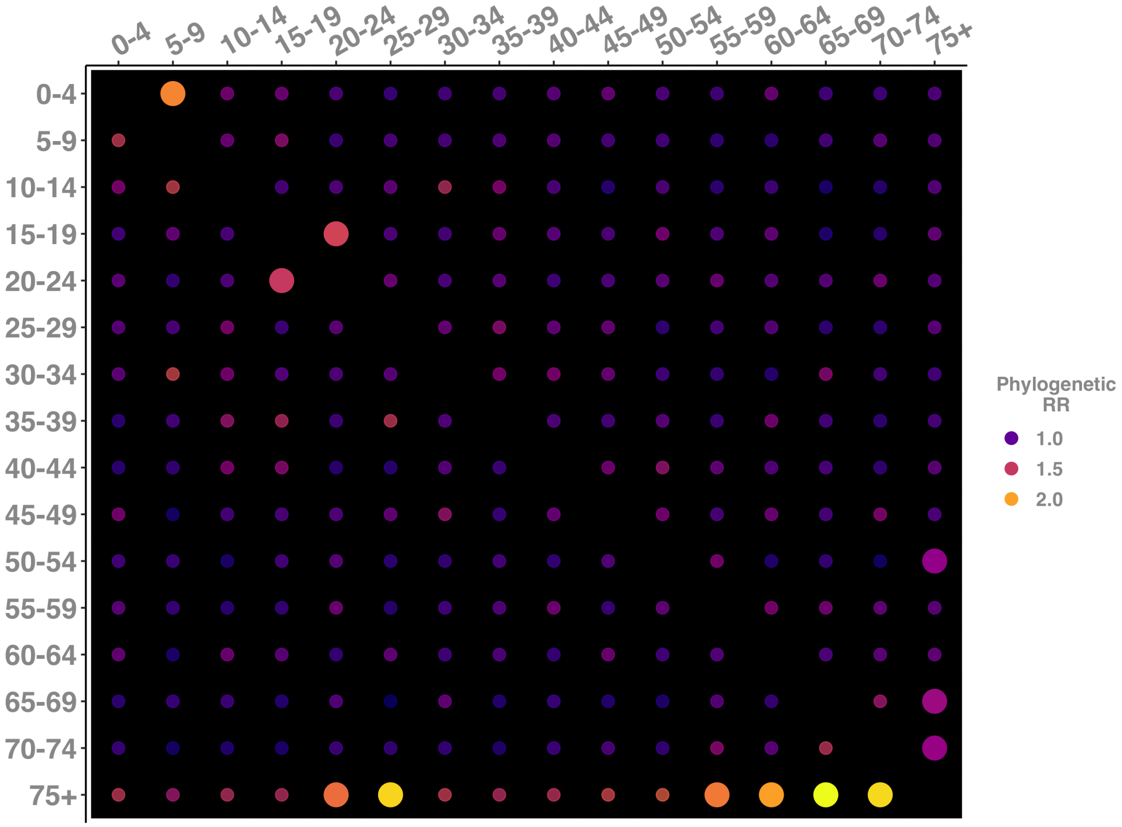


Figure 6. Visual summary of phylogenetic risk ratio results for US H3N2 influenza epidemic. We show the phylogenetic relative risk inferred from our GLM analyses. We found significant risks (indicated by larger, bold circles in the matrix) for seniors (75+ years) as the group of origin for individuals aged 20-29 and 55+ years, notably excluding adults ages 35-50. The lack of significant transmission to individuals aged 35-50 years may be driven by HA imprinting since during 1968-1977 H3N2 was the only circulating influenza strain. After this period, it remained the dominating circulating strain in years where these birth cohorts would be exposed to their first influenza infections.

1.4.5 Regional Heterogeneity in Age Structured Transmission Patterns

Given we used asymmetric Markov models to describe viral transmission among various age groups, we are able to quantify the relative source-sink dynamics between age groups, offering a complement to traditional, surveillance-based approaches for inferring important groups during epidemics. Using phylogenetic relative risks (Figure 2), we investigated age structured transmission patterns at the national and HHS regional levels. Nationally, we recovered patterns consistent with known influenza epidemiology. For example, we see an elevated risk of transmission between school aged children (ages 15-24), which reflects their high within group social contact rates relative to other age groups (Mossong et al. 2008b; Prem, Cook, and Jit 2017). Our results also show that seniors (65+ years) tended to be sources of infection for other senior individuals, again mirroring the known assortative nature of social contacts. Specifically, we find that individuals aged 75+ years are frequently sinks for infection arising in other senior individuals. Additionally, we find that individuals aged 75+ years transmitted to numerous groups. First, we see that they tended to transmit infection more frequently to individuals aged 55+ years. The transmissions from individuals ages 75+ years to the 55-59 year age group could represent transmissions to those individuals caring for them in professional or personal capacities. Interestingly, we also see frequent transmission to individuals ages 20-29 years, which could represent the grandparent-grandchild transmission effect noted by Towers et al. (Towers and Feng 2012).

When we examine HHS regions independently, the specific patterns of risk diverge from the general patterns detected at the national scale. For example, in regions 3, 4 and 7, we find significant phylogenetic risks for transmissions from school-aged individuals (15-24 years) to seniors (75+ years) (Figure 7). Alternatively, in region 2 we found an increased frequency of transmissions between school-aged individuals (Figure 7). In Region 9, we see significant transmission clustering between individuals aged 60+ years. Regions 1, 5, 6, 8 and 10 exhibit 'well mixed' transmission patterns where between groups takes place at roughly the same frequencies (Figures 7). Overall, these analyses highlight the inferential power of phylogenetic methods and the utility of using virus sequences for evaluating differences in regional influenza risks. Overall, we see stark heterogeneity between regions in the magnitude and distribution of risks among various age groups when compared to the national analysis.

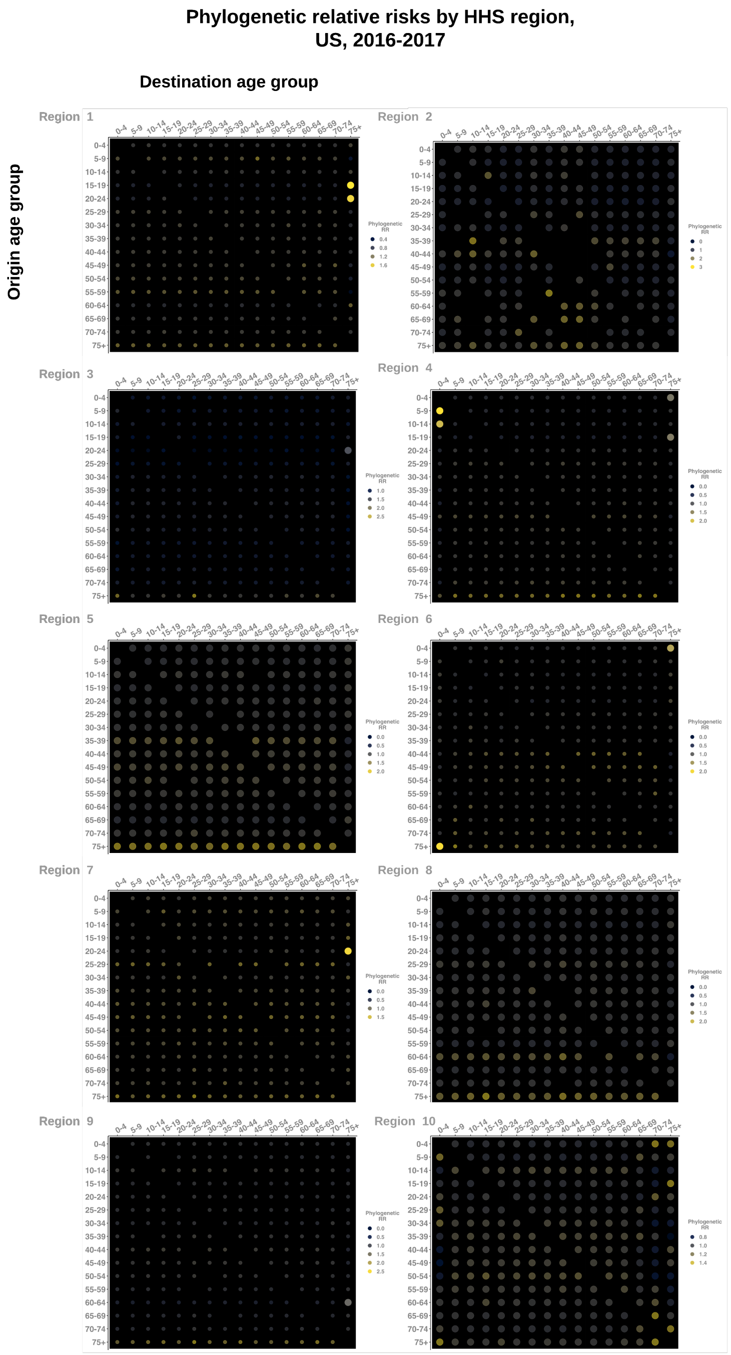


Figure 7. Visual summary of regional phylogenetic risk ratio results for the 2016-2017 US H3N2 influenza epidemic. We show significant risks as bubbles that are opaque and larger than corresponding spots within the same region. This allows us to quickly identify regions, such as regions 1, 5, 6, 8 and 10, which have well mixed epidemics in which all age groups transmit influenza to each other at similar magnitudes. In contrast, regions 2, 3, 4, 7, and 9 show evidence of significant age related structure to transmission patterns inferred from phylogenetic reconstructions.

1.5 Discussion

1.5.1 2016-2017 US Influenza A/H3N2 Epidemic through the Lens of Genomic Epidemiology

In this study, we quantified role of immunologic, demographic, age and social contact factors on the age associated diffusion on influenza A/H3N2 while simultaneously estimating the risk of transmission between various age groups by using molecular sequence data annotated with host age information. We identified, toddlers aged 0-4 years, school-aged individuals (aged 15-24 years) and seniors aged 75+ years as key age groups driving national influenza A/H3N2 transmissions using two independent, though complementary, modeling approaches. Though we take care to distinguish this national trend from patterns generated by region level phylogenetic analysis.

Statistical analyses of traditional surveillance case count data (Katelyn M Gostic et al. 2016; Katelyn M. Gostic et al. 2019) suggest a critical role for childhood immune imprinting in shaping case age distributions during influenza A outbreaks in the US. We corroborate that claim with both our phylogenetic and standard GLM analyses based on collected sequence data. However, the geographic resolution of our phylogenetic results paints a more detailed picture of the heterogeneous patterns of risk which characterize different HHS regions. Nationally, we see an elevated risk of transmission from seniors aged 75+ years to individuals aged 20-29 and 55+ years. Notably, this excludes individuals aged 35-50 years; corresponding to a period of time in which influenza A/H3N2 viruses were the only or dominant circulating influenza strains in the US (Katelyn M Gostic et al. 2016). Combined with the notion of subsequent immune protective effects conferred by such early life influenza exposures, we suspect these transmission patterns may driven by high probabilities of early life exposure to homosubtypic H3N2 viruses for these birth cohorts.

Previous studies suggest the nature of school age childrens' highly assortative contact networks are important for propagating initial seasonal epidemics (Glass and Glass 2008). Analyses of social contact and household structure data suggests young adults' and seniors' social contact network structures, especially in industrialized countries, show strong assortative features which support influenza transmission (Prem, Cook, and Jit 2017). To test these hypotheses, we augmented our phylogenetic and standard GLM analyses with social contact data (Mossong et al. 2008b) projected to HHS regional demographic structure (Arregui et al. 2018). We fail to detect significant effects due to total, daily social contacts, though this could be obfuscated by a strong signal for immunologic imprinting present in our data set. By calculating phylogenetic relative risks using the reconstructed transmission events between groups, we recover elevated transmission risks in associative groups, as expected by social contact patterns. Synthesizing our inferred transmission risks with previous reports that peak incidence occurs earlier in seniors than other groups (Schanzer, Vachon, and Pelletier 2011), we postulate that seniors may have played a critical role in seeding the initial phase of the 2016-2017 US epidemic.

There is concern among researchers about potential bias in phylogenetic reconstructions due to convenience sampling of isolates for sequencing (Lemey et al. 2014; Magee, Suchard, and Scotch 2017; Magee and Scotch 2018). To measure the degree of phylogeny-trait association in our data sets, we calculated the Parsimony Score (PS) and Association Index (AI) for each HHS regions. For this analysis, the null hypothesis is that discrete traits are *not* uncorrelated with genetic distance, as inferred by a phylogeny. Using both the AI and PS statistics, we reveal limited phylogeny-trait association for regions 3, 7 & 10. Regions 5 & 8 show limited phylogeny-trait association under the PS criterion, however, the AI indicated that there is, indeed, detectable phylogeny-trait association for these regions. Together, we interpret these results as showing detectable phylogeny-trait associations in all regions except 3, 7 & 10. Therefore, we are confident that our phylogenetic models for these regions are detecting true, underlying signals in the data. We suspect low sequence counts in region 10 and highly imbalanced sampling for regions 3 & 7 contributed to the lack of detectable phylogeny-trait association and reduces our confidence in the modeling results from these regions. We take additional steps to validate our phylogenetic GLM estimates via comparison against another Bayesian GLM fit to a line-list representation of our full sequence data set. Overall, we demonstrate congruence between the inferences rendered by each modeling framework, however, differences in method granularity prohibited more direct comparison between model estimates. We interpret the congruence and similar magnitude of coefficient effect size estimates as evidence that our phylogeographic models capture the underlying signal present in the data, bolstering their utility as actionable public health evidence.

1.5.2 Public Health Implications

Our findings suggest genomic epidemiological evidence is potentially useful for informing disease control efforts. Assuming that the age groups mix assortatively (i.ei with each other more than others), the optimal target for vaccination is the group with the highest transmission risk (Keeling and Rohani 2008). This has clear implications for informing disease control efforts if genomic epidemiology methods are used as evidence for decision making within public health contexts. Overall, our results suggest that regionally tailored approaches to influenza control may be warranted. For example, public health practitioners could use phylogenetic relative risk measures to focus control efforts on specific groups with elevated transmission risks. Applying the example using our national level analysis, our results suggest focusing generally on increasing vaccination coverage of school-aged individuals (aged 15-24 years) and seniors aged 75+ years, which may curb further spread to both other senior adults and individuals aged 20-29 years (Figure 7), both of which our results implicate as important groups for influenza spread during the 2016-17 season.

Extending the example to the regional level, different vaccination strategies tailored to specific regional risk scenarios, for example focusing on increasing vaccination coverage of children ages 5-14 in HHS regions 2, 3, 4 and 7 may lead to better influenza control. Similarly, our results suggest that increasing vaccination coverage among seniors in region 9 may bolster regional control efforts. Mathematical studies of optimal vaccine allocation during influenza epidemics have suggested prioritization of school-aged children (Mbah et al. 2013) as well as adults aged 30-39 years (Medlock and Galvani 2009), though these models are often based exclusively on social contact data collected (Mossong et al. 2008a) or inferred (Prem, Cook, and Jit 2017) from survey data. Alternatively, our approach integrates multiple sources of evidence, including social contacts, to arrive at coherent risk estimates offering complementary benefits to both surveillance based and mathematical approaches for risk determination and public health decision making.

While genomic epidemiology abound in the literature (Swetnam et al. 2018; Dudas et al. 2017a; Grubaugh et al. 2017; Trovão et al. 2015), descriptions of implementations of these approaches in public health agencies are paradoxically scant. We find only one implementation case study (Poon et al. 2016) reported in the literature, yet, there is no clear reference to implementation frameworks or accepted constructs in their treatment. Clear implementation strategies are required to facilitate uptake of these methods into routine public health workflows. Otherwise, their potential disruptive impact on public health practice will likely remain unrealized as they move from “bench to bookshelf”. Brownson et al. (Brownson, Fielding, and Maylahn 2009) describe two key tenets of evidence-based public health as i) using the best available peer-reviewed evidence and ii) systematically using data and information systems. We suggest genomic epidemiological approaches to risk quantification satisfy these requirements, especially considering the increasing interest (Gwinn, MacCannell, and Khabbaz 2017) in using molecular sequences for public health investigations. To facilitate the uptake of these methods, we plan to make these methods available in our open-source ZooPhy tool (Scotch et al. 2010; Scotch, Magge, and Vaiente 2019), alleviating, at least partially, technical barriers to the implementation of these methods. This is a critical first step in making outcomes of phylogenetic analyses accessible and actionable for public health workers, toward the aim of translating genomic epidemiological evidence into routine public health practice.

1.5.3 Limitations and Future Work

The nature of isolate sampling by reference laboratories for influenza HA gene molecular sequencing is a concern for epidemiological investigations using secondary data sources, such as GenBank. There is an understanding among researchers that imbalanced sampling schemes may lead to biased reconstructions of discrete trait evolution (Lemey et al. 2014; Magee, Suchard, and Scotch 2017; Magee and Scotch 2018). We motivated our use of HA molecular sequence as representative of infections in the underlying population by relying on guidance provided to state laboratories to establish systematic, statistically sound influenza surveillance systems. Our per region proportion of included sequences ranged between 45% (region 5) and 80% (regions 5 and 10, data not shown) of all available sequences in GenBank. Previous results suggest these sampling proportions should be sufficient to recover root trait signals embedded in molecular sequence data (Magee and Scotch 2018). However, we are not assured this same result when considering the total count and pattern of MJ between groups. Thus, differences in hospitalization and sampling rates between age groups may contribute to the observed results. Similarly, we only included sequences from the US in this analysis which may influence our reconstruction results. We also acknowledge that our work focuses exclusively on influenza A/H3N2 viruses. Other circulating seasonal variants of influenza A, as well as influenza B, may exhibit different age associated diffusion patterns. Future work is necessary to determine the statistical performance of these methods under various sampling intensities and epidemic scenarios. In this paper, we have shown how molecular epidemiology can shed light on subpopulation transmission dynamics during influenza outbreaks. We urge for continued reporting of relevant clinical metadata with pathogen genome sequences to enable further study of subpopulation transmission dynamics during influenza and other viral disease outbreaks. We envision studies which systematically apply tested implementation frameworks as a necessary next step in translation of these methods. Under the RE-AIM implementation framework (Glasgow, Vogt, and Boles 1999), quantifying reach, the proportion of the total population that receives benefit from an intervention and effectiveness (reduction of disease due to application of an intervention) are important steps for translating evidence into practice. These measurements must necessarily be made in real world contexts. Thus, future work is needed to quantify the effectiveness of molecular epidemiological methods in for quantifying risk and informing control efforts in such settings.

1.6 Acknowledgments

This research was supported by a grant from the National Library of Medicine of the National Institutes of Health under award number R01LM012080 to MS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to thank the ASU Department of Biomedical Informatics for providing computational resources to perform phylogenetic experiments. We would also like to thank the NVIDIA GPU Grant program for providing GPU hardware which was used for phylogenetic experiments with large state spaces.

1.7 Supplementary Methods

1.7.1 Bayesian GLM for Case Counts Observed in Sequence Databases

In Figure 8 we show the overall structure of our Bayesian GLM. We modeled the case counts, Y_i in each age group as multinomial random variables given by parameter pi describing the probability of a case in each age group. Following Gostic et al. (Katelyn M Gostic et al. 2016) we model this probability as a linear combination of a null expectation D_0 and covariates of interest. We discuss the derivation of the null expectation in the following section. We incorporated social contact, probability of matched immune imprinting and two dummy variables allowing for increased risk for young and elderly groups grouped into the design matrix X. We defer discussion of the details regarding these covariates until the following section. We take α as the complement of matched immune imprinting to ensure π sums to one. To achieve variable selection, we use a mixture of "spike and slab" normal priors (George and McCulloch 1997). To enable Gibbs sampling, we introduce binomial latent variables r_k which describe the mixture component membership of each β_k. We set the prior for r_k such that each variable had a 50% prior inclusion probability in the model. Taking the approach in (Lemey et al. 2014), we can determine variable importance via Bayes Factors (BF) using the ratio of posterior to prior inclusion odds. We fit our Bayesian GLMs using the full set of 2812 sequences via Gibbs sampling. We ran our MCMC for 150000 iterations, sampling at every step and discarded the first 50000 iterations as burn-in.

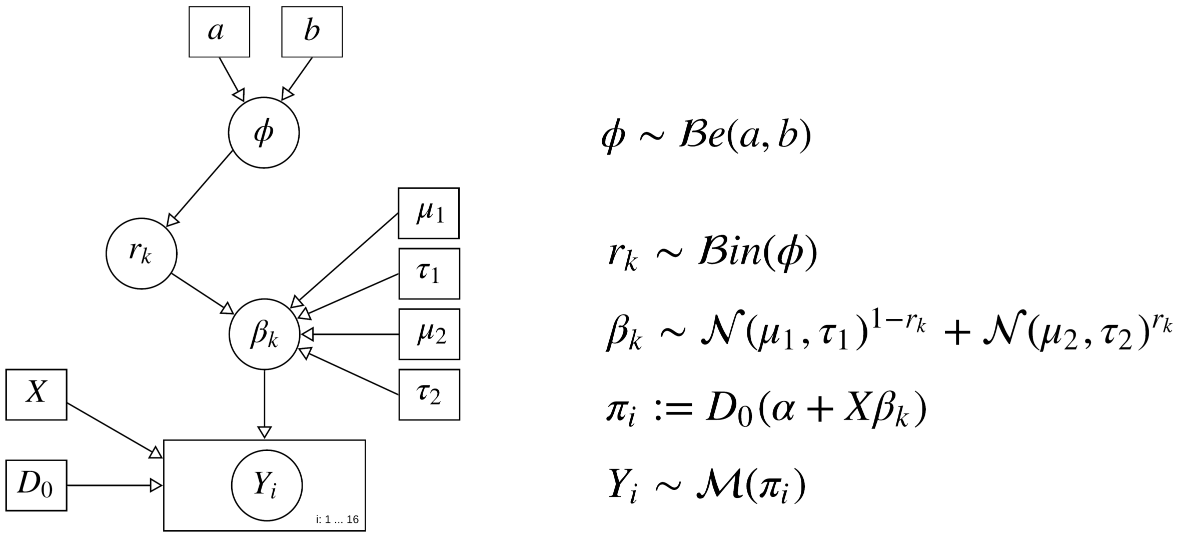

a
b
ϕ
r_k
μ_1
τ_1
β_k
μ_2
τ_2
X
D_0
Y_i
i: 1 ... 16
$\phi \sim \mathcal{Be}(a,b)$
$r_k \sim \mathcal{Bin}(\phi)$
$\beta_k \sim \mathcal{N}(\mu_1,\tau_1)^{1-r_k} + \mathcal{N}(\mu_2,\tau_2)^{r_k}$
$\pi_i := D_0(\alpha + X\beta_k)$
$Y_i \sim \mathcal{M}(\pi_i)$

Figure 8. Directed acyclic graph (DAG) describing the the Bayesian GLM for sequence counts.

1.7.2 Null Expectation for Influenza A/H3N2 Case Age Distribution

To generate a null expectation for influenza case age distributions, we first note that the distribution of case ages during an influenza outbreak will be determined by nonlinear transmission dynamics distributed across an age stratified social contact network. To reflect these assumptions, we employed an age-structured SIR model with 16 compartments representing the 16 non-overlapping 5-year age intervals described for our phylogeographic models.

$$\dot{S}_i = -\beta S_i \sum_j C_{ij} \frac{I_j}{N_j} \tag{1.1}$$

$$\dot{I}_i = \beta S_i \sum_j C_{ij} \frac{I_j}{N_j} - \gamma I_i \tag{1.2}$$

$$\dot{R}_i = \gamma I_i \tag{1.3}$$

We parameterized our SIR model using 2017 population estimates from the US Census and age stratified social contact rates (C_{ij}) estimated for the US population (Prem, Cook, and Jit 2017). Then, to complete our model specification, we assumed an R_0 of 1.4 and an infectious period (γ^{-1}) of 5 days, commensurate with previous studies of influenza H3N2 dynamics. We used these assumptions to calculate the per contact transmission probability, β, as 1.81%

Lemma 1.

The transmission probability per contact β is given by the expression $\frac{R_0 \gamma}{s(K)}$ where s is the spectral radius operator and K is the next-generation matrix (Driessche 2017).

Lemma 2.

The next generation matrix K is given by $\frac{\beta}{\gamma} * C_{ij} \frac{f_i}{f_j}$ where C_{ij} is the rate of contact between group i and group j and f_i is the fraction of the total population comprised of age group i.

We computed the null expectation (or equivalently, the prior) by simulating from our model over the course of one influenza season (365 days, beginning on week 40). Model simulations were initiated assuming one infected individual in each age group. Then, we calculated the proportion of infected individuals belonging to each age compartment at the end of the simulations. That is,

$$D_{0_i} = \int_0^T \frac{R_i}{\sum_i R_i} \tag{1.4}$$

1.7.3 Phylogenetic GLM for Age-associated Influenza A/H3N2 Diffusion

Bayesian phylogeographic models treat diffusion in discrete space as a K dimensional continuous time Markov chain, parameterized by a $K \times K$ infinitesimal rate matrix describing state transitions (Lemey et al. 2009). Recent work extends the parameterization of these models such that the individual rate parameters of the Markov transition matrix are assumed to be a log linear function of predictors of interest (Lemey et al. 2014). In our case we consider $K = 16$, where 16 is the number of 5 year age groupings that we consider. Our full model takes the form:

$$\log(\Lambda_{ij}) = \beta_1\delta_1 C_{i,j} + \beta_2\delta_2 f_i + \beta_3\delta_3 f_j + \beta_4\delta_4 A_i + \beta_5\delta_5 A_j + \beta_6\delta_6\tau_i + \beta_7\delta_7\tau_j \tag{1.5}$$

1.7.3.1 Predictors for Phylogeographic GLM of Age Associated Influenza A/H3N2 Diffusion

We tested several predictors of age-associated influenza diffusion in our GLM treatment.

- **Social contact patterns,** C_{ij}. Considering the total densities of possible hosts is natural in epidemiological modeling as SIR-type models often assumes that mass action transmission (Begon et al. 2002). We tested the impact of daily social contact patterns on age-associated influenza diffusion by estimating social contact matrices for each HHS region. We do so by transforming a social contact matrix estimated for the US by Prem et al (Prem, Cook, and Jit 2017) by applying a density correction method proposed in (Arregui et al. 2018) using demographic data from the US Census.

- **Population density,** f. To test the influence of population density on influenza diffusion, we calculated the proportion of total US population comprised of a particular age group using 2017 population size estimates from the US Census. We allow for separate predictors for the origin and destination groups by including both origin and destination predictors values (indexed by i, j, respectively).
- **Increased age group risk,** A. A basic tenet of influenza epidemiology states that the very young and elderly are at increased risk for infection. To reflect this assumption in our modeling scheme, we defined a variable A that is defined by an indicator function with value one for 0-4 and 65+ year age groups and is zero for all others. The approach emulates the variables included by in other work on age-associated influenza diffusion.
- **H3 imprinting probabilities,** τ. Recent work in modeling age-associated influenza diffusion suggests that immune dynamics play an important role in shaping the severity and age distribution of cases during seasonal epidemics (Katelyn M Gostic et al. 2016; Ranjeva et al. 2019; Arevalo et al. 2019). Specifically, the antigenic type of an individuals' first influenza exposure shapes their future ability to mount immune responses to subsequent influenza exposures, a process known as 'original antigenic sin' (Smith et al. 1999). We follow the examples of (Katelyn M Gostic et al. 2016; Arevalo et al. 2019) and include the probability that an individual born in year y was first infected (i.e. imprinted) by an H3-type influenza virus. *A priori*, we expect that individuals with low probabilities of H3 imprinting will transmit H3-type influenza at increased rates due to reduced ability to produce anti-H3 influenza immune responses. We use H3 imprinting probabilities as calculated by Arevalo et al. (Arevalo et al. 2019)

and estimate an age-group wide imprinting probability by averaging individual birth cohort probabilities.

Chapter 2

DIVERSITY, DILUTION AND WEST NILE VIRUS: INTERROGATING ECO-EPIDEMIOLOGICAL HYPOTHESES WITH VIRUS SEQUENCE DATA

2.1 Abstract

In recent years, genomic surveillance and phylodynamic analysis has emerged as the premier tool with which emerging disease outbreaks are studied. Though West Nile virus (WNV) has been studied extensively since its emergence in the US, our understanding of the factors shaping early patterns of WNV spread are paradoxically scant. Previous phylodynamics studies are focused on reconstructing spatial history of WNV without quantifying the roles that specific factors had in shaping observed epidemic patterns. In particular, the roles of avian species density and overall avian diversity in shaping patterns of WNV spread in the US remain uncertain. In this study, we employ Bayesian phylogeographic generalized linear models (GLMs) to interrogate the roles of avian species density and diversity in shaping early epidemic expansion of WNV in the US. We show that early WNV dispersal was driven primarily by American Robin (AMRO) density and that WNV was less likely to disperse between regions with high avian biodiversity, supporting the hypothesis that dilution effects are prominent in the WNV disease system. To our knowledge, our study represents the first effort to utilize genomic epidemiology to examine the association between avian species density and diversity and WNV dispersal in the US. Future work should focus on identifying correlates of highly granular WNV spread by linking appropriate molecular sequences sequences with detailed predictor and geographic metadata.

2.2 Introduction

West Nile virus (WNV) is a mosquito-borne, zoonotic RNA flavivirus capable of causing severe, neuroinvasive disease in human, avian and equine populations. WNV endemicity is maintained in an enzoonotic transmission cycle involving amplifying bird hosts and *Culex* species mosquitoes, though researchers have shown that vertical transmission within mosquitoes (Nelms et al. 2013) and direct transmission between birds (Komar et al. 2003) is also possible. After its introduction to the US via New York City (NYC) in 1999 (Lanciotti et al. 1999), WNV spread rapidly across the US and by the end of 2004 had been detected in 48 states (Malkinson et al. 2002; Swetnam et al. 2018). Several previous works focused on reconstructing the spatial history of WNV in the US (Pybus et al. 2012; Di Giallonardo et al. 2016; Swetnam et al. 2018) reach the general consensus that the virus spread rapidly from its origin in NYC, traveling simultaneously down the East Coast and across the Midwest. Researchers generally believe these patterns to be consistent with migratory bird movements (Pybus et al. 2012; Di Giallonardo et al. 2016; Swetnam et al. 2018; Hadfield et al. 2019), though explicit testing of these assumptions remains an open problem. WNV is distinguished by its relatively broad host range with infections reported in over 300 bird (Komar 2003; Marra et al. 2004; Reisen 2013) and numerous mammalian species (Reisen 2013). Though WNV infection occurs in several species, experimental evidence indicates that Passerine birds (those belonging to the order *Passeriformes*) are both highly competent WNV hosts (Komar et al. 2003) and are preferred by biting *Culex* species mosquito vectors in the US (Kilpatrick et al. 2006; Molaei et al. 2006; Hamer et al. 2009; Simpson et al. 2011) and Europe (Rizzoli et al. 2015). Of probable Passerine amplifying hosts, recent studies suggest that American Robins

(*Turdus migratorius*) and House Sparrows (*Passer domesticus*) (Komar 2003; Komar et al. 2003; Komar et al. 2005; Kilpatrick et al. 2006; Nemeth et al. 2009; Wheeler, Vineyard, et al. 2012; Wheeler, Langevin, et al. 2012) are among the most important hosts for WNV amplification and maintenance. American Robins (AMRO) are highly competent, migratory birds which tend to be the preferred host for biting *Culex* mosquitoes (Komar et al. 2003; Kilpatrick et al. 2006; Simpson et al. 2011). Their widespread distribution and migratory patterns have led researchers to hypothesize a key role of this species for WNV dissemination in the US. Conversely, House Sparrows (HOSP) are also highly competent hosts (Komar 2003) capable of sustaining viremia for several weeks post infection (Nemeth et al. 2009; Wheeler, Langevin, et al. 2012; Wheeler, Vineyard, et al. 2012). Field studies report seroprevalence rates of up to 69 percent (Komar et al. 2001; McLean 2006) within the species, indicating widespread exposure to WNV. Further, work by Duggal et al. (Duggal et al. 2014) shows that modern circulating WNV strains tend to produce higher viremia titers than the founder strain (NY99 genotype) in HOSP. Together, these results underscore the potential importance of these two species for viral amplification and maintenance in the US.

Another potential factor which influences the disease ecology of pathogens capable of infecting multiple host species of varying competence is the “dilution effect”; a mechanism which posits that increased host species biodiversity attenuates disease risk in multi-host pathogen systems. The dilution effect hypothesis was formalized by Ostfeld & Keesing in a seminal paper describing reduced Lyme disease prevalence in diverse host communities (Ostfeld and Keesing 2000), however, the concept has been generalized and tested in several multi-host pathogens (Ezenwa et al. 2005; Clay et al. 2009). In the case of WNV, empirical studies offer conflicting evidence regarding the associations between avian biodiversity and WNV. While early works report observation of dilution effects (Ezenwa et al. 2005; Swaddle and Calos 2008; Allan et al. 2009), some recent studies either report failing to detect dilution effects (Loss et al. 2009) or, in one case, detection of an an amplification effect (Levine et al. 2017); where disease risk is maximized (instead of minimized) in highly diverse host communities (Miller and Huppert 2013; Levine et al. 2017).

Though WNV has been studied extensively in the 20 years since its emergence, the specific factors enabling its rapid invasion and migration across the US remain unclear. In this paper, we take a Bayesian phylogeographic approach to quantifying the contribution of avian population and biodiversity factors on the spatiotemporal spread of WNV in the US. The specific aims of this study are to identify avian population and diversity factors that drove on the spatiotemporal spread of WNV in the US, while simultaneously reconstructing its early epidemic spread in the US. We will employ the phylogeographic generalized linear modeling framework introduced by Lemey et al (Lemey et al. 2014) by modeling WNV migration between Health and Human Services (HHS) regions as a log-linear combination of avian population and diversity predictors derived from Christmas Bird Count data (Bock and Root 1981). Adopting this framework permits statistical assessment of whether a predictor was supportive of or protective against WNV migration between specific HHS regions. We argue that a phylogeographic investigation of WNV ecology is especially prudent since pathogen genome sequencing is increasingly used to support emerging viral disease outbreak responses (Grubaugh et al. 2017; Dudas et al. 2017b; Dudas et al. 2018). We complement previous studies of WNV ecology by synthesizing evidence reported across the literature into a single modeling framework. Previous phylogeographic analyses of WNV are generally focused on reconstructing the migration history of

WNV without explicitly testing for associations between factors potentially driving the inferred spatial spread. While informative, these primarily descriptive studies do not attempt to quantify the contribution of various factors influencing WNV disease ecology (Pybus et al. 2012; Di Giallonardo et al. 2016; Swetnam et al. 2018; Hadfield et al. 2019). To our knowledge, this study represents the first application of phylogeographic GLM techniques to understanding the ecologic factors driving WNV migration in the US.

2.3 Methods

2.3.1 Sequence and Metadata Collection and Processing

We downloaded all published WNV molecular sequences available through GenBank (Dennis A. Benson et al. 2012). We excluded sequences from known dead-end hosts (humans and equines) and minor vector genera (*Coquillettidia*) which do not contribute significantly to viral propagation since our analysis is primarily focused on identifying ecologic factors associated with WNV spread. The resulting final data set was comprised of sequences obtained from birds and *Culex* mosquitoes. We then applied the following inclusion criteria to obtain a final sequence data set: a) sequences had a state-level location of sampling reported in their GenBank record b) sequences included the full coding region (CDS) of the WNV genome c) sequences had the year of isolate sampling available in their GenBank record and d) sequences had the host genus and species from which the isolate was derived reported in their associated GenBank record. We annotated each sequence with sampling date, location and host genus and species information to create the full data set. We generated independent analysis data sets by sampling 300 sequences and 500 sequences from the full set without replacement, respectively, in addition to sampling the full set of sequences. Our sampling scheme was motivated by previous work on ancestral reconstruction efficiency indicating that sampling 50% of available sequences provides convergent root state estimates (Magee and Scotch 2018) and computational expedience. We provide a visual summary of the sequence inclusion algorithm in Figure 9. We aligned each set of 300 sequences using MAFFT v. 7.407 (Katoh et al. 2002) using the default settings and inspected the results manually in Seqotron (Fourment and Holmes 2016).

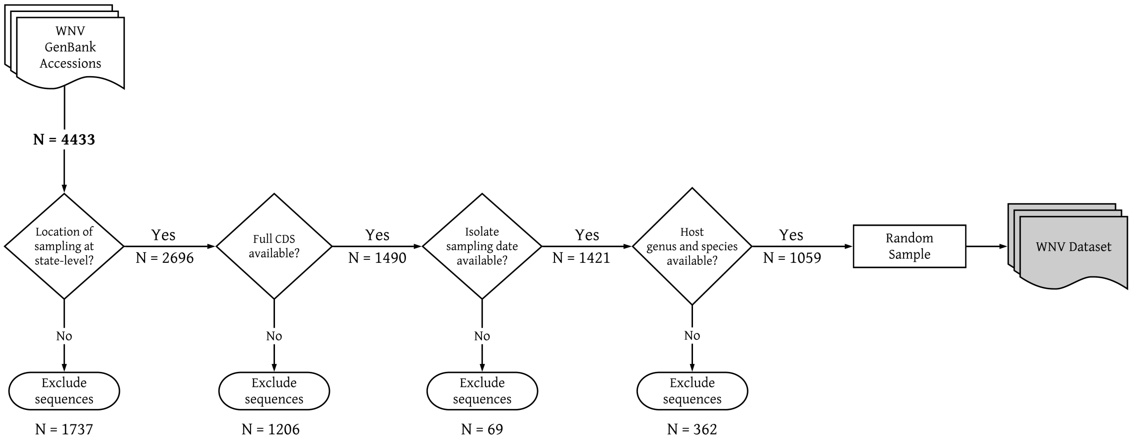


Figure 9. Overview of sequence inclusion algorithm. We downloaded all available full length, CDS sequences from GenBank and applied the inclusion criteria represented by decision nodes. The total sequences remaining after each step is shown below the arrow to the next decision node. The number of sequences discarded after each filtering step is shown below the corresponding exclude sequences node.

2.3.2 Bayesian Phylogenetic Analysis

We modeled molecular evolution using the GTR + Γ nucleotide substitution model with 4 rate categories and uncorrelated, log-normal relaxed molecular clock (Drummond et al. 2006), taking cues from other studies of WNV molecular evolution (Di Giallonardo et al. 2016; Swetnam et al. 2018). Since localized WNV epidemics have been reported periodically in several states across the US, we specified the non-parametric Bayesian skyline model as a prior for the tree generating model reflecting these *a priori* expectations about WNV demographic history. Since sampling time metadata are not required when uploading molecular sequences to GenBank, the resolution with which sampling dates are reported tend to be heterogeneous. To ameliorate this heterogeneity, we fixed tip dates as the year of sampling reported in GenBank for each sequence in order estimate divergence times. We note that other alternatives, such as a data-augmentation approach which treats the unknown sampling times as parameters within the model are possible. As we are primarily focused on reconstructing broad historical patterns in WNV, we feel that this choice is merited and supported by other work in this area (Di Giallonardo et al. 2016; Swetnam et al. 2018). We estimated the unknown parameters of our phylogenetic models using Bayesian Markov Chain Monte Carlo (MCMC) as implemented in BEAST v1.10 (Suchard et al. 2018). For each data set, we ran our MCMC for 350 million iterations, sampling every 35,000 steps and removed the first 25% as burn-in. We diagnosed convergence of the MCMC procedure using Tracer v1.7.1 (Rambaut et al. 2018a) checking that all model parameters had Effective Sample Sizes (ESS) of 200 or greater. We summarized posterior tree distributions as Maximum Clade Credibility (MCC) trees using TreeAnnotator (Suchard et al. 2018), discarding the first 2000 trees as burn-in.

2.3.3 Avian Predictor Data Collection and Processing

Conflicting results from previous empirical studies make it unclear whether avian biodiversity is broadly supportive of or protective against WNV disease risk (Ezenwa et al. 2005; Swaddle and Calos 2008; Levine et al. 2017). In other words, *is WNV ecology driven by a dilution effects or, conversely, does avian diversity increase risk?* To address on this question, we interrogate the dilution effect hypothesis using a phylogenetic GLM framework. We obtained Christmas Bird Count (CBC) data (Bock and Root 1981) from years 1999 to 2016 for the 39 avian species represented in our molecular sequence data set, generously provided by the National Audubon society. This data set included species count data for CBC years 2016 as well as survey area effort data that records the number of observers and observer hours for each reported count. We normalized total counts for each survey area, for each year, by the total number of observers contributing to each corresponding survey area count. This is motivated by the fact that observation effort heterogeneity will bias observed count data between survey areas. We computed HHS region level estimates of AMRO and HOSP density by averaging normalized species specific survey area counts within each HHS region. Previous empirical work focused on detecting dilution effects for WNV examined both Passerine and Non-Passerine diversity and their association with WNV disease risk. We follow this example and compute the Shannon diversity for Passerine and Non-Passerine species (Tramer 1969; Spellerberg and Fedor 2003) using the normalized counts for each survey area. We followed the same procedure of averaging estimates of Shannon diversity for all survey areas located within each HHS region to generate average HHS regional Shannon diversity predictors.

2.3.4 Identifying Ecological correlates of WNV Spatial Diffusion

We modeled the rate of WNV diffusion between Health and Human Serivces (HHS) regions in the US as a log-linear combination of avian predictors calculated from Christmas Bird Count data (Bock and Root 1981) using the Bayesian phylogeographic generalized linear model (GLM) framework introduced by Lemey et al (Lemey et al. 2014). Epidemiological reports indicate that WNV was first detected in the US 1999 (Lanciotti et al. 1999) and rapidly expanded to the West Coast where it was first detected in the summer of 2003 (Reisen et al. 2004). Based on these reports, we expect that the WNV migration in the US is divided into two phases; an early period in which WNV was actively spreading across the US (years 1999-2004) and a later period in which WNV became endemic is suitable localities (from 2004-

present). To reflect these assumptions, we employed the epoch modeling concepts introduced by (Bielejec et al. 2014) to allow for predictor temporal heterogeneity in our phylogeographic reconstructions. This approach involves dividing evolutionary history along a phylogeny into discrete partitions (that is, non-overlapping subsets) for which separate phylogeographic GLMs are specified. Following epidemiological reports of WNV, we defined two epochs corresponding to the periods in which WNV was actively spreading across the US (years 1999-2004) and when it had successfully established endemicity across several regions in the US and began to establish locally adapted strains (years 2004-present)

Given the role of amplifying host density for successful establishment of WNV in previously unaffected areas, we included predictors for each American Robins (AMRO) and House Sparrows (HOSP) motivated by previous work showing these species to be among the most important amplifying hosts for WNV. To test hypotheses related to avian biodiversity and WNV migration (which we take as a proxy for disease risk), we included predictors for Passerine, Non-Passerine and Total avian diversity in our phylogeographic GLM, following the examples of previous analyses of WNV risk and avian biodiversity (Ezenwa et al. 2005; Loss et al. 2009; Levine et al. 2017). For all predictors, we follow standard practices and normalized values prior to inclusion in the phylogeographic GLM. We additionally log transformed values for AMRO and HOSP density predictors, as is standard practice. Since Passerine and Non-Passerine Shannon diversity are calculated on the log scale, we include these predictors in our model without log transformation. We use the following abbreviations for predictors included in our model: total avian diversity (TD), Non-Passerine diversity (ND), Passerine diversity (PD), HOSP density (S) and AMRO density (R). We use separate predictors for origin and destination locations to incorporate predictor spatial heterogeneity into our analysis (Lemey et al. 2014). Epochs are denoted by the subscript t which separates the predictors and their inferred coefficients into the temporal periods described above. Practically, this is similar to change-point regression, a commonly used technique in classical statistics; except, in our case, the time at which the predictors change values is assumed to be known *a priori*.

Our "epochize" phylogeographic GLM takes the following form:

$$\Lambda_{ijt} = TD_{it}\beta_{1t}\delta_{1t} + ND_{it}\beta_{2t}\delta_{2t} + PD_{it}\beta_{3t}\delta_{3t} + S_{it}\beta_{4t}\delta_{4t} + R_{it}\beta_{5t}\delta_{5t} + \tag{2.1}$$

$$TD_{jt}\beta_{6t}\delta_{6t} + ND_{jt}\beta_{7t}\delta_{7t} + PD_{jt}\beta_{8t}\delta_{8t} + S_{jt}\beta_{9t}\delta_{9t} + R_{jt}\beta_{10t}\delta_{10t} \tag{2.2}$$

where each β is the coefficient for the corresponding predictor in the full GLM and each δ a binomial variable used to indicate the inclusion of a specific predictor in the model. The binomial indicator variables can be used in a Bayesian Stochastic Search Variable Selection (BSSVS) procedure as developed by (George and McCulloch 1997) and introduced into the phylogenetic GLM by (Lemey et al. 2014). Since we divided the our GLM into two epochs, we can perform BSSVS separately for each epoch allowing us to determine which predictors were important during the early epidemic expansion of WNV in the US. We performed inference for our phylogeographic GLM using Bayesian MCMC as implemented in BEAST v1.10.1 (Suchard et al. 2018). We ran our MCMC for 10 million steps, sampling every 1000 steps, conditioning on the last 1000 posterior trees from the associated phylogenetic inference for each sample.

2.3.5 Evaluation of Predictor Support

The key metric when evaluating predictor support the phylogeographic GLM are the posterior inclusion probabilities of each individual predictor. For a specific predictor, the posterior inclusion probability is calculated as the Monte Carlo expectation of the binomial indicator variables included in the GLM parameterization described above. Intuitively, a higher inclusion probability suggests greater statistical support for models including the corresponding predictor relative to models which omit this predictor. Following standard practices, we specified a prior inclusion probability for each predictor such that there is a 50% prior probability that, overall, no predictors are included in the model. We quantified the support of each predictor via calculation of Bayes factors (BFs) which are defined as the ratio of posterior and prior inclusion odds (Kass and Raftery 1995; Lemey et al. 2014). We use a cutoff of ≥3 as the baseline with which we compare posterior predictor support, since this is considered to be substantial evidence in support of a specific model hypothesis (Kass and Raftery 1995). Additionally, these cutoffs consistent with previous work in Bayesian phylogeography for evaluating evidence in support of specific predictors of GLM parameterized discrete trait diffusion (Lemey et al. 2009; Lemey et al. 2014; Magee et al. 2015).

2.4 Results

2.4.1 Timing, Origin and Evolution of WNV in the US

The final data set included 1059 WNV sequences annotated with host group, year and HHS region of sampling. We provide a summary of the sequence collection results in Figure 9. Of note, our final data set did not contain sequences from HHS region 10 which is reflected in Figure 10. In Table 5, we show a summary of the phylogenetic analysis results for the two independent samples. Overall, we inferred a mean substitution rates on the order of 10^{-4} substitutions per site, per year which is commensurate with other studies of WNV molecular evolution (May et al. 2011; Swetnam et al. 2018; Hadfield et al. 2019). Bayesian molecular clock dating of WNV genomes indicates that the most recent common ancestor (MRCA) occurred between late September to early October of 1997 (posterior mean TMRCA: 1997.54-1997.77, Table 5). We estimated the mean (95% HPD) TMRCA for each tree as 1997.74 (1996.82-1998.56) and 1997.77 (1996.98-1998.45) for the 300 and 500 sequence models, respectively. When including all available WNV genomes, we recover a slighly earlier mean estimate of 1997.54 (95% HPD region: 1996.15-1998.52). Though the earliest case of WNV was observed in the US in 1999 (Lanciotti et al. 1999), this is a reasonable TMRCA estimate given known monophyly and, indeed, sequence similarity between the US founding strain (NY99) and IS98; a strain isolated from a White Stork (*Ciconia ciconia*) during an outbreak in Israel in the summer of 1998 (Lanciotti et al. 1999; Malkinson et al. 2002; Charrel et al. 2003).

	Posterior mean (95% HPD)	
	Root age (TMRCA)	Substitution rate
300 sequence sample	1997.74 (1996.82-1998.56)	4.38e-04 (3.95e-04, 4.83e-04)
500 sequence sample	1997.77 (1996.98-1998.45)	4.70e-04 (4.42e-04, 4.97e-04)
Full sequence sample	1997.54 (1996.15-1998.52)	5.074e-04 (4.83e-04, 5.33e-04)

Table 5. **Posterior summary of Bayesian phylogenetic analysis**. We show the mean and 95% highest posterior density regions (95% HPD) for the root age, hierarchical mean substitution rate for each data set sample.

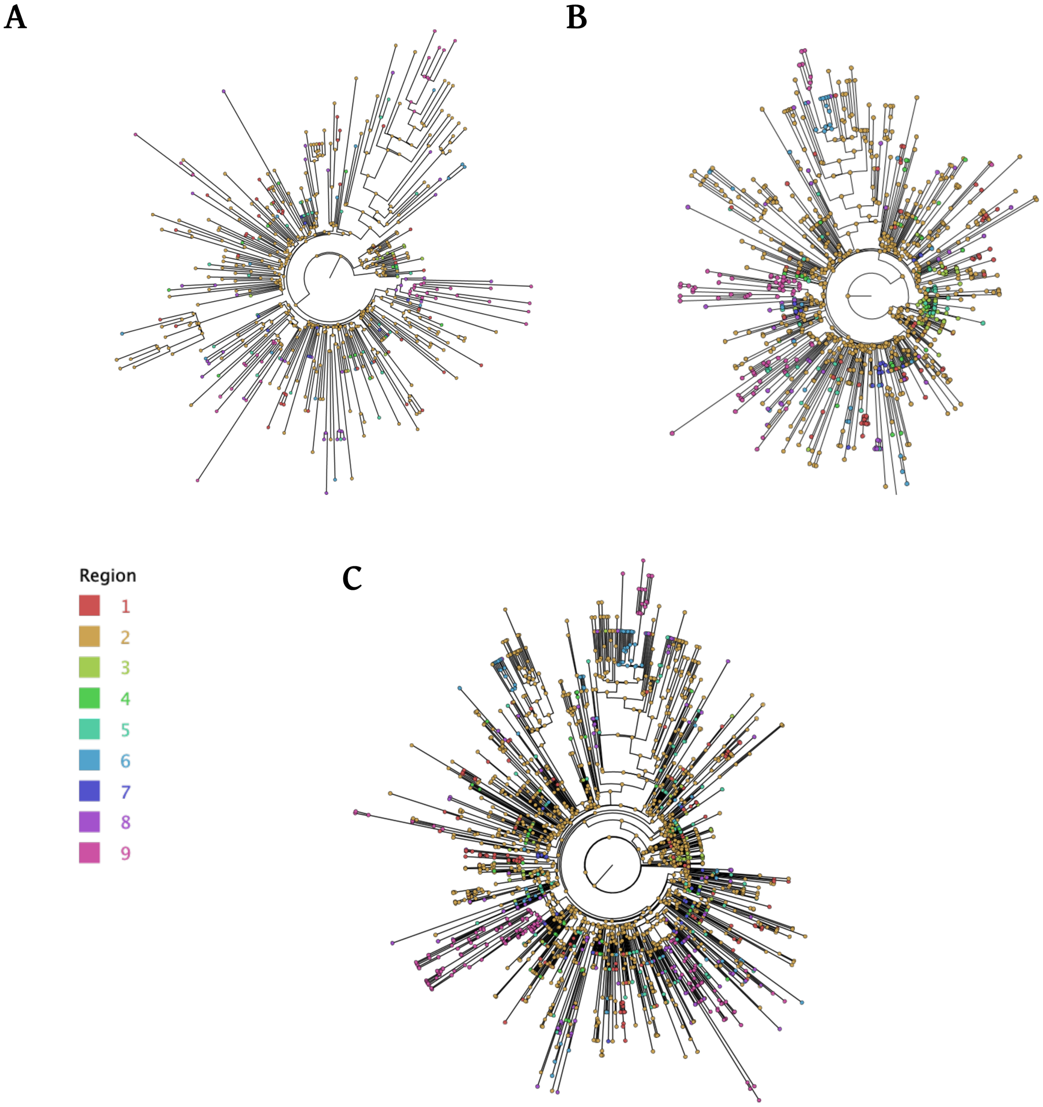


Figure 10. **Maximum Clade Credibility trees for 300 (A), 500 (B) and total (C) sequence samples**. As a note, the final data set did not contain any WNV genomes from HHS region 10 (Washington, Idaho, Alaska, Oregon) as discussed in the main text.

2.4.2 Avian Density and Diversity Drive Early WNV Migration in the US.

To determine the factors influencing the geographic dispersal of WNV in the US, we used an epochized phylogeographic GLM fitted to viral genomic data obtained from GenBank. In Figure 11, we show the posterior inclusion probabilities and coefficient effect size estimates for our epochized phylogeographic GLM. Panels A and B show the results from the 300 and 500 sequence samples, respectively, while panel C corresponds to results obtained for models fit to the full set of WNV genomes (Figure 11). We find that predictor support for WNV dispersal was primarily concentrated during the early epoch (years 1999-2004), congruent with our *a priori* expectations. Of the 10 predictors assessed, we found substantial evidence for 3 predictors which were supported categorically across sequence data samples (Tables 6, 8 & 10).

Our models show that WNV dispersal tended to originate from HHS regions with high total avian diversity (BF support for inclusion: >100, Tables 6 & 8). Specifically, we estimated strong supportive effects for total avian diversity (TD) at the region of origin (O). For the 300 sequence sample, we estimated a posterior mean coefficient effect size of 5.997 (95% HPD interval, 3.882-7.929, Table 6). We similarly observed a strong positive associations for models fit the 500 sequence sample; where we estimated a posterior mean coefficient effect size of 7.309 (95% HPD: 5.079-9.860, Table 8). When we fitted our phylogeographic GLMs to the the full set of WNV genomes, we see that total avian diversity is similarly, strongly supported with a posterior mean coefficient of 10.036 (95% HPD region: 8.061-12.134, Table10). The positive effect of total avian diversity on WNV dispersal and disease risk is consistent with the notion of an amplification effect, as previously described for WNV (Levine et al. 2017). However, this can be due to myriad mechanisms. One plausible explanation is that highly diverse communities are more likely to contain sufficient densities of those highly competent avian hosts which are preferred by biting *Culex* vectors (i.e American Robins, Northern Cardinals, House Sparrows and others) necessary to initiate and sustain enzoonotic WNV transmission cycles. A related hypothesis suggests that avian host diversity may be associated with *Culex* vector abundance by providing numerous bloodmeals on WNV incompetent hosts which indirectly leads to higher infection rates (Levine et al. 2017). Regardless of the specific mechanism, our results indicate a clear signal that total avian diversity in the region of origin was the primary driver of regional WNV dispersal from 1999 to 2004.

We found American Robin density at the origin to be strongly positively associated with viral dissemination between HHS regions in the US (BF support for inclusion: >100, Tables 6 & 8. For models fit to the 300 sequence sample, we estimated a posterior mean coefficient effect size of 3.32 (95% HPD: 1.96-4.829, Table 6) and a BF associated with "decisive" support (Kass and Raftery 1995). Similarly, we estimated a strong positive association for phylogeographic models fit to the 500 sequence sample. Here, we estimated a posterior mean coefficient effect size of 1.397 (95% HPD region: 0.413-3.217, Table 8) which, albeit lower than our model estimates for the 300 sequence sample, corroborates the presence and direction of this association. The BF for this predictor also corresponded to "decisive" posterior support (BF: 581.542, Table 8). Finally, when fit to all available WNV sequence data, our modeling results confirm the presence of a positive association between regional WNV dispersal and American Robin density (Table 10). Our phylogeogrpahic models provide evidence which directly suggests that AMRO density was broadly supportive of regional WNV dispersal within the US from 1999-2004, after which its effect diminished, reflecting reduced inter-regional WNV dispersal after 2004. The importance of American Robins in maintaining and amplifying local WNV epidemics is well established (Kilpatrick et al. 2006; Marm Kilpatrick et al. 2006). Here, we provide substantial evidence that AMRO density is positively correlated with the regional dissemination of WNV from 1999-2004; similar to other supported predictors, we see that the effect of this predictor is negligible after 2004 (BF for inclusion: <1, Tables 7 and 9).

Since Non-Passerines are known to be relatively incompetent hosts compared to their Passerine counterparts and are occasional bloodmeal hosts for biting *Culex* vectors (Molaei et al. 2006; Simpson et al. 2011), disease ecological theory predicts that avian communities with high Non-Passerine diversity will be less capable of sustaining enzoonotic WNV transmission, resulting in a net protective effect (Ostfeld and Keesing 2000; Ezenwa et al. 2005) on WNV disease risk. We tested whether this protective effect extended to regional WNV spread by including it as a predictor in our phylogeographic GLM. Consistent with ecological theory, we inferred a protective effect on regional WNV dispersal in our phylogeographic models. This suggests a reduced likelihood of WNV dispersal between regions with high Non-Passernie avian diversity. We estimated posterior mean coefficient effect sizes of -4.777 and -4.349 for models fit to the 300 and 500 sequence samples, respectively (Tables 6 & 8). These strong negative associations are accompanied by strictly negative 95% HPD regions and decisive posterior support (BF: >100, Tables 6 & 8). For models fit to the 300 and 500 seuqence samples, we note that this effect was localized to the region of origin (Tables 6 & 8). However, when increasing the data available to our phylogeographic models, we recover a protective effect for both origin and destination regions. For the full sequence models, we estimated mean posterior coefficients of -6.106 and -1.261 for the origin and destination regions, respectively, along with strictly negative 95% HPD regions (Tables 10). Together, this offers strong evidence of a negative association between WNV dispersal and Non-Passerine diversity and offers empirical support for the dilution effect during early WNV expansion in the US.

We found a significant propensity for reduced regional WNV dispersal from regions with highly diverse Passerine avian communities. Though models fit to the 300 and full sequence samples are associated with strong evidence according to the BF, the coefficient 95% HPD region estimates cross zero, which is commonly interpreted as a lack of association in phylogeographic models (Table 6 & Table 10). However, we found a negative association which meets both standards of evidence when we fitted our model to the 500 sequence sample. In particular, we estimated a posterior mean coefficient effect size of -3.111 alond with a strictly negative 95% HPD region, consistent with the notion of a protective, dilution effect (Table 8). It is expected that highly diverse Passerine communities results in sufficiently low densities of key amplifying hosts such that it reduces the likelihood of sustained epizootic transmission in these communities. Our models corroborate theoretical expectations by demonstrating that this protective effect extends to geographic dispersal of WNV. We observed numerous predictors in both epochs which correspond to strong or decisive support when considering their BF alone. However, in concordance with standard interpretations of phylogeographic GLMs, we only considered predictors with BFs greater than 3 with strictly positive (or negative) 95% HPD regions as providing substantial statistical evidence of an association with WNV dispersal. Predictors with substantial BF support and 95% HPD regions which contain zero are: Non-Passerine density in destination region, total avian diversity in the destination region and Passerine diversity in the destination region (Tables 6, 7, 8 & 9).

We found moderate support for total avian and Passerine diversity influencing regional WNV migration during the second epoch (Figure 11). Interestingly, we observed a shift from increased probability of WNV diffusion from regions with highly diverse (total) avian communities, to that of a reduced probability of diffusion between these regions, signaling a shift in WNV disease ecology. During this epoch, we estimated that total avian diversity acts to reduce geographic dissemination of WNV, which is possible through the mechanisms discussed above (Table 11). The increase in dispersal propensity for regions with high Passerine diversity may be related to the co-evolution between WNV and resident Passerine birds. If so, the mechanisms for increasing dispersal probability would be similar to those discussed for total avian diversity; more diverse Passerine communities are more likely to contain high densities of those few, highly competent Passerine hosts.

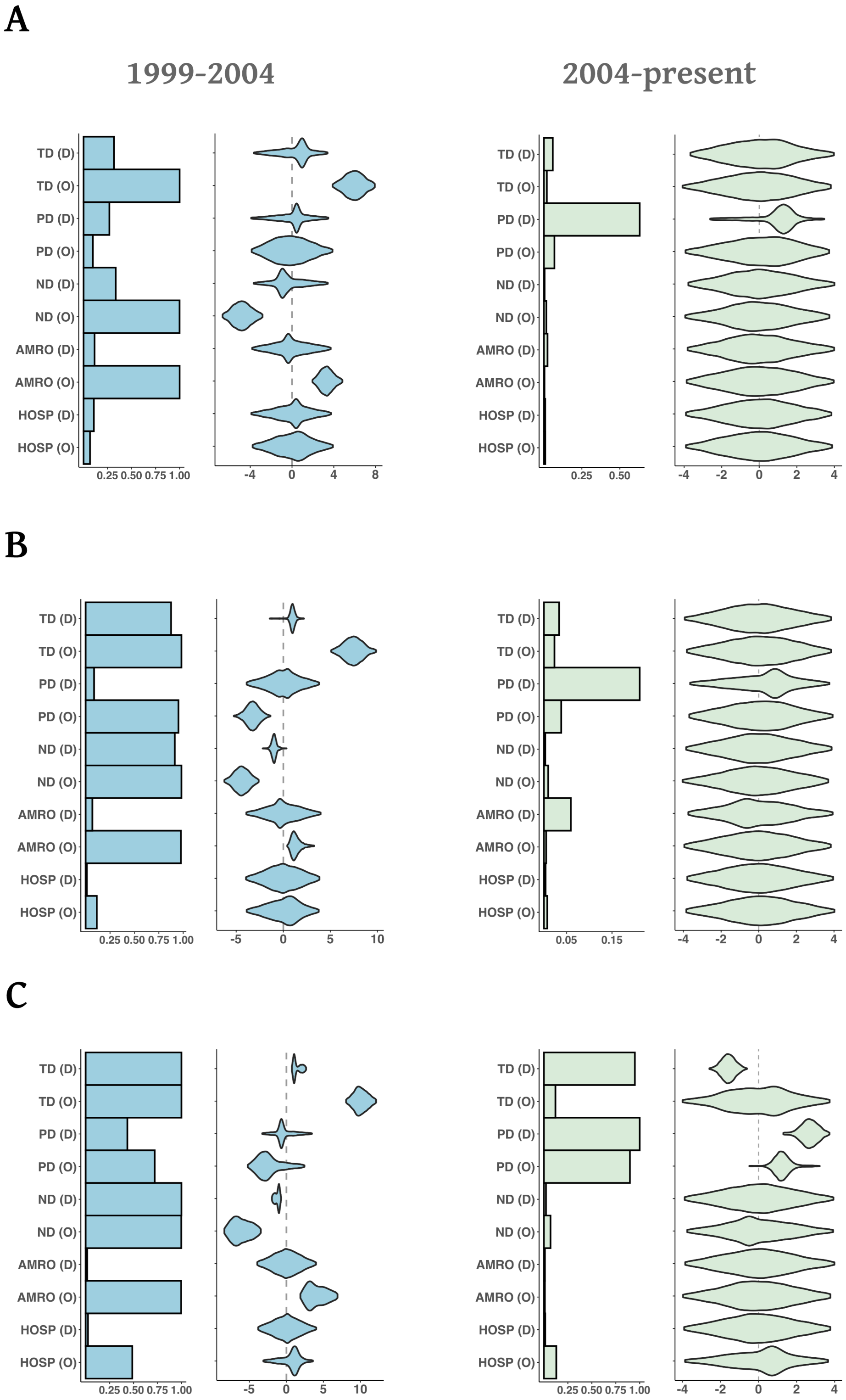


Figure 11. **Avian density and diversity support regional WNV migration**. We fit "epochized" phylogeographic GLMs to 300 (A), 500 (B) and full (C) sequence samples of WNV genomes obtained from GenBank. Overall, we find strong support for three predictors (AMRO, ND, TD) across data samples indicating avian density and diversity helped shape WNV dispersal patterns during its early epidemic expansion in the US.

	Summary Statistics		
Predictor	Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.067	0.066 (-3.767, 3.918)	1.0
House Sparrow Density (d)	0.107	0.048 (-3.891, 3.727)	1.6
Robin Density (o)	**0.999**	**3.320 (1.960, 4.829)**	**-**
Robin Density (d)	0.114	-0.081 (-3.807, 3.707)	1.8
Non-Passerine Diversity (o)	**0.999**	**-4.777 (-6.645, -2.819)**	**-**
Non-Passerine Diversity (d)	0.335	-0.281 (-3.668, 3.430)	7.0
Passerine Diversity (o)	0.096	-0.139 (-3.862, 3.878)	1.5
Passerine Diversity (d)	0.268	0.132 (-3.910, 3.478)	5.1
Total Avian Diversity (o)	**0.999**	**5.977 (3.882, 7.929)**	**-**
Total Avian Diversity (d)	0.317	0.278 (-3.647, 3.406)	6.5

Table 6. **Posterior summary of 300 sequence Bayesian phylogeographic GLM, 1999-2004**. Here, we bold rows associated with predictors for which we estimated BF greater than 3 that have 95% HPD regions which do not contain zero. Dashes reepreseent Bayes' Factors greater than 1000

	Summary Statistics		
Predictor	Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.009	-0.005 (-3.909, 3.888)	0.128
House Sparrow Density (d)	0.008	-0.003 (-3.902, 3.848)	0.118
Robin Density (o)	0.004	0.006 (-3.865, 3.974)	0.056
Robin Density (d)	0.024	-0.004 (-3.813, 3.993)	0.340
Non-Passerine Diversity (o)	0.016	0.006 (-3.944, 3.741)	0.228
Non-Passerine Diversity (d)	0.004	0.021 (-3.778, 4.005)	0.063
Passerine Diversity (o)	0.069	0.100 (-3.929, 3.723)	1.037
Passerine Diversity (d)	0.631	0.868 (-2.615, 3.466)	23.839
Total Avian Diversity (o)	0.019	0.018 (-4.082, 3.815)	0.273
Total Avian Diversity (d)	0.059	0.017 (-3.665, 3.975)	0.869

Table 7. **Posterior summary of 300 sequence Bayesian phylogeographic GLM, 2004-2016**.

	Summary Statistics		
Predictor	Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.115	0.097 (-3.888, 3.758)	1.814
House Sparrow Density (d)	0.013	0.010 (-3.961, 3.839)	0.179
Robin Density (o)	**0.977**	**1.397 (0.413, 3.271)**	**581.5**
Robin Density (d)	0.069	-0.023 (-3.928, 3.963)	1.037
Non-Passerine Diversity (o)	**0.982**	**-4.340 (-6.260, -2.606)**	**755.9**
Non-Passerine Diversity (d)	0.914	-0.905 (-2.209, 0.345)	147.9
Passerine Diversity (o)	**0.951**	**-3.111 (-5.253, -1.364)**	**270.4**
Passerine Diversity (d)	0.086	-0.040 (-3.890, 3.804)	1.317
Total Avian Diversity (o)	**0.982**	**7.309 (5.079, 9.860)**	**755.9**
Total Avian Diversity (d)	0.875	0.883 (-1.444, 2.202)	97.80

Table 8. **Posterior summary of 500 sequence Bayesian phylogeographic GLM, 1999-2004**. Here, we bold rows associated with predictors for which we estimated BF greater than 3 that have 95% HPD regions which do not contain zero.

	Summary Statistics		
Predictor	Posterior Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.007	0.043 (-3.852, 4.023)	0.10
House Sparrow Density (d)	0.004	-0.019 (-3.800, 3.957)	0.05
Robin Density (o)	0.005	-0.024 (-3.955, 3.832)	0.07
Robin Density (d)	0.059	-0.024 (-3.742, 3.932)	0.88
Non-Passerine Diversity (o)	0.009	-0.002 (-4.046, 3.698)	0.13
Non-Passerine Diversity (d)	0.003	0.001 (-3.863, 3.875)	0.04
Passerine Diversity (o)	0.038	0.041 (-3.691, 3.930)	0.55
Passerine Diversity (d)	0.211	0.206 (-3.650, 3.756)	3.72
Total Avian Diversity (o)	0.023	0.034 (-3.821, 3.847)	0.33
Total Avian Diversity (d)	0.033	0.018 (-3.930, 3.850)	0.48

Table 9. **Posterior summary of 500 sequence Bayesian GLM, 2004-2016**.

	Summary Statistics		
Predictor	Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.48	0.577 (-3.117, 3.578)	13.2
House Sparrow Density (d)	0.02	-0.032 (-3.817, 4.036)	0.35
Robin Density (o)	**0.99**	**4.178 (1.893, 6.915)**	**-**
Robin Density (d)	0.02	-0.006 (-3.940, 4.008)	0.23
Non-Passerine Diversity (o)	**0.99**	**-6.106 (-8.389, -3.452)**	**-**
Non-Passerine Diversity (d)	**1.0**	**-1.261 (-1.954, -0.719)**	**-**
Passerine Diversity (o)	0.721	-2.196 (-5.218, 2.454)	36.0
Passerine Diversity (d)	0.435	-0.297 (-3.265, 3.449)	10.7
Total Avian Diversity (o)	**1.0**	**10.036 (8.061, 12.134)**	**-**
Total Avian Diversity (d)	**1.0**	**1.502 (0.746, 2.660)**	**-**

Table 10. **Posterior summary of 1059 sequence Bayesian phylogeographic GLM, 1999-2004**. Here, we bold rows associated with predictors for which we estimated BF greater than 3 that have 95% HPD regions which do not contain zero. Additionally, we show BFs with associated posterior inclusion probabilities of 1 as dashes.

	Summary Statistics		
Predictor	Inclusion Prob.	Mean (95% HPD)	BF
House Sparrow Density (o)	0.129	0.078 (-3.867, 3.657)	2.067
House Sparrow Density (d)	0.014	-0.021 (-3.919, 3.775)	0.194
Robin Density (o)	0.010	0.010 (-3.972, 3.787)	0.145
Robin Density (d)	0.011	-0.002 (-3.859, 3.990)	0.151
Non-Passerine Diversity (o)	0.069	-0.005 (-3.760, 3.925)	1.031
Non-Passerine Diversity (d)	0.022	0.006 (-3.886, 3.940)	0.310
Passerine Diversity (o)	**0.898**	**1.161 (-0.490, 3.215)**	**122.6**
Passerine Diversity (d)	**1.000**	**2.646 (1.302, 3.731)**	**-**
Total Avian Diversity (o)	0.121	0.024 (-3.997, 3.738)	1.916
Total Avian Diversity (d)	**0.950**	**-1.519 (-2.588, -0.608)**	**263.6**

Table 11. **Posterior summary of 1059 sequence Bayesian phylogeographic GLM, 2004-2016**. Here, we bold rows associated with predictors for which we estimated BF greater than 3 that have 95% HPD regions which do not contain zero. Additionally, we show BFs with associated posterior inclusion probabilities of 1 as dashes.

2.5 Discussion

The recent emergence of numerous zoonotic viruses such as Ebola, Zika and SARS-CoV-2 are serious public health threats which have caused widespread health and economic devastation. Modern tools and methods are urgently needed to adapt to these emerging and converging disease risks. For zoonotic diseases, an understanding of specific factors which contribute to disease dispersal is a prerequisite to developing potential interventions to curb disease spread. A trend in recent years toward rapid, genomic sequencing of viral pathogens has proven to be a powerful tool for reconstructing the timing and spread of emerging disease outbreaks while simultaneously assessing factors which support disease dispersal (Dudas et al. 2017b; Deng et al. 2020). The introduction and spread of WNV in the Americas provides an ideal model system with which to study the ecologic factors which support or suppress emerging disease migration. Previous phylogeographic studies of WNV have shown rapid early spread that is consistent with migratory bird movements (Pybus et al. 2012; Di Giallonardo et al. 2016; Swetnam et al. 2018; Hadfield et al. 2019), though this work is largely descriptive, offering few insights into factors shaping observed migration patterns. In this work, we modeled the regional diffusion of WNV using Bayesian phylogeography which allowed us to simultaneously evaluate the contribution of avian demographic and diversity factors on WNV spread in the US. Numerous field and experimental studies show Passerine birds are highly competent WNV (Komar et al. 2003) and frequent (Molaei et al. 2006; Hamer et al. 2009; Simpson et al. 2011; Rizzoli et al. 2015) bloodmeal hosts for biting *Culex* sp. mosquitoes suggesting that these birds are critical for dissemination of WNV in the US. In particular, researchers consider two species to be the most important amplifying hosts: the House Sparrows (Komar et al. 2005; Nemeth et al. 2009; Wheeler, Vineyard, et al. 2012) and American Robins (Marm Kilpatrick et al. 2006; Kilpatrick et al. 2006; Molaei et al. 2006; Simpson et al. 2011).

Similar to previous works, our phylogeographic inference shows rapid, westward expansion of WNV emanating from the northeastern US (HHS region 2, notably containing NYC) beginning in 1999 which continued until 2004. We see this reflected in phylogeographic reconstructions of WNV dispersal. These models suggest that long range WNV dispersals are expected to be more frequent during this early epidemic period by the inclusion of more predictors with large coefficient effect sizes (Figure 11). After 2004, long range WNV dispersals were relatively rare and locally adapted WNV strains began to emerge. For this period, we found little posterior support for any predictor, suggesting (in the context of the underlying CTMC) long waiting times between migration events such that they were not observed. From 1999 to 2004, phylogeographic analysis shows that AMRO density was the principle determinant of WNV diffusion in the US, suggesting an increased likelihood of dispersal from regions with high American Robin densities. While the importance of AMRO hosts for WNV ecology is echoed elsewhere in the literature, here we provide direct evidence in support of the hypothesis that this species was involved in the westward diffusion of WNV in the US. American Robins have been frequently implicated as the preferred bloodmeal host for competent, biting WNV mosquito vectors (Molaei et al. 2006; Marm Kilpatrick et al. 2006; Kilpatrick et al. 2006; Simpson et al. 2011). Recent work revealing that Texas, New York and Illinois were critical diffusion loci for WNV (Swetnam et al. 2018) also lend support to these conclusions since these states tend to be associated with high AMRO densities as reported in eBird (Sullivan et al. 2009), a citizen science database of bird sightings. Together these results support the notion that migratory American Robins were integral in disseminating WNV across the US.

Though there is compelling evidence of of co-evolution between HOSP and WNV (Duggal et al. 2014), we fail to detect an association with WNV dispersal in the US (Figure 11). This could be due to several reasons. House Sparrows are primarily a resident species in urban environments; therefore, though we expect this species to be important for local WNV amplification and maintenance they may have limited contributions to the long-range geographic dissemination of WNV.

Conflicting evidence regarding the associations between avian biodiversity and WNV risk (Ezenwa et al. 2005; Swaddle and Calos 2008; Allan et al. 2009; Loss et al. 2009; Levine et al. 2017) motivated our phylogeographic investigation in which we quantify the direction of association between avian biodiversity and WNV geographic diffusion; a proxy measure for WNV disease risk. Specifically, two dominant hypotheses related to biodiveristy effects on WNV disease ecology have been proposed in the literature. The first posits that biodiversity has a protective effect on disease risk; reducing vector-borne disease transmission in the presence of moderately competent and incompetent hosts. Conversely, the notion that host biodiversity can serve to increase disease transmission has been demonstrated to be theoretically (Begon 2008; Miller and Huppert 2013) possible and, in one case, detected for the WNV system (Levine et al. 2017). We tested whether avian biodiversity was broadly supportive of or protective against WNV dispersal using a phylogeographic models. Overall, we find a significant protective effect of Non-Passerine avian biodiversity on WNV migration (mean coefficient effect sizes: -4.77 & -4.34, Tables 6 & 8) across all model replicates. As predicted by disease ecological theory, this protective effect is likely due to the removal of otherwise infectious bites from more competent, Passerine hosts (Ostfeld and Keesing 2000; Ezenwa et al. 2005).

We similarly find a reduced propensity of WNV dispersal from regions with high

Passerine biodiversity. We fail to detect significant posterior evidence supporting the inclusion of Passerine biodiversity in our 300 sequence models since the 95% HPD region crosses zero (Table 6). However, when fit to a larger set of sequence data, we find substantial evidence of reduced WNV diffusion for regions with highly diverse Passerine avian communities (Table 8). Again, we expect increased Passerine host biodiversity reduces the likelihood that a given bloodmeal host for biting *Culex* sp. vectors is a highly competent WNV host. It is entirely possible that our phylogeographic models lacked sufficient power to detect protective effects associated with Passerine avian biodiversity when fit to only 300 WNV genomes. In our analyses, we found that WNV transmission was more likely when originating from regions with high total avian diversity, which may seem paradoxical given the aforementioned negative associations between biodiversity and WNV dispersal. In terms of disease ecologic hypotheses, this is consistent with an 'amplification effect' where disease risk is maximized, instead of minimized, in mixed host communities (Miller and Huppert 2013). Levine et al (Levine et al. 2017) also detect positive associations between avian diversity and WNV risk, consistent with an amplification effect, though astutely argue that this association may be confounded with community composition; since increased total avian diversity would subsequently increase the occurrence of highly competent WNV hosts. We conclude that this may be driving the observed associations in our phylogeographic analyses; American Robins are relatively rare in many regions across the US, increased avian diversity in these regions increases the probability that these hosts are present to amplify and disseminate WNV infections.

In this paper, we have demonstrated the use of phylogeographic GLMs for identifying the ecological drivers of WNV dispersal in the US. Overall, our results suggest that AMRO are the dominant amplifying species responsible for the westward diffusion of

WNV in the US and that protective, dilution-type effects were prominent in shaping early WNV spatial dynamics. To our knowledge, our study is the first to demonstrate a direct role of American Robins in driving geographic diffusion while simultaneously quantifying the association between biodiversity and WNV spread in a statistical framework which is based on viral molecular sequence data. Collectively, this study highlights the value of phylogeographic evidence for examining disease ecologic and epidemiological hypotheses about factors related to the emergence and spread of zoonotic infectious diseases. As geographic metadata become available at increasing resolution for virus sequences, a promising avenue of research will be modeling local differences in WNV risk patterns using highly granular bird abundance data coupled with experimentally derived, species-level competence indices. We believe the value of phylogeographic techniques for testing disease ecologic and epidemiological hypotheses will continue to increase as molecular sequences and highly granularity geographic metadata become increasingly available for myriad pathogens. Although these models are not amenable to predicting future dispersal patterns, they do provide valuable, albeit retrospective information about factors driving dispersal in emerging disease epidemics. Understanding where transmission hotspots are likely to occur, or which hosts are likely to be actively disseminating infections, is critical for effective disease surveillance and intervention programs. Finally, viral phylogeography studies are limited by scant metadata reporting with respect to location and species of infected host. We elected to model WNV on the HHS regional scale due to heterogeneity in sequencing availability at the state level (Di Giallonardo et al. 2016; Swetnam et al. 2018). Therefore, our models do not capture fine scale differences in community competence associated with diverse community compositions since we model diversity using a single summary measure (Kain and Bolker 2019). In our study, we discarded nearly half of otherwise suitable sequences due to missing host species metadata. Future studies will benefit from increased effort to include relevant epidemiological metadata along with molecular sequences when uploading these data to publicly available databases.

2.6 Acknowledgements

This research was supported by grants from the National Library of Medicine and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award numbers R01LM012080 and R01AI117011, respectively. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank the ASU Department of Biomedical Informatics for providing computational resources to perform phylogenetic experiments. We also thank the NVIDIA GPU Grant program for providing GPU hardware which was used to accelerate phylogenetic experiments with large state spaces.

Chapter 3

GOING BACK TO THE ROOTS: EVALUATING BAYESIAN PHYLOGEOGRAPHIC MODELS WITH DISCRETE TRAIT UNCERTAINTY

3.1 Abstract

Phylogeography is a popular way to analyze virus sequences annotated with discrete, epidemiologically-relevant, trait data. For applied public health surveillance, a key quantity of interest is often the state at the root of the inferred phylogeny. In epidemiological terms, this represents the geographic origin of the observed outbreak. Since determining the origin of an outbreak is often critical for public health intervention, it is prudent to understand how well phylogeographic models perform this root state classification task under various analytical scenarios. Specifically, we investigate how discrete state space and sequence data set influence the root state classification accuracy. We performed phylogeographic inference on several simulated DNA data sets while i) increasing the number of sequences and ii) increasing the total number of possible discrete trait values. We show that phylogeographic models tend to perform best at intermediate sequence data set sizes. Further, we demonstrate that a popular metric used for evaluation of phylogeographic models, the Kullback-Leibler (KL) divergence, both increases with discrete state space and data set sizes. Further, by modeling phylogeographic root state classification accuracy using logistic regression, we show that KL is not supported as a predictor of model accuracy, indicating its limited utility for assessing phylogeographic model performance on empirical data. These results suggest that relying solely on the KL metric may lead to artificially

inflated support for models with finer discretization schemes and larger data set sizes. These results will be important for public health practitioners seeking to use phylogeographic models for applied infectious disease surveillance.

3.2 Introduction

For the last decade, researchers have used Bayesian phylogeography (Lemey et al. 2009) to investigate the epidemiology of rapidly evolving viral pathogens with the aim of elucidating the contributions of discrete traits, often geographic location, to the propagation and persistence of disease outbreaks. Numerous examples are available in the literature and recent compelling studies have focused on recent Ebola (Dudas et al. 2017b), Zika (Grubaugh et al. 2017), West Nile (Swetnam et al. 2018) and influenza H3N2 (Magee, Suchard, and Scotch 2017), H9N2 (Yang et al. 2019) and H5N2 (Hicks et al. 2020) virus outbreaks. Bayesian phylogeographic discrete trait diffusion models require both a set of molecular sequences annotated with isolate sampling times and metadata describing a discrete traits of interest. Then, discrete trait diffusion is modeled as a continuous time Markov chain which evolves across a phylogenetic tree topology. Modeling discrete trait diffusion in this way enables computation of the model likelihood via Felsenstein's pruning algorithm. (Felsenstein 1981). Briefly, the algorithm proceeds via a post-order tree traversal and calculates the partial likelihood, backwards in time, for all trait states at internal tree nodes using the aforementioned Markov model. In a standard analysis, sequence records with discrete trait metadata are assumed to have a probability mass function (PMF) which assigns all mass to the observed trait. Concretely, the partial likelihood vectors at the tips are one-hot encoded as a vector with dimension equal to the cardinality of the discrete trait state space; the total number of distinct values a discrete trait may take.

For many researchers, the predominant method of obtaining publicly available molecular sequences for phylogeographic analysis is through the use of GenBank (Dennis A Benson et al. 2018), a nucleotide sequence database maintained by the National Center for Biotechnology Information or NCBI (Sayers et al. 2020). Usually, researchers parse the *country* field in a GenBank record in order to obtain geographic metadata for phylogeography studies. However, metadata representing geographic locations, host age and species, and other discrete characteristics are not required when submitting new molecular sequences to GenBank databases leading to numerous records with missing metadata. For example, previous work by Scotch et al. (Scotch et al. 2011) which linked virus sequence records to geographical entities in the GeoNames ontology (Vatant and Wick 2012) found that 80% of GenBank records contain "insufficient" geographic metadata. In this case, they defined geographic metadata insufficiency as data regarding the location of infected host (LOIH) at 1st-level administrative division (ADM1) or greater granularity. This means geographic metadata were typically informative for the LOIH at the state (province) or country level but seldom contained information on finer geographic entities such as counties or cities. Similarly, Tahsin et al. (Tahsin et al. 2014) reported the proportion of GenBank virus records with insufficient geographic data to be between 64% and 90%. Many real-world public health tasks require modeling transmission patterns at high geographic granularity to inform control strategies necessary to curb disease spread, such as modeling viral diffusion between counties within a state's boundary. Therefore, the insufficiency of GenBank metadata represents a major barrier to the implementation of virus phylogeography for applied public health surveillance.

This paucity of high resolution geographic metadata has inspired researchers to develop new methods and tools to ascertain the LOIH for viral sequences represented in GenBank records (Tahsin et al. 2014; Tahsin et al. 2017; Magge et al. 2018). Indeed, available pipelines for discerning the LOIH are configured such that they output not only the most probable location for a specific sequence, but also a vector of other possible locations along with their relative probabilities (Magge et al. 2018). Building on the availability of these new pipelines, Scotch et al. (Scotch et al. 2019) introduced the notion of incorporating sampling uncertainty into phylogeographic analyses. This parameterization of the standard discrete trait diffusion model involves assigning a prior PMF to the set of possible geographic locations for each tip with an uncertain LOIH. The additional uncertainty in LOIH is easily incorporated into the likelihood calculation using the standard pruning algorithm (Felsenstein 1981) by defining the partial likelihood vectors at the tips to be the desired PMF.

We note that since phylogeographic discrete trait diffusion models can be applied to general discrete traits, so too the phylogeographic uncertain trait model (UTM) introduced by Scotch et al. (Scotch et al. 2019) can be used to assign prior PMFs to tips missing arbitrary discrete trait information. In the case of non-geographic discrete traits, where relatively little attention has been paid to resolving insufficient metadata, this provides two distinct advantages to standard analysis workflows: it provides researchers with a coherent method of specifying *a priori* beliefs about unobserved traits and effectively increases the data set size by including sequences which would otherwise be excluded from an analysis due to missing metadata. Previously, phylogeographic researchers studying non-geographic discrete traits, such as host species or age, were left with two options for sequences with missing metadata: to manually curate locations for each unresolved record, or, to exclude these sequences from phylogeographic analysis (Magee and Scotch 2018; Dellicour et al. 2019). The former option is extremely labor intensive, difficult to replicate, and cannot be scaled to large data sets. Conversely, the latter has the disadvantage of reducing the amount of data included in a given phylogeographic analysis, which may induce biases in rate matrix parameters if the records with particular discrete traits are selectively over/underrepresented in the sample (De Maio et al. 2015);

Though phylogeographic discrete trait diffusion models remain a popular and promising tool for epidemiological inference, relatively few studies aim to quantify the statistical performance of these methods under various analysis conditions Lemey et al. 2014; De Maio et al. 2015; Magee, Suchard, and Scotch 2017; Magee and Scotch 2018. Particularly, phylogeographic discrete trait diffusion models are increasingly used for inference on large discrete state spaces and data set sizes, especially as pathogen genome sequencing continues to become a routine part of outbreak response. For example, recent studies commonly use state space sizes ranging from 10 to 56 discrete entities (Lemey et al. 2014; Magee, Suchard, and Scotch 2017; Dudas et al. 2017b). Paradoxically, a rigorous examination of model performance with respect to increasing state space and data set sizes is currently absent from the literature (Lemey et al. 2009; Lemey et al. 2014; De Maio et al. 2015; Magee, Suchard, and Scotch 2017). Further, given its recent introduction, the statistical performance of the phylogenetic UTM (Scotch et al. 2019) compared to other established model parameterizations is yet to be established. Since the quantity of interest from phylogeographic discrete trait diffusion models is often the most likely state at the root of the phylogeny, we select this *root state classification task* as the primary axis on which we evaluate model performance. In this paper, we take a simulation-based approach to investigating the performance of phylogeographic discrete trait diffusion models, paying special attention to the roles of data set and discrete state space size for performance on the root state classification task. Simultaneously, we compare the performance of the alternative phylogenetic UTM parameterizations against a reference model which omits sequences with missing metadata. This work represents, to our knowledge, a unique contribution to understanding the performance of popular phylogeographic discrete trait diffusion models under various analysis conditions and will be useful to researchers and public health practitioners tasked with designing phylogeographic studies using publicly available pathogen sequences.

3.3 Methods

3.3.1 Study Design

There are several ways in which the UTM can be implemented depending on the prior beliefs of the analyst for the missing discrete state values. For example, if no information *a priori* is available with respect to a discrete trait of interest with a molecular sequence, a reasonable choice may be to use a uniform prior over all possible trait values ("uniform"). On the other hand, it may be the case that a researcher wants to incorporate their prior beliefs on the relative probability of each state into the analysis. While this prior PMF can take many forms (indeed, there are infinitely many of them), we focused on two possibilities expected to arise frequently in practice: the researcher assigns most of the prior mass to the correct discrete trait ("informed“), or, alternatively, most of the mass is assigned to the incorrect state ("misspecified"). Concretely, for "informed“ models, we assigned 50% of the prior mass to the correct discrete trait, and divided the remaining mass uniformly across the remaining states.

Conversely, for “uninformed” models, we reverse the parameterization such that 50% of the prior mass is placed on an incorrect discrete state (chosen uniformly from the set of incorrect discrete traits) and the remaining mass distributed uniformly among the remaining traits. We believe these three options (uniform, informed, misspecified) are representative of choices likely to be made in practice. Prior to the introduction of the phylogenetic UTM (Scotch et al. 2019), researchers often exclude sequences with missing metadata from phylogeographic analyses. We specified this modeling approach (“drop“) as the reference to which we compared alternative UTM parameterizations.

We utilized a fully factorial, completely randomized design to quantify the relationships between discrete state space size, data set size and phylogeographic model performance. We defined 150, 250 and 500 sequences, respectively, as the factor levels for data set size. Similarly, we defined discrete state space sizes of 4, 8, and 16 states as factor levels for discrete state space size. We then simulated 25 replicate data sets under for each of 9 combinations of the aforementioned factor levels resulting in 225 data sets. We analyzed each data set using the phylogeographic UTM with either: i) informed ii) misspecified or iii) uniform prior PMFs. We also analyzed each data set after excluding sequences with missing metadata to serve as a reference for model comparison. Using this design, we analyzed 225 data sets under each of the 4 alternative model parameterization for a total of 900 independent model analyses. We discuss the data simulation procedure including generation of missing discrete traits in the following sections and provide a visual summary in Figure 12.

3.3.2 Data Simulation

3.3.2.1 Phylogenetic Trees

Since virus sequences represent isolates from individuals infected during epidemics, we simulated phylogenetic trees using the serially-sampled birth-death SIR model (SSBD-SIR) (Stadler et al. 2013). The SSBD-SIR model requires specification of 3 parameters: β, γ, ϕ representing the transmission (birth), recovery (death) and sampling rates, respectively. An equivalent specification can be made in terms of R0, the basic reproduction number, by using a fixed recovery rate. We selected simulation parameters to be similar to general, seasonal influenza outbreaks with an R0 value of 1.4 and assuming an infectious period (ϕ^{-1}) of one week, consistent with observed epidemiological patterns (Connolly 2005). Finally, we specified a sampling rate of 20%, reflecting a densely sampled epidemic scenario. We simulated trees until either 150, 250 or 500 tips were sampled. We performed tree simulation using the TreeSim package in R (Stadler 2011)

3.3.2.2 Sequence Data

We converted branch lengths to units of substitutions by assuming a strict molecular clock model with a rate of 1×10^{-3} substitutions per site, per year to allow for sequence simulation on each tree. We utilized an HKY + Γ model of nucleotide substitution with 4 rate categories, as is commonly used for modeling influenza molecular sequence evolution. We simulated 1750 base-pair (bp) sequences using the aforementioned parameters using Phyx (Brown, Walker, and Smith 2017).

3.3.2.3 Discrete and Missing Trait Data Simulation

We simulated the evolution of discrete traits on each phylogenetic tree by assuming traits evolved with a rate of 0.1 substitutions per site per year. Since a key goal of our study is to estimate the performance of phylogeographic trait models on a variety of state spaces, we simulated traits with 4, 8 or 16 states. We used random symmetric Markov matrices with gamma distributed rate parameters. To generate missing traits, we used a binomial sampling process on the observed traits where each trait is dropped with 20% probability. A final data set includes sequences written in FASTA format with discrete trait and sampling time information annotated in the description line.

3.3.3 Bayesian Phylogenetic and Phylogeographic Inference

We performed phylogenetic and phylogeographic inference using BEAST v 1.10.1 (Suchard et al. 2018) . We modeled molecular evolution using an HKY + Γ model with 4 rate categories, reflecting the conditions under which the data were simulated. We employ a flexible nonparametric skygrid prior since we know *a priori* that the population of infected individuals follow non-linear SIR-type dynamics. We specified a symmetric Markov model for inference of discrete trait evolution, again driven by our choice of data simulation conditions. To estimate divergence times, we fixed tip dates as the dates of sampling recorded during each simulation. We ran the each MCMC for 100 million iterations, sampling every 10,000 steps and removed the first 20% as burn-in. We diagnosed convergence of the MCMC procedure using Tracer v1.7.1 (Rambaut et al. 2018b) checking that all model parameters had Effective Sample Sizes (ESS) of 200 or greater.

3.3.4 Model Evaluation

Inferring the most likely state at the root of the phylogeny, akin to identifying the location or host species where an outbreak started, is a key output of phylogeographic discrete trait diffusion models. We can evaluate the performance of popular phylogeographic techniques by treating the root state identification problem as a *classification* problem, borrowing terms from the machine learning literature. Several metrics are available to summarize the a classification model with respect to its performance on a classification task. Given a classification model and labeled test data one can compute the *accuracy* of a classification model: the proportion of instances it classifies correctly. In phylogeography, a central task is to correctly classify the most likely state at the root of a phylogeny. We recorded the root state from which each simulation was initialized and calculated the accuracy of phylogeographic models when given more data (sequences) or when performing inference over increasing discrete state spaces. Since the result of our Bayesian phylogeographic analysis is a posterior distribution over root states, we follow standard practice in classification model evaluation and selected the most likely posterior state j as the root state "prediction" output by our models.

$$\hat{j} = \max_{j} \mathcal{P}(X = j|\theta)$$

Though informative, accuracy does not fully describe the characteristics of a given classification model. A common measure of classification model performance is the cross entropy. This is generally interpreted at the number of bits needed to transmit data from a source distribution when using a model of that distribution. In the context of classification model evaluation, we can interpret cross entropy as a kind of "distance" between the posterior distribution estimated by our model and the true root state distributions. Defining the true root state distribution P_j as a one-hot encoded vector permits computation of the cross entropy using:

$$C = -\sum_{j \in J} P_j \log \mathcal{P}(X = j|\theta)$$

Another useful metric which measures the efficiency of classification models is the Kullback-Leibler (KL) divergence. Here, it represents the amount of information we gain about the distribution of the root state by using our model output relative to our *a priori* assumptions. We defined our prior P_j as a uniform distribution over all possible root states. Then, the KL divergence was calculated using the posterior distribution over root states $\mathcal{P}(X = j|\theta)$ output by our phylogeographic models.

$$KL = \sum_{j \in J} P_j \log P_j - \log \mathcal{P}(X = j|\theta)$$

For each combination of simulation parameters, we recorded the state at the root of the phylogeny and calculated the root state accuracy, cross entropy and KL divergence to measure the performance of the standard and uncertain phylogenetic discrete trait models. We analyzed the impact of model parameterization, data set size and discrete state space size on model performance metrics using ANOVA.

3.3.5 Data Availability

We provide the simulated data as analysis-ready BEAST XML files along with files containing the parameters associated with each data set.

3.3.6 Factors Influencing Model Accuracy

We modeled root state classification accuracy for our 900 models using logistic regression by defining factors related to phylogeographic study design choices as predictors. For these analyses, we set the reference levels of each factor variables to be: i) 4 discrete states ii) the “drop” model design (where sequences with missing metadata are excluded) and iii) 150 molecular sequences, respectively.

3.4 Results

3.4.1 Phylogeographic Models show Strong Performance on Moderately Sized Data Sets

In Figure 13, we show the mean and 95% confidence intervals for each of the non-reference level factors included in our analysis. Using the standard interpretation of the odds ratio, we show that increased discrete state space sizes are associated with weaker model performance with respect to root state classification (Figure 13), p-values < 0.01). We found that, for our analysis, increasing data set size does not significantly improve phylogeographic root classification performance (13). Interestingly, we see increased performance at for models with 250 sequences, relative to other data set sizes, as shown by the positive odds ratios associated with these models (Figure 13, p-values < 0.05). Overall, we find no significant effects of model implementation method on root state classification performance.

3.4.2 Phylogeographic Information Gain Increases with State Space and Data Set Size

In the phylogeographic context, KL divergence is often used to quantify the amount of information gained from an analysis with respect to a prior distribution. Concretely, we are interested in quantifying the amount of information that the root state posterior contains relative to a uniform (uninformative) prior over all possible traits. We performed this calculation such that our KL divergence is expressed in units of bits; representing the total amount of information gained by an analyst from performing phylogeographic analysis to identify the root state. For many empirical analyses, since the true root state (and any root states of internal nodes) are unknown *a priori*, it is unclear how information gain is related to model accuracy and if increased information gain translates directly to improved classification performance. By including KL divergence as a predictor in our logistic regression analysis, we were able to infer the respective relationship between this metric and model accuracy. We found that KL divergence was not associated with root state classification performance (Figure 13, p-value: 0.248). In Figure 14 we present the mean Kullback-Leibler (KL) divergence (from a uniform prior) for 25 model replicates stratified by model implementation method as well as state space and data set size. Using ANOVA, we find that information gain and discrete state space size were significantly related to KL divergence and that KL divergence tended to increase along with discrete state space size ($p < 0.001$, F-score: 255.84, Table 13). Further, we also found KL divergence to be significantly associated with data set size ($p < 0.001$, F-score: 255.84, Table 13). We visualize the results of this analysis and show interactions between various design factors in Figure 15. We utilized Tukey's HSD post-hoc test to identify that this effect is primarily driven by the increase in information gain occurring when increasing data set sizes from 250 to 500 tips ($p < 0.001$). Echoing the results of our logistic regression analysis, we find no significant differences in information gain between model implementation methods (p-value: 0.866, F-score: 0.242, Table 13). However, for models with large discrete state spaces, we observed a leveling off in KL divergence with increasing data set size (Figure 15) indicating a functional limit to the information a phylogeographic model can extract about the root state given sufficient data. We also found significant interaction effects between state space size and data set size (Table 13, $p = 9.98 \times 10^{-1}$).

3.4.3 Phylogeographic Cross Entropy increases with Discrete State Space Size

By casting the phylogeographic root state inference problem in a classification framework, we gain access several established metrics for use in quantifying classification model performance. We select cross-entropy due to its usage in a wide variety of substantive areas. In Figure 16, we show the cross entropy mean and and 95% confidence interval stratified by model implementation method as well as state space and data set size. We observed that cross entropy tends to increase with the size of the state space; this is intuitive since the complexity of the classification task is related to the size of the state space. In Figure 17, we show the interaction plots between design factors and cross entropy, noting that cross entropy tended to increase with data set size which we confirmed using ANOVA (Table 12). Again, we employ Tukey's HSD post-hoc testing to show that the sequences ($p < 0.001$). This is congruent with the results obtained from our logistic regression analysis which indicates that models fit to intermediate data set sizes tend to perform better than models with larger data set sizes. Since we generally expect classification model performance to generally increase with data set size, we offer an explanation of the apparent increase in model complexity arising from increasing data set sizes. We expect that classification performance diminished on larger data set sizes since phylogeographic classification models perform trait state estimation for all $n-1$ internal nodes before making a final classification for the root trait state; if any errors are made at intermediate nodes, these errors are propagated back toward the root.

3.5 Discussion

3.5.1 Performance of Phylogeographic Models for Root State Classification

Pathogen molecular sequence data are being created at an unprecedented rate. So too has interest increased in methods and tools which leverage this new data stream for public health application. Examples in the literature include evaluating the impact of hypothetical interventions on epidemic spread (Dellicour et al. 2018) as well as identifying specific groups or locations responsible for driving epidemic spread (Lemey et al. 2009; Lemey et al. 2014; Dudas et al. 2017b; Grubaugh et al. 2017; Magee, Suchard, and Scotch 2017; Swetnam et al. 2018). With increasing metadata availability, the resolution with which phylogeographic analyses are performed is increasing (Dudas et al. 2017b; Grubaugh et al. 2017; Dellicour et al. 2018). Concretely, this translates into specification of models with large number of discrete states and sequences. Reconstructing epidemiological patterns of infectious disease spread using pathogen genomes is often achieved by modeling the epidemiological trait of interest as a continuous time Markov chain which evolves across a phylogeny. The ability of these models to accurately reconstruct these traits of interest is paramount to their use in applied public health settings for modeling infectious disease outbreaks.

In this paper, we took a simulation-based approach and quantified the role of discrete state space and sequence data set sizes on the root state classification performance of modern phylogeographic models. We focused specifically on root state classification since, in infectious disease epidemiology, this task is analogous to identifying the discrete trait (i.e. geographic location, host species, etc.) associated with the origin of an outbreak. We simulated 225 data sets which we then analyzed using standard and uncertain phylogeographic discrete trait diffusion models. For the uncertain trait models, we performed analyses using three distinct prior specifications: i) a uniform prior across states ii) an informed prior which assigns most of the prior mass to the correct state and iii) a misspecified prior, which assigns most of its mass to an incorrect state. We compared characteristics of each model's MCC phylogeny to characterize model performance on the root state classification task. We found no significant differences between model implementation methods and model performance, suggesting that while the phylogeographic UTM does not substantially increase or decrease model performance. Therefore, it remains an attractive alternative for researchers wanting to include sequences with missing metadata in their analyses. Interestingly, a misspecified prior for the tip trait states did not seem to substantially effect root state predictive accuracy. We expect this is similarly due to errors in the state at each node being propagated back through the phylogenetic tree during inference. We expect that while the tip prior misspecification may influence the classification error at proximal internal nodes, as the model is applied backward in time toward the root, the partial likelihood vector begins to resemble the stationary distribution of the associated Markov model. This is especially likely for fast evolving traits, since the total evolutionary time for the model is the sum of all branch lengths across the tree.

Though phylogeographic models are popular epidemiological tools in an era of pathogen genomes aplenty, relatively few studies have characterized the performance of these methods under various analysis conditions (Lemey et al. 2014; De Maio et al. 2015; Magee and Scotch 2018; Scotch et al. 2019). Indeed, much of this previous work is concerned with empirical analyses of virus sequence data sets (Magee, Suchard, and Scotch 2017; Magee and Scotch 2018; Scotch et al. 2019) and often compares model root state posterior probabilities as a proxy for performance. The informativeness of the analysis is then typically assessed by calculating the Kullback-Leibler (KL) divergence between the root state prior and posterior distributions (Lemey et al. 2014; Magee, Suchard, and Scotch 2017; Magee and Scotch 2018). Work by de Maio et al (De Maio et al. 2015) established performance characteristics for several phylogeographic models using 200 tips and either two to eight discrete states, while focusing on the role of migration rates and sampling bias on inference quality. In contrast, we focus on the combined roles of data set and discrete state space sizes and how they impact discrete trait diffusion model inference.

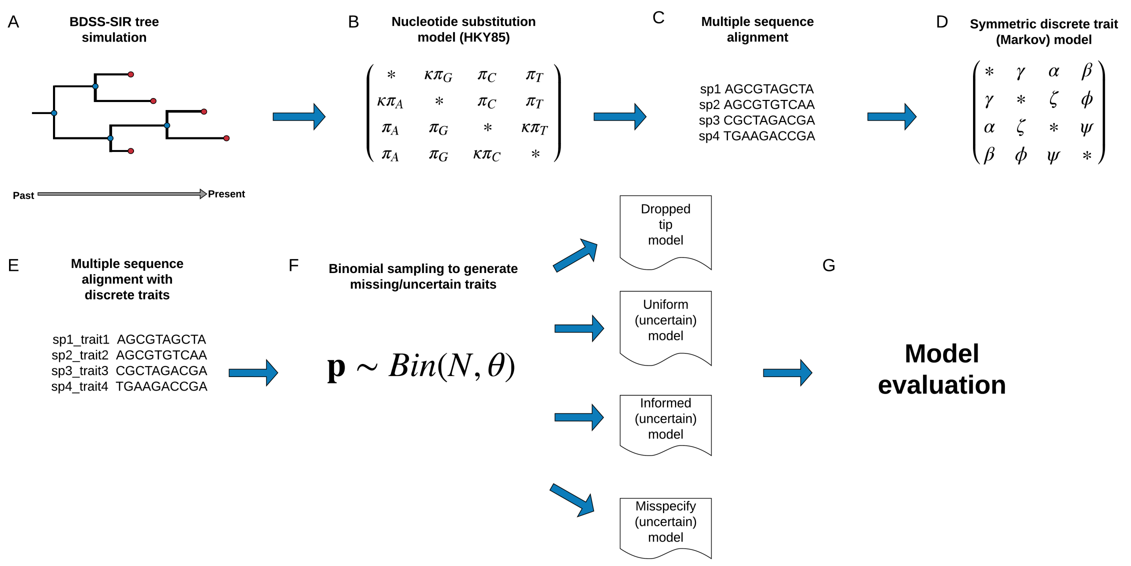
We found that KL divergence is significantly positively associated with both discrete state space and data set size. This suggests caution when relying on this metric as it may erroneously suggest more granular discretization schemes or reward more data intensive models though this may not translate to increased performance on the root classification task. For example, though Scotch et al. (Scotch et al. 2019) found that the phylogeogephic UTM improved performance relative to other popular heuristics, these conclusions were based on a empirical comparison between model posteriors. Additionally, we show that KL divergence was not predictive of root state classification performance suggesting more informative models may still ultimately produce incorrect results. So, while empirical studies are informative for assessing congruence between root state inferences drawn by different phylogeographic methods, they are not informative with respect to the absolute accuracy (i.e. classification) of these methods for root state inference. We also find that root classification performance is the best at intermediate data set sizes. We believe that our models show poorer performance on larger data sets since as data set size increases, the number of internal nodes for which trait reconstruction must occur also increases. We expect that any errors in internal node classification (that is, internal node distributions which assign the most mass to an incorrect trait state) are propagated back toward the root. However, this could also be influenced by uncertainty in the phylogenetic tree topology. Changes in fast evolving traits, such as host age or geography during disease outbreaks, will be sensitive to uncertainty in branch lengths since as time increases, the partial likelihood vectors at internal nodes begin to more closely resemble the stationary distribution of their evolutionary Markov models. Since tree space is known to grow factorially (Felsenstein and Felenstein 2004) with respect to tip number, it is likely that a combination of posterior tree uncertainty, mediated through the effect of increasing tip numbers, also impacted our results. Following this line of reasoning, we expect that increases in molecular sequence length will improve model performance since increasing the data available to models (via including more sites independently evolving across a tree) will reduce tree topological uncertainty.

3.5.2 Limitations and Future Work

Phylogeographic discrete trait diffusion models have emerged as the primary statistical tool for analyzing pathogen genomes annotated with discrete trait metadata. Given the increasing interest in the application of genome sequencing for public health outbreak response, it is prudent to establish the performance of phylogeographic models on different size data sets. This is of direct interest to public health practitioners who may be tasked with designing molecular epidemiological studies within budgetary, computational or data constraints. Overall, this study aimed to evaluate the performance of popular phylogeographic models under various analysis conditions, focusing on the roles of discrete state space and data set size on phylogeographic model performance. While we find that model performance is significantly increased at intermediate data set sizes, our results paint suggest caution when relying solely on KL divergence and other metrics calculated from purely empirical studies. However, our study is not without limitations. We limited our simulation study to discrete traits simulated from symmetric Markov rate matrices. This represents the simplest of the phylogeograhic models; we focused on this case to estimate a baseline for model performance. In reality, there are several ways in which trait states models are specified and inferred. Of particular note is the use of Bayesian Stochastic Search Variable Selection (BSSVS) which augments the model state space such that each instantaneous rate parameter r_{ij} is multiplied by a binomial random variable whose value represents the inclusion (or conversely, exclusion) of a given rate parameter in the matrix. The BSSVS parameterization effectively reduces the number of estimated transition rates which may lead to increased model performance on root state classification. Another popular approach for parameterizing discrete trait diffusion models is to model each transition rate as linear combination of covariates of interest. This reduces the problem of estimating transition rates to estimating the coefficients of the resulting generalized linear model (GLM). Clearly, our results do not extend to these parameterization methods. Finally, we quantified model performance with respect to root state classification only. It may be the case that the UTM increases classification performance on intermediate nodes in the phylogeny and that phylogeographic methods in general perform better on inferring the discrete states for proximal ancestral nodes. Quantifying the *treewide* classification performance of phylogeographic models under various conditions remains an open area of research.

3.6 Acknowledgements

This research was supported by a grant from the National Library of Medicine of the National Institutes of Health under award number R01LM012080 to MS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would also like to thank ASU Biomedical Informatics for providing computational resources to perform phylogenetic experiments. We would also like to thank the NVIDIA GPU Grant program for providing GPU hardware which was used to accelerate phylogenetic experiments.


A
BDSS-SIR tree simulation
Past
Present
B
Nucleotide substitution model (HKY85)
C
Multiple sequence alignment
sp1 AGCGTAGCTA
sp2 AGCGTGTCAA
sp3 CGCTAGACGA
sp4 TGAAGACCGA
D
Symmetric discrete trait (Markov) model
E
Multiple sequence alignment with discrete traits
sp1_trait1 AGCGTAGCTA
sp2_trait2 AGCGTGTCAA
sp3_trait3 CGCTAGACGA
sp4_trait4 TGAAGACCGA
F
Binomial sampling to generate missing/uncertain traits
$\mathbf{p} \sim Bin(N, \theta)$
Dropped tip model
Uniform (uncertain) model
Informed (uncertain) model
Misspecify (uncertain) model
G
Model evaluation

$$\begin{pmatrix} * & \kappa\pi_G & \pi_C & \pi_T \\ \kappa\pi_A & * & \pi_C & \pi_T \\ \pi_A & \pi_G & * & \kappa\pi_T \\ \pi_A & \pi_G & \kappa\pi_C & * \end{pmatrix}$$

$$\begin{pmatrix} * & \gamma & \alpha & \beta \\ \gamma & * & \zeta & \phi \\ \alpha & \zeta & * & \psi \\ \beta & \phi & \psi & * \end{pmatrix}$$

Figure 12.

Figure 12. Visual summary of data simulation procedure. We simulated phylogenetic trees under the serially sampled birth-death SIR model using an R0 of 1.4 and an infectious period of 7 days. We simulated molecular sequence evolution on each tree topology using an HKY85 model and a strict molecular clock with a rate of 1×10^{-3} substitutions per site, per year. We also simulated discrete traits on each tree topology, using symmetric Markov rate matrices with rate parameters drawn from a gamma distribution and a strict molecular clock with a rate of 0.1 substitutions per site, per year. This results in a set of molecular sequences annotated with discrete traits and sampling time information. We simulated missing traits using a binomial sampling process for each tip, indicating the presence, or conversely, the absence of discrete trait metadata. Finally, each data set was analyzed using one of four phylogeographic model parameterizations.

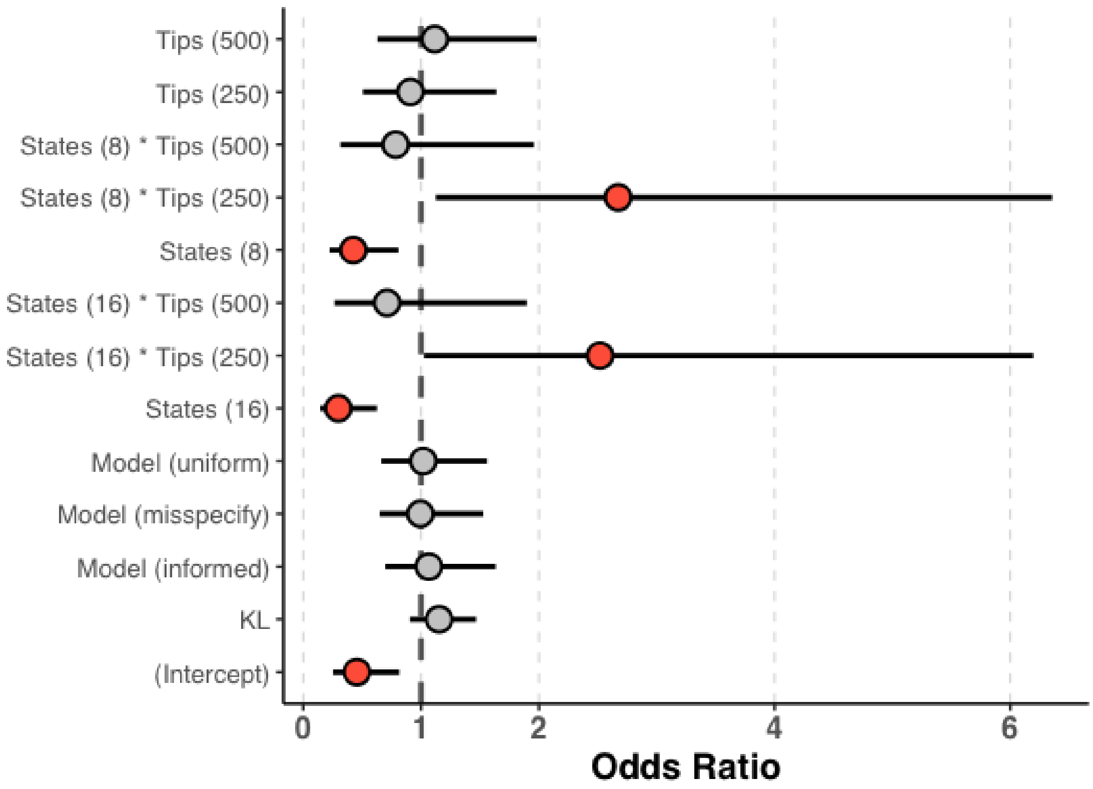


Figure 13.

Odds ratios show the effect of design factors on model accuracy. We used logistic regression analysis to estimate the effects of design choices on phylogeographic model accuracy. For the purposes of analysis, we defined our reference factor levels to be 4 state, 150 sequence and drop model design, respectively. We show the factor found to be significant as red points, where grey points represent insignificant factors. Our analysis shows that relative to this reference level that increasing discrete state space size reduces the root state classification accuracy of phylogeographic models. We find that, independently, data set size and implementation method have no significant effects on model accuracy. However, our analysis shows increased root state classification performance for models with 250 sequences, suggesting that phylogeographic models may perform most favorably at intermediate data set sizes.

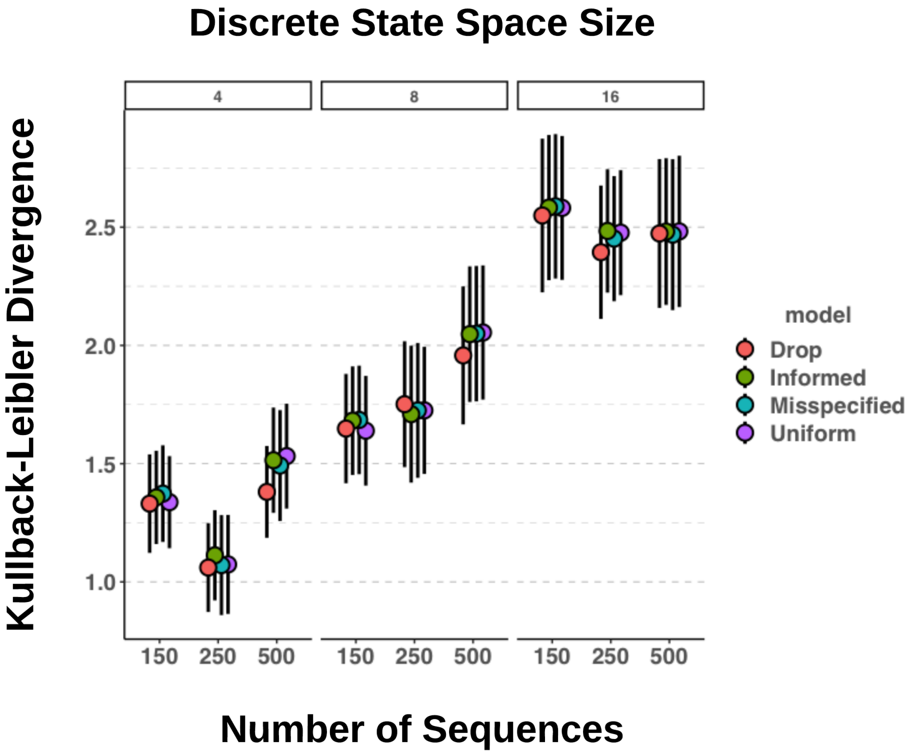


Figure 14.

Comparison of Kullback Leibler (KL) divergence stratified by model design factor . Here, we show the mean and 95% confidence intervals for KL divergence arranged by increasing data set and discrete state space size. We observed an upward trend in information gain associated with both increasing discrete state space and data set sizes. We confirmed the presence of this trend using ANOVA (Table 13). As suspected, ANOVA suggests no statistically significant differences in posterior information gains between various model implementation heuristics (Table

13). We observed a tendency for information gain to increase when increasing data set size from 250 to 500 sequences.

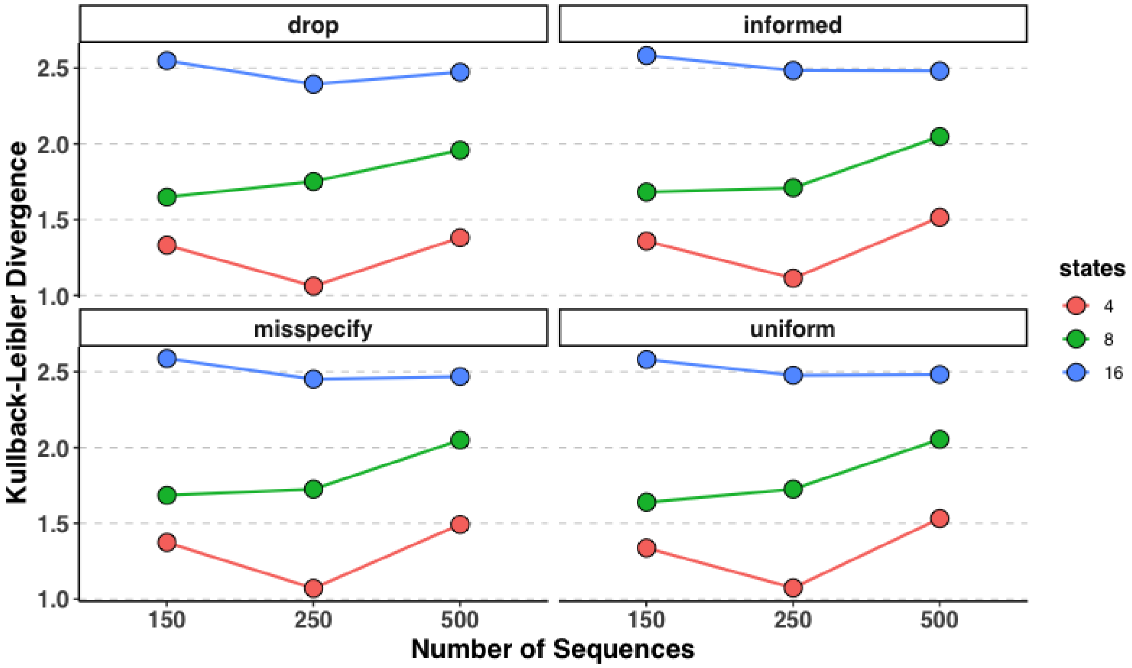


Figure 15.

Interaction effects between model design factors and information gain. We show that estimated mean KL divergence tended to increase when increasing the data set size from 250 to 500 sequences and that this effect was generally consistent across model implementations. From this perspective, it is further illustrated that we find information gain tended to increase with discrete state space size.

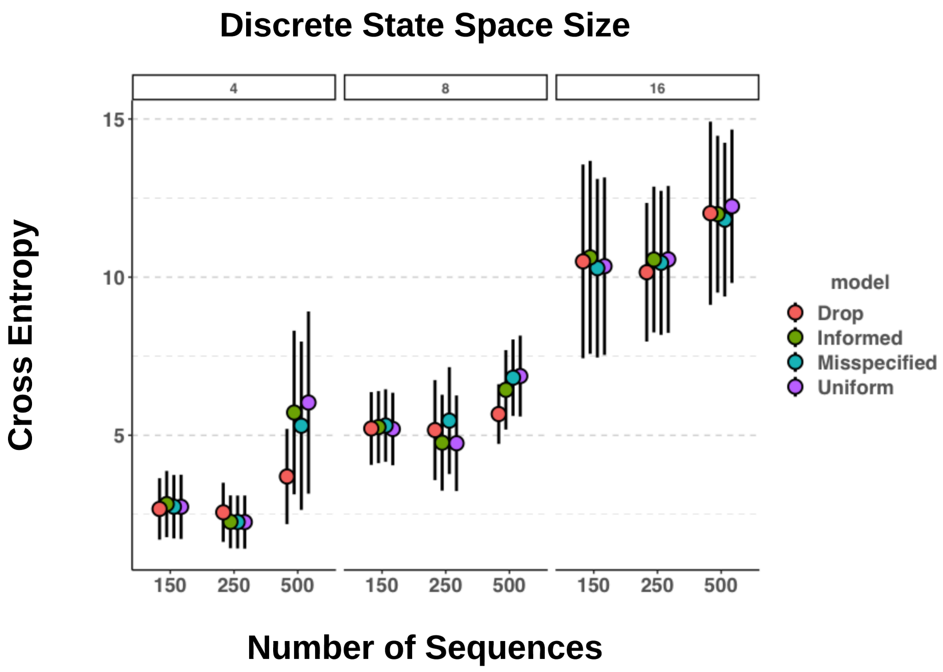


Figure 16.

Comparison of Cross Entropy stratified by model design factor.
We present the mean and 95% CIs of the cross-entropy stratified by data set and discrete state space size. Similar to KL divergence (Figure 14), we show that cross entropy tends to increase with discrete state space size. This is expected since the classification problem becomes more challenging as the total number of states increases. We also find that cross entropy tends to increase with data set size. This could be due to phylogegraphic the fact that phylogeographic root state classification first requires the model infer the discrete state probabilities at all $n - 1$ intermediate tree nodes. We expect that inference for an increasing number of internal tree nodes similarly increases the difficulty of the phylogeographic root state classification task.

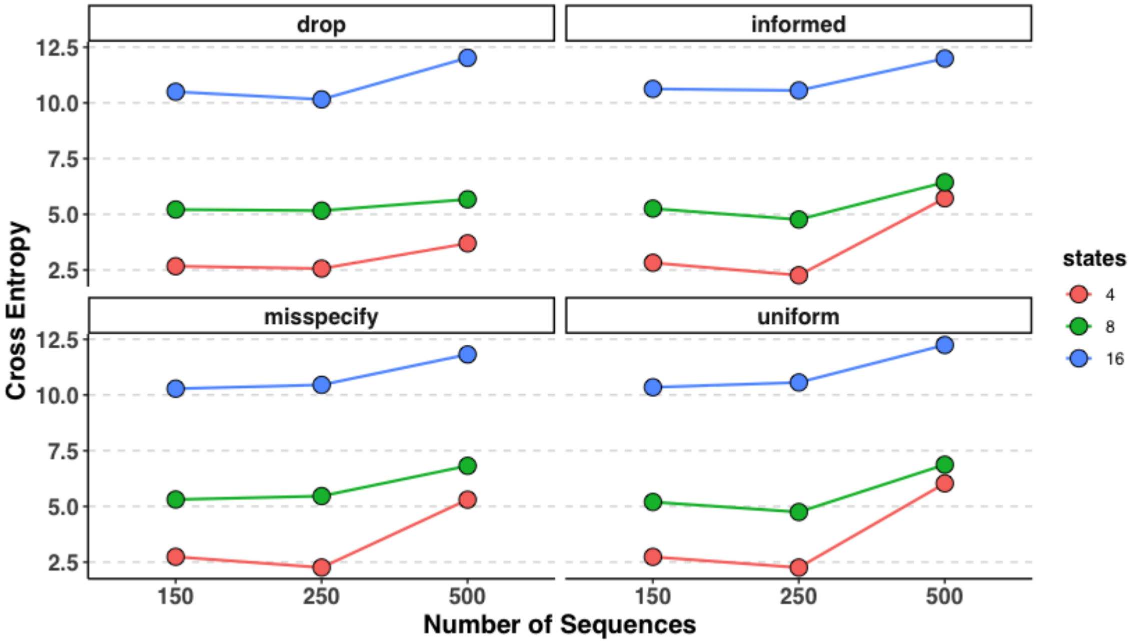


Figure 17.

Interaction effects between cross entropy and model design factors. By visualizing the interaction between each model design factor, we can observe that cross entropy remains relatively consistent between models with 150 and 250 sequences. However, it sharply increases significantly when models increase from 250 to 500 sequences (Tukey's HSD post-hoc analysis, p $= 4.2 \times 10^{-6}$) similar to the trends observed with KL divergence.

Table 12. Analysis of Variance: Cross Entropy

Factor	Deg. Freedom	Sum Sq	Mean Sq	F-value	p-value
Model	3	19	6	0.27	0.846
Tips	2	667	333	14.498	**6.36** $\times 10^{-7}$
States	2	8900	4450	193.481	$<$ **2** $\times \mathbf{10}^{-16}$
Tips * States	4	70	18	0.762	0.550
Residual	888	20147	23	-	-

Table 13. Analysis of Variance: Kullback-Leibler Divergence

Factor	Deg. Freedom	Sum Sq	Mean Sq	F-value	p-value
Model	3	0.3	6	0.27	0.867
Tips	2	8.2	4.09	14.498	**6.39** $\times 10^{-5}$
States	2	214.3	107.16	255.884	$<$ **2** $\times 10^{-16}$
Tips * States	4	7.8	1.95	4.662	**9.98** $\times 10^{-4}$
Residual	888	366.9	0.42	-	-

Chapter 4

BUILD-A-BEAST: A PIPELINE FOR PRODUCING BEAST XML DOCUMENTS

4.1 Introduction

Bayesian phylogeographic discrete trait diffusion models have emerged as powerful data analysis tools for pathogen genomes annotated with sampling time and discrete trait metadata. Many extensions to these models exist, and in particular we have focused on the utility of modeling "Markov Jumps" which record the relative magnitude and number of transitions between discrete traits in a given phylogeographic model. In Chapters 2 & 3, we transformed raw phylogeographic model output to into actionable information via the calculation of "phylogenetic relative risks" which compute the relative probabilities that a particular discrete trait, such as a geographic location or particular species, served as a sink or source of infections. Further, we showed that "phylogenetic relative risk" analyses tend to produce hypotheses similar to other popular epidemiological modeling approaches. In the case that discrete traits are not available for sequences of interest, we showed in Chapter 4 that implementing the phylogeographic UTM is a viable alternative to other popular heuristics for assigning missing discrete trait data and that this additional uncertainty does not substantially influence model performance. Unfortunately, while BEAUTi offers the capability to record the total number of Markov Jumps, as of version 1.10.X, this does not include the capability to record jumps counts between each pairwise combination of traits specified in a model. Enumerating these pairwise jumps is central to computing phylogenetic relative risks as introduced and utilized in Chapters 2 & 3. To do so, researchers must manually manipulate BEAST XML files to implement this functionality. This represent a large barrier-to-entry for using these models in public health practice and for making these types of analyses available in public health informatics applications such as ZooPhy. In this Chapter, we introduce **Build-A-BEAST**, a set of extensible python classes designed for BEAST XML modification. We then describe our implementations of classes for generating BEAST XML with Markov Jumps and the phylogenetic UTM specification.

4.2 Program Requirements

The most recent version of this software can be obtained from Github at https://github.com/matteo-V/Build-A-BEAST. This repository contains a README, code and example XML files that can be used with the pipeline. Build-A-BEAST is written using python v3.X and utilized the built-in packages "argparse", "xml", and "sys". There is one external requirement for the "pandas" package. We list instructions for installing this package in the README and additionally make a virtual environment available which packages the necessary dependencies alongside a compatible python implementation.

4.3 System Architecture

While researchers often implement new phylogeographic models in popular software software packages like BEAST, there is often a delay in the release of software tools which eases the burden of specifying these models manually. For example, though the generalized linear model parameterization of the standard phylogeographic model was

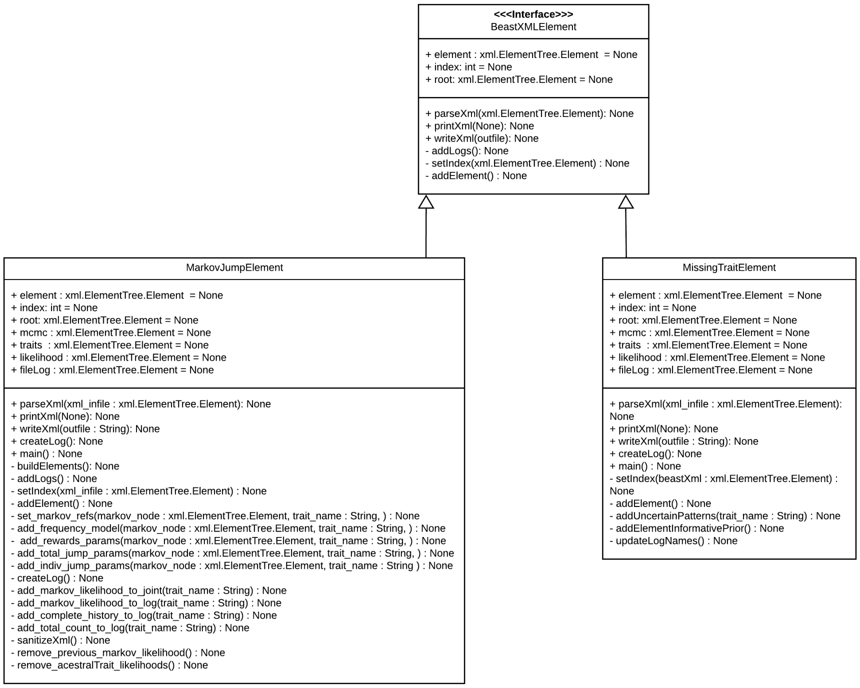

<<<Interface>>>
BeastXMLElement
+ element : xml.ElementTree.Element = None
+ index: int = None
+ root: xml.ElementTree.Element = None
+ parseXml(xml.ElementTree.Element): None
+ printXml(None): None
+ writeXml(outfile): None
- addLogs(): None
- setIndex(xml.ElementTree.Element) : None
- addElement() : None
MarkovJumpElement
+ element : xml.ElementTree.Element = None
+ index: int = None
+ root: xml.ElementTree.Element = None
+ mcmc : xml.ElementTree.Element = None
+ traits : xml.ElementTree.Element = None
+ likelihood : xml.ElementTree.Element = None
+ fileLog : xml.ElementTree.Element = None
+ parseXml(xml_infile : xml.ElementTree.Element): None
+ printXml(None): None
+ writeXml(outfile : String): None
+ createLog(): None
+ main() : None
- buildElements(): None
- addLogs() : None
- setIndex(xml_infile : xml.ElementTree.Element) : None
- addElement() : None
- set_markov_refs(markov_node : xml.ElementTree.Element, trait_name : String,) : None
- add_frequency_model(markov_node : xml.ElementTree.Element, trait_name : String,) : None
- add_rewards_params(markov_node : xml.ElementTree.Element, trait_name : String,) : None
- add_total_jump_params(markov_node : xml.ElementTree.Element, trait_name : String,) : None
- add_indiv_jump_params(markov_node : xml.ElementTree.Element, trait_name : String) : None
- createLog() : None
- add_markov_likelihood_to_joint(trait_name : String) : None
- add_markov_likelihood_to_log(trait_name : String) : None
- add_complete_history_to_log(trait_name : String) : None
- add_total_count_to_log(trait_name : String) : None
- sanitizeXml() : None
- remove_previous_markov_likelihood() : None
- remove_acestralTrait_likelihoods() : None
MissingTraitElement
+ element : xml.ElementTree.Element = None
+ index: int = None
+ root: xml.ElementTree.Element = None
+ mcmc : xml.ElementTree.Element = None
+ traits : xml.ElementTree.Element = None
+ likelihood : xml.ElementTree.Element = None
+ fileLog : xml.ElementTree.Element = None
+ parseXml(xml_infile : xml.ElementTree.Element): None
+ printXml(None): None
+ writeXml(outfile : String): None
+ createLog(): None
+ main() : None
- setIndex(beastXml : xml.ElementTree.Element) : None
- addElement() : None
- addUncertainPatterns(trait_name : String) : None
- addElementInformativePrior() : None
- updateLogNames() : None

Figure 18. **Build-A-BEAST Class Architecture**.
Here we show the UML class diagram describing the architecture of our BEAST XML modification system. We based our implementation around a unified interface which defines the public methods for all BEAST XML modifications. We define classes which implement this interface to generate Markov Jump and uncertain trait model (UTM) specifications in an input BEAST XLM file.

introduced in 2014 (Lemey et al. 2014), specification of this model was not available in BEAUTi software until v.1.10.X released (Suchard et al. 2018). Since BEAST XML is highly idiosyncratic, manual modification of these files is typically only possible phylogeographic researchers. Further, we expect that manual modifications will not scale to models with large data set and discrete state space sizes. Therefore, a lack of a simple, extensible framework to implement these modifications represents a strong barrier to rapid adoption and implementation of these methods. We provide an overview of the architecture of our object-oriented system below.

4.3.1 Program Inputs

Build-A-BEAST is designed using a pipeline architecture which applied a linear chain of transformation to a BEAST XML document and output another BEAST XML document with the specified modifications. As such, it requires as input a BEAUTi-generated BEAST XML document with a user-specified discrete trait diffusion model. Any discrete trait diffusion model available within BEAUTi is compatible with this pipeline.

4.4 Program Execution

We designed Build-A-BEAST to be both a command line tool and set of classes and interfaces which developers can use to implement their own custom BEAST XML modification pipelines. We demonstrate general use cases and discuss command line options for our two implemented components below.

```
usage: MarkovJumpElement.py [--help] --infile BEASTXML [--outfile OUTFILE]

optional arguments:
  -h, --help            show this help message and exit
  -i BEASTXML, --infile BEASTXML
                        BEAST XML file
  -o OUTFILE, --outfile OUTFILE
                        modified BEAST XML location. If none provided, XML

usage: MissingTraitElement.py [--help] --infile BEASTXML [--outfile OUTFIL
```

```
optional arguments:
  -h, --help            show this help message and exit
  -i BEASTXML, --infile BEASTXML
                        BEAST XML file
  -o OUTFILE, --outfile OUTFILE
                         modified BEAST XML location. If none provided, XML
  --apriori             BEAST XML contains prior information about missing
```

For all use cases and tools, we expect that the user will supply the path to a valid BEAST XML file with a discrete trait diffusion model specified. Any valid paramerizations is compatible with this pipeline, including the poA user may optionally specify a path (outfile) where the modified BEAST XML will be saved. If the user does not enter a path for the output XML file, a prompt will be issued asking the user if they want to overwrite the XML at the current path location.

4.5 Conclusion

I developed this program to automate the currently idiosyncratic, tedious nature of manipulating BEAST XML files to implement both total Markov jump counts and the phylogenetic UTM. As previosuly mentioned, these functions are not currently supported in the most recent version of BEAUti. Ideally, these tools will enable researchers to easily produce these files for investigating viral diffusion between discrete groups of interest. Although the program is built to handle several anticipated errors, it is possible that more will be discovered by users. In this event, users are encouraged to report any perceived bugs or issues to GitHub. To my knowledge, this framework produces files compliant with BEAST v1.10.X standards. Users are encouraged to modify, extend and utilize this framework for their applications.

Chapter 5

DISCUSSION

5.1 Summary of Chapters

The purpose of this dissertation was to develop analytical tools and to expand the conceptual framework for learning salient features RNA viral disease dynamics from a set of molecular sequences. While myriad methods exist for reconstructing patterns of discrete trait evolution along a phylogeny, few of these studies attempt to translate the results of phyloegographic analyses into actionable metrics that can be used by public health agencies to direct the development of interventions aimed at reducing pathogen spread. In Chapter 1, I developed a comprehensible, epidemiologically-relevant metric, the phylogenetic risk ratio and applied it to studying age associated diffusion of influenza A/H3N2 during the 2016-2017 US epidemic. The results show distinct regional age associated transmission patterns and that nationally, transmission tended to occur within similar age groups. By modelling age associated diffusion as a continuous time Markov chain, parameterized by a generalized linear model (GLM), I show a strong association between age-specific hemagglutinin (HA) immune imprinting and the inferred rates of age associated diffusion, echoing results reported previously in the literaure. As expected, immune impriting with homosubtypic HA was associated with reduced rates of age associated diffusion, which we interpret as either i) reduced transmission from imprinted individuals or ii) reduced disease severity (and . Under both the partial and full protection hypotheses, which suggest that prior exposure to homosubtypic HA antigens either: i) confers partial immune protection that prevents severe disease ii) confers total protective immunity against a specific HA subtype. Currently, results from several studies suggest that immune imprinting confers partial immune protection. In the context of phylogeographic models, both partial and full immune protection would be associated with reduced rates of age associated diffusion, so the results are plausible. Also supported for inclusion in the phylogeographic models was age-specific population density and we show a trend toward inclusion of social contact patterns, though we fail to detect a statistically significant effect. Though these predictors are commonly invoked when discussing determinants of influenza epidemiology, there have been few attempts to quantify the roles of these factors on directly shaping observed disease dynamical patters within a phylogenetic framework. Therefore, this study represents a major conceptual leap in considering age as a discrete trait in phylogenetic models, as well as proposes methods for translating output of phylogeographic analyses into actionable information for to public health agencies tasked with influenza disease surveillance and control.

Encouraged by these results, in Chapter 2, I then applied the phylogenetic risk ratio and Bayesian phylogeographic GLMs to study ecologic factors underlying the rapid geographic dissemination of WNV in the US. I collected and annotated nearly 1000 WNV genomes with sampling location (at state-level resolution) and host species excluding genomes collected from human and non-Culicine mosquito hosts. Then, I modeled the geographic diffusion of WNV in the US as a log-linear combination of bird host density and diversity across two epochs: early epidemic expansion (1999-2003) and endemic (2003-current) periods. I obtained bird density and diversity data for all species represented in the sequence data set from the National Audobon Society's Christmas Bird Count and aggregated predictor data into Health and Human Services (HHS regions). The results indicate that during early epidemic expansion period, Passerine and total bird diversity were broadly supportive of WNV dissemination while Non-Passerine diversity was negatively associated with WNV geographic dissemination, suggesting a dilution type effect. During the endemic period, there was a similar role for Passerine diversity in supporting geographic dissemination. However, total bird diversity was associated with reduced WNV dissemination during this period, indicating a prominent role for silution effects in shaping WNV dissemination in this epoch. To uncover the patterns of host-associated diffusion of WNV after its introduction to the US, I reconstructed the patterns of host associated diffusion and calculated phylogenetic relative risks to determine which hosts were responsible for a majority of transmissions during early WNV epidemic expansion. The results shows frequent and rapid WNV dissemination between Passerine and Non-Passerine hosts during the epidemic expansion period; implying that bird-to-bird transmission was an important feature of early epidemic expansion of WNV in the US.

Though Chapters 1 & 2, I demonstrate how epidemiologically-relevant discrete traits (other than geographic location) can be used in genomic epidemiology studies to provide actionable public health information in the form of the phylogenetic risk ratio. However, for many non-geographic traits of interest, metadata in publicly available molecular sequence databases is quite sparse. Indeed, several otherwise suitable sequences were excluded from phylogeographic analysis due to missing metadata for discrete traits of epidemiological relevance. While newer techniques for incorporating trait uncertainty into discrete phylogenetic models exist: the phylogenetic uncertain trait model (UTM), a rigorous evaluation of the statistical performance of these methods relative to standard parameterizations was lacking. In Chapter 3, I perform a simulation study of the standard and uncertain trait models and characterize their statistical performace with respect to the root state classification task. Here, the goal was to characterize the performance of the phylogeographic UTM. My results showing that model parameterization does not necessarily increase (nor does it decrease) classification performance with respect to root state identification in the phylogeographic context. Therefore, we expect the phylogentic UTM to enable both highly granular genomic epidemiology studies of viral molecular sequences as well as increasing molecular data set sizes for sequences with sparse metadata reporting for epidemiological traits of interest.

In Chapter 4, I discuss the software tools made available to implement these methods since manual manipulation of BEAST XML is idiosyncratic, error-prone, and difficult to scale for studies examining discrete traits with large state spaces. The PRR framework may be used to address complex epidemiological questions and enjoys efficient algorithms for computing labeled transitions in evolutionary models. To ensure the technical complexity of implementing these methods is not a barrier to their wider use in public health surveillance, it is prudent to develop software and visualization tools which automate these processes. Although there is a tutorial provided which explicity details the XML modifications necessary to implement Markov jumps (on which the PRR is based), this process is extremely tedious, error-prone and cumbersome for discrete traits with large state spaces. Therefore, in Chapter 4, I introduced a pipeline that may facilitate expanded use of the PRR and phylogenetic UTM by other researchers, public health departments and within public health informatics applications designed to simplify the use of viral phylogeography for viral surveillance (Scotch, Magge, and Vaiente 2019). The pipeline that I created and have made public (https://github.com/matteo-v/Build-A-BEAST), allows individuals to simply pass a BEAST XML file, the desired output path, and specify the desired options using a command line prompt.

5.2 Future Research

There are several opportunities to extend and apply the research presented in this dissertation. First and foremost, the phylogenetic risk ratio (PRR) (which was introduced in Chapter 1) represents a powerful analytical tool which can be used to translate the results of genomic epidemiological analyses of viral pathogens into actionable public health information. For example, a promising application of the PRR would be to quantify the age-associated diffusion of SARS-CoV-2 in the US. The results of this study could then be used to focus public health resources and intervention efforts to reduce transmission types with PRR greater than one. Increasingly, public health and governmental agencies rely on the output of infectious disease models to guide policy decisions and recommendations aimed at improving infectious disease outcomes. During the global SARS-CoV-2 pandemic, genomic epidemiological approaches have received renewed interest. Currently, these efforts are concentrated around estimating dates of introduction (i.e divergence time dating) of SARS-CoV-2 into certain localities across the US. Extending these efforts to include phylogenetic risk estimation for particular age, sociodemographic, and other groups is a particularly promising area. Additionally, systematic efforts to disseminate and implement genomic epidemiological methods into routine public health surveillance activities are paradoxically absent from the literature. We envision studies which systematically apply tested implementation frameworks as a necessary next step in translation of these methods. Under the RE-AIM implementation framework (Glasgow, Vogt, and Boles 1999), quantifying reach, the proportion of the total population that receives benefit from an intervention and effectiveness (reduction of disease due to application of an intervention) are important steps for translating evidence into practice. These measurements must necessarily be made in real world contexts. Thus, future work is needed to quantify the effectiveness of molecular epidemiological methods in for quantifying risk and informing control efforts in such settings.

NOTES

REFERENCES

Allan, Brian F, R Brian Langerhans, Wade A Ryberg, William J Landesman, Nicholas W Griffin, Rachael S Katz, Brad J Oberle, Michele R Schutzenhofer, Kristina N Smyth, Annabelle de St Maurice, et al. 2009. “Ecological correlates of risk and incidence of West Nile virus in the United States.” *Oecologia* 158 (4): 699–708.

APHL. 2013. “Influenza virologic surveillance right size roadmap.” July. https://www.aphl.org/programs/infectious_disease/influenza/Influenza-Virologic-Surveillance-Right-Size-Roadmap/Pages/default.aspx.

Apolloni, Andrea, Chiara Poletto, and Vittoria Colizza. 2013. “Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic.” *BMC infectious diseases* 13 (1): 176.

Arevalo, Philip, Huong Q. McLean, Edward A. Belongia, and Sarah Cobey. 2019. “Earliest infections predict the age distribution of seasonal influenza A cases.” *medRxiv.* doi:10.1101/19001875. eprint: https://www.medrxiv.org/content/early/2019/09/08/19001875.full.pdf.

Arregui, Sergio, Alberto Aleta, Joaquın Sanz, and Yamir Moreno. 2018. “Projecting social contact matrices to different demographic structures.” *PLoS computational biology* 14 (12): e1006638.

Bahl, Justin, Martha I Nelson, Kwok H Chan, Rubing Chen, Dhanasekaran Vijaykrishna, Rebecca A Halpin, Timothy B Stockwell, Xudong Lin, David E Wentworth, Elodie Ghedin, et al. 2011. “Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans.” *Proceedings of the National Academy of Sciences* 108 (48): 19359–19364.

Bahl, Justin, Truc T Pham, Nichola J Hill, Islam TM Hussein, Eric J Ma, Bernard C Easterday, Rebecca A Halpin, Timothy B Stockwell, David E Wentworth, Ghazi Kayali, et al. 2016. “Ecosystem interactions underlie the spread of avian influenza a viruses with pandemic potential.” *PLoS pathogens* 12 (5): e1005620.

Basta, Nicole E, Dennis L Chao, M Elizabeth Halloran, Laura Matrajt, and Ira M Longini Jr. 2009. “Strategies for pandemic and seasonal influenza vaccination of schoolchildren in the United States.” *American journal of epidemiology* 170 (6): 679–686.

Bedford, Trevor, Sarah Cobey, Peter Beerli, and Mercedes Pascual. 2010. “Global migration dynamics underlie evolution and persistence of human influenza A (H3N2).” *PLoS pathogens* 6 (5): e1000918.

Begon, Michael. 2008. “Effects of host diversity on disease dynamics.” *Infectious disease ecology: effects of ecosystems on disease and of disease on ecosystems:* 12–29.

Begon, Michael, Malcolm Bennett, Roger G Bowers, Nigel P French, SM Hazel, and Joseph Turner. 2002. “A clarification of transmission terms in host-microparasite models: numbers, densities and areas.” *Epidemiology & Infection* 129 (1): 147–153.

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2012. “GenBank.” *Nucleic Acids Research* 41, no. D1 (November): D36–D42. doi:10.1093/nar/gks1195. eprint: http://oup.prod.sis.lan/nar/article-pdf/41/D1/D36/3680750/gks1195.pdf.

Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D Pruitt, and Eric W Sayers. 2018. “GenBank.” *Nucleic acids research* 46 (D1): D41–D47.

Bielejec, Filip, Philippe Lemey, Guy Baele, Andrew Rambaut, and Marc A Suchard. 2014. “Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography.” *Systematic biology* 63 (4): 493–504.

Bock, Carl E, and Terry L Root. 1981. “The Christmas bird count and avian ecology.” *Studies in Avian Biology* 6:17–23.

Brown, Joseph W, Joseph F Walker, and Stephen A Smith. 2017. “Phyx: phylogenetic tools for unix.” *Bioinformatics* 33 (12): 1886–1888.

Brownson, Ross C, Jonathan E Fielding, and Christopher M Maylahn. 2009. “Evidence-based public health: a fundamental concept for public health practice.” *Annual review of public health* 30:175–201.

Charrel, RN, AC Brault, P Gallian, J-J Lemasson, B Murgue, S Murri, B Pastorino, H Zeller, R De Chesse, P De Micco, et al. 2003. “Evolutionary relationship between Old World West Nile virus strains: evidence for viral gene flow between Africa, the Middle East, and Europe.” *Virology* 315 (2): 381–388.

Clay, Christine A, Erin M Lehmer, Stephen St Jeor, and M Denise Dearing. 2009. “Testing mechanisms of the dilution effect: deer mice encounter rates, Sin Nombre virus prevalence and species diversity.” *EcoHealth* 6 (2): 250–259.

Connolly, Máire A. 2005. *Communicable disease control in emergencies: a field manual.* World health organization.

Darriba, Diego, Guillermo L Taboada, Ramón Doallo, and David Posada. 2012. “jModelTest 2: more models, new heuristics and parallel computing.” *Nature methods* 9 (8): 772.

De Luca, Giancarlo, Kim Van Kerckhove, Pietro Coletti, Chiara Poletto, Nathalie Bossuyt, Niel Hens, and Vittoria Colizza. 2018. “The impact of regular school closure on seasonal influenza epidemics: a data-driven spatial transmission model for Belgium.” *BMC infectious diseases* 18 (1): 29.

De Maio, Nicola, Chieh-Hsi Wu, Kathleen M O’Reilly, and Daniel Wilson. 2015. “New routes to phylogeography: a Bayesian structured coalescent approximation.” *PLoS genetics* 11 (8): e1005421.

Dellicour, Simon, Guy Baele, Gytis Dudas, Nuno R Faria, Oliver G Pybus, Marc A Suchard, Andrew Rambaut, and Philippe Lemey. 2018. “Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak.” *Nature communications* 9 (1): 2222.

Dellicour, Simon, Sebastian Lequime, Bram Vrancken, Mandev S Gill, Paul Bastide, Karthik Gangavarapu, Nate Matteson, Yi Tan, Louis du Plessis, Alexander A Fisher, et al. 2019. “Phylogeographic and phylodynamic approaches to epidemiological hypothesis testing.” *bioRxiv:* 788059.

Deng, Xianding, Wei Gu, Scot Federman, Louis du Plessis, Oliver G Pybus, Nuno Faria, Candace Wang, Guixia Yu, Brian Bushnell, Chao-Yang Pan, et al. 2020. “Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California.” *Science.*

Di Giallonardo, Francesca, Jemma L Geoghegan, Douglas E Docherty, Robert G McLean, Michael C Zody, James Qu, Xiao Yang, Bruce W Birren, Christine M Malboeuf, Ruchi M Newman, et al. 2016. “Fluid spatial dynamics of West Nile virus in the United States: rapid spread in a permissive host environment.” *Journal of virology* 90 (2): 862–872.

Driessche, Pauline van den. 2017. “Reproduction numbers of infectious disease models.” *Infectious Disease Modelling* 2 (3): 288–303.

Drummond, Alexei J, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. 2006. "Relaxed phylogenetics and dating with confidence." *PLoS biology* 4 (5): e88.

Drummond, Alexei J, Oliver G Pybus, Andrew Rambaut, Roald Forsberg, and Allen G Rodrigo. 2003. "Measurably evolving populations." *Trends in ecology & evolution* 18 (9): 481–488.

Dudas, Gytis, Luiz Max Carvalho, Trevor Bedford, Andrew J Tatem, Guy Baele, Nuno R Faria, Daniel J Park, Jason T Ladner, Armando Arias, Danny Asogun, et al. 2017a. "Virus genomes reveal factors that spread and sustained the Ebola epidemic." *Nature* 544 (7650): 309.

———. 2017b. "Virus genomes reveal factors that spread and sustained the Ebola epidemic." *Nature* 544 (7650): 309.

Dudas, Gytis, Luiz Max Carvalho, Andrew Rambaut, and Trevor Bedford. 2018. "MERS-CoV spillover at the camel-human interface." *Elife* 7:e31257.

Duggal, Nisha K, Angela Bosco-Lauth, Richard A Bowen, Sarah S Wheeler, William K Reisen, Todd A Felix, Brian R Mann, Hannah Romo, Daniele M Swetnam, Alan DT Barrett, et al. 2014. "Evidence for co-evolution of West Nile Virus and house sparrows in North America." *PLoS neglected tropical diseases* 8 (10).

Ezenwa, Vanessa O, Marvin S Godsey, Raymond J King, and Stephen C Guptill. 2005. "Avian diversity and West Nile virus: testing associations between biodiversity and infectious disease risk." *Proceedings of the Royal Society B: Biological Sciences* 273 (1582): 109–117.

Fauver, Joseph R, Mary E Petrone, Emma B Hodcroft, Kayoko Shioda, Hanna Y Ehrlich, Alexander G Watts, Chantal BF Vogels, Anderson F Brito, Tara Alpert, Anthony Muyombwe, et al. 2020. "Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States." *Cell.*

Felsenstein, Joseph. 1981. "Evolutionary trees from DNA sequences: a maximum likelihood approach." *Journal of molecular evolution* 17 (6): 368–376.

Felsenstein, Joseph, and Joseph Felenstein. 2004. *Inferring phylogenies.* Vol. 2. Sinauer associates Sunderland, MA.

Fourment, Mathieu, and Edward C Holmes. 2016. "Seqotron: a user-friendly sequence editor for Mac OS X." *BMC research notes* 9 (1): 106.

George, Edward I, and Robert E McCulloch. 1997. “Approaches for Bayesian variable selection.” *Statistica sinica:* 339–373.

Glasgow, Russell E, Thomas M Vogt, and Shawn M Boles. 1999. “Evaluating the public health impact of health promotion interventions: the RE-AIM framework.” *American journal of public health* 89 (9): 1322–1327.

Glass, Laura M, and Robert J Glass. 2008. “Social contact networks for the spread of pandemic influenza in children and teenagers.” *BMC public health* 8 (1): 61.

Goldstein, Edward, Hieu H Nguyen, Patrick Liu, Cecile Viboud, Claudia A Steiner, Colin J Worby, and Marc Lipsitch. 2017. “On the relative role of different age groups during epidemics associated with respiratory syncytial virus.” *The Journal of infectious diseases* 217 (2): 238–244.

Gostic, Katelyn M., Rebecca Bridge, Shane Brady, Cécile Viboud, Michael Worobey, and James O. Lloyd-Smith. 2019. “Childhood immune imprinting to influenza A shapes birth year-specific risk during seasonal H1N1 and H3N2 epidemics.” *PLOS Pathogens* 15, no. 12 (December): 1–20. doi:10.1371/journal.ppat.1008109.

Gostic, Katelyn M, Monique Ambrose, Michael Worobey, and James O Lloyd-Smith. 2016. “Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting.” *Science* 354 (6313): 722–726.

Grubaugh, Nathan D, Jason T Ladner, Moritz UG Kraemer, Gytis Dudas, Amanda L Tan, Karthik Gangavarapu, Michael R Wiley, Stephen White, Julien Thézé, Diogo M Magnani, et al. 2017. “Genomic epidemiology reveals multiple introductions of Zika virus into the United States.” *Nature* 546 (7658): 401.

Gwinn, Marta, Duncan R MacCannell, and Rima F Khabbaz. 2017. “Integrating advanced molecular technologies into public health.” *Journal of clinical microbiology* 55 (3): 703–714.

Hadfield, James, Anderson F Brito, Daniele M Swetnam, Chantal BF Vogels, Ryan E Tokarz, Kristian G Andersen, Ryan C Smith, Trevor Bedford, and Nathan D Grubaugh. 2019. “Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain.” *PLoS pathogens* 15 (10).

Hamer, Gabriel L, Uriel D Kitron, Tony L Goldberg, Jeffrey D Brawn, Scott R Loss, Marilyn O Ruiz, Daniel B Hayes, and Edward D Walker. 2009. “Host selection by Culex pipiens mosquitoes and West Nile virus amplification.” *The American journal of tropical medicine and hygiene* 80 (2): 268–278.

Hicks, Joseph T, Dong-Hun Lee, Venkata R Duvuuri, Mia Kim Torchetti, David E Swayne, and Justin Bahl. 2020. “Agricultural and geographic factors shaped the North American 2015 highly pathogenic avian influenza H5N2 outbreak.” *PLoS pathogens* 16 (1): e1007857.

Holmes, Edward C, and Bryan T Grenfell. 2009. “Discovering the phylodynamics of RNA viruses.” *PLoS computational biology* 5 (10): e1000505.

Kain, Morgan P, and Benjamin M Bolker. 2019. “Predicting West Nile virus transmission in North American bird communities using phylogenetic mixed effects models and eBird citizen science data.” *Parasites & vectors* 12 (1): 395.

Kass, Robert E, and Adrian E Raftery. 1995. “Bayes factors.” *Journal of the american statistical association* 90 (430): 773–795.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.” *Nucleic acids research* 30 (14): 3059–3066.

Keeling, Matt J., and Pejman Rohani. 2008. *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press.

Kilpatrick, A Marm, Laura D Kramer, Matthew J Jones, Peter P Marra, and Peter Daszak. 2006. “West Nile virus epidemics in North America are driven by shifts in mosquito feeding behavior.” *PLoS Biol* 4 (4): e82.

Komar, Nicholas. 2003. “West Nile virus: epidemiology and ecology in North America.” *Advances in virus research* 61:185–234.

Komar, Nicholas, Stanley Langevin, Steven Hinten, Nicole Nemeth, Eric Edwards, Danielle Hettler, Brent Davis, Richard Bowen, and Michel Bunning. 2003. “Experimental infection of North American birds with the New York 1999 strain of West Nile virus.” *Emerging infectious diseases* 9 (3): 311.

Komar, Nicholas, Nicholas A Panella, Joseph E Burns, Stephen W Dusza, Tina M Mascarenhas, and Thomas O Talbot. 2001. “Serologic evidence for West Nile virus infection in birds in the New York City vicinity during an outbreak in 1999.” *Emerging Infectious Diseases* 7 (4): 621.

Komar, Nicholas, Nicholas A Panella, Stanley A Langevin, Aaron C Brault, Manuel Amador, Eric Edwards, and Jennifer C Owen. 2005. “Avian hosts for West Nile virus in St. Tammany Parish, Louisiana, 2002.” *The American journal of tropical medicine and hygiene* 73 (6): 1031–1037.

Lanciotti, RS, JT Roehrig, V Deubel, J Smith, M Parker, K Steele, B Crise, KE Volpe, MB Crabtree, JH Scherret, et al. 1999. "Origin of the West Nile virus responsible for an outbreak of encephalitis in the northeastern United States." *Science* 286 (5448): 2333–2337.

Lee, Nelson, Paul KS Chan, David SC Hui, Timothy H Rainer, Eric Wong, Kin-Wing Choi, Grace CY Lui, Bonnie CK Wong, Rita YK Wong, Wai-Yip Lam, et al. 2009. "Viral loads and duration of viral shedding in adult patients hospitalized with influenza." *The Journal of infectious diseases* 200 (4): 492–500.

Lemey, Philippe, Andrew Rambaut, Trevor Bedford, Nuno Faria, Filip Bielejec, Guy Baele, Colin A Russell, Derek J Smith, Oliver G Pybus, Dirk Brockmann, et al. 2014. "Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2." *PLoS pathogens* 10 (2).

Lemey, Philippe, Andrew Rambaut, Alexei J Drummond, and Marc A Suchard. 2009. "Bayesian phylogeography finds its roots." *PLoS computational biology* 5 (9).

Levine, Rebecca S, David L Hedeen, Meghan W Hedeen, Gabriel L Hamer, Daniel G Mead, and Uriel D Kitron. 2017. "Avian species diversity and transmission of West Nile virus in Atlanta, Georgia." *Parasites & vectors* 10 (1): 62.

Loss, Scott R, Gabriel L Hamer, Edward D Walker, Marilyn O Ruiz, Tony L Goldberg, Uriel D Kitron, and Jeffrey D Brawn. 2009. "Avian host community structure and prevalence of West Nile virus in Chicago, Illinois." *Oecologia* 159 (2): 415–424.

Magee, Daniel, Rachel Beard, Marc A Suchard, Philippe Lemey, and Matthew Scotch. 2015. "Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion." *Archives of virology* 160 (1): 215–224.

Magee, Daniel, and Matthew Scotch. 2018. "The effects of random taxa sampling schemes in Bayesian virus phylogeography." *Infection, Genetics and Evolution* 64:225–230.

Magee, Daniel, Marc A Suchard, and Matthew Scotch. 2017. "Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction." *PLoS computational biology* 13 (2): e1005389.

Magge, Arjun, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. "Deep neural networks and distant supervision for geographic location mention extraction." *Bioinformatics* 34 (13): i565–i573.

Malkinson, Mertyn, Caroline Banet, Yoram Weisman, Shimon Pokamunski, Roni King, et al. 2002. “Introduction of West Nile virus in the Middle East by migrating white storks.” *Emerging infectious diseases* 8 (4): 392.

Marm Kilpatrick, A, Peter Daszak, Matthew J Jones, Peter P Marra, and Laura D Kramer. 2006. “Host heterogeneity dominates West Nile virus transmission.” *Proceedings of the Royal Society B: Biological Sciences* 273 (1599): 2327–2333.

Marra, Peter P, Sean Griffing, Carolee Caffrey, Marm A Kilpatrick, Robert McLean, Christopher Brand, EMI Saito, Alan P Dupuis, Laura Kramer, and Robert Novak. 2004. “West Nile virus and wildlife.” *BioScience* 54 (5): 393–402.

May, Fiona J, C Todd Davis, Robert B Tesh, and Alan DT Barrett. 2011. “Phylogeography of West Nile virus: from the cradle of evolution in Africa to Eurasia, Australia, and the Americas.” *Journal of virology* 85 (6): 2964–2974.

Mbah, Martial L Ndeffo, Jan Medlock, Lauren Ancel Meyers, Alison P Galvani, and Jeffrey P Townsend. 2013. “Optimal targeting of seasonal influenza vaccination toward younger ages is robust to parameter uncertainty.” *Vaccine* 31 (30): 3079–3089.

McLean, Robert G. 2006. “West Nile virus in North American birds.” *Ornithological Monographs:* 44–64.

Medlock, Jan, and Alison P Galvani. 2009. “Optimizing influenza vaccine distribution.” *Science* 325 (5948): 1705–1708.

Miller, Ezer, and Amit Huppert. 2013. “The effects of host diversity on vector-borne disease: the conditions under which diversity will amplify or dilute the disease risk.” *PLoS One* 8 (11): e80279.

Minin, Vladimir N, Erik W Bloomquist, and Marc A Suchard. 2008. “Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics.” *Molecular biology and evolution* 25 (7): 1459–1471.

Minin, Vladimir N, and Marc A Suchard. 2008. “Counting labeled transitions in continuous-time Markov models of evolution.” *Journal of mathematical biology* 56 (3): 391–412.

Molaei, Goudarz, Theodore G Andreadis, Philip M Armstrong, John F Anderson, and Charles R Vossbrinck. 2006. “Host feeding patterns of Culex mosquitoes and West Nile virus transmission, northeastern United States.” *Emerging infectious diseases* 12 (3): 468.

Mossong, Joël, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, Jacco Wallinga, et al. 2008a. "Social contacts and mixing patterns relevant to the spread of infectious diseases." *PLoS medicine* 5 (3): e74.

———. 2008b. "Social contacts and mixing patterns relevant to the spread of infectious diseases." *PLoS medicine* 5 (3): e74.

Nelms, Brittany M, Ethan Fechter-Leggett, Brian D Carroll, Paula Macedo, Susanne Kluh, and William K Reisen. 2013. "Experimental and natural vertical transmission of West Nile virus by California Culex (Diptera: Culicidae) mosquitoes." *Journal of medical entomology* 50 (2): 371–378.

Nemeth, Nicole, Ginger Young, Christina Ndaluka, Helle Bielefeldt-Ohmann, Nicholas Komar, and Richard Bowen. 2009. "Persistent West Nile virus infection in the house sparrow (Passer domesticus)." *Archives of virology* 154 (5): 783–789.

Oster, Alexandra M, Joel O Wertheim, Angela L Hernandez, M Cheryl Bañez Ocfemia, Neeraja Saduvala, and H Irene Hall. 2015. "Using molecular HIV surveillance data to understand transmission between subpopulations in the United States." *Journal of acquired immune deficiency syndromes (1999)* 70 (4): 444.

Ostfeld, Richard S, and Felicia Keesing. 2000. "Biodiversity and disease risk: the case of Lyme disease." *Conservation biology* 14 (3): 722–728.

Parker, Joe, Andrew Rambaut, and Oliver G Pybus. 2008. "Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty." *Infection, Genetics and Evolution* 8 (3): 239–246.

Poon, Art FY, Réka Gustafson, Patricia Daly, Laura Zerr, S Ellen Demlow, Jason Wong, Conan K Woods, Robert S Hogg, Mel Krajden, David Moore, et al. 2016. "Near real-time monitoring of HIV transmission hotspots from routine HIV genotyping: an implementation case study." *The lancet HIV* 3 (5): e231–e238.

Prem, Kiesha, Alex R Cook, and Mark Jit. 2017. "Projecting social contact matrices in 152 countries using contact surveys and demographic data." *PLoS computational biology* 13 (9): e1005697.

Putri, Wayan CWS, David J Muscatello, Melissa S Stockwell, and Anthony T Newall. 2018. "Economic burden of seasonal influenza in the United States." *Vaccine* 36 (27): 3960–3966.

Pybus, Oliver G, Marc A Suchard, Philippe Lemey, Flavien J Bernardin, Andrew Rambaut, Forrest W Crawford, Rebecca R Gray, Nimalan Arinaminpathy, Susan L Stramer, Michael P Busch, et al. 2012. “Unifying the spatial epidemiology and molecular evolution of emerging epidemics.” *Proceedings of the National Academy of Sciences* 109 (37): 15066–15071.

Rambaut, Andrew, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. 2018a. “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.” *Systematic Biology* 67, no. 5 (April): 901–904. doi:10.1093/sysbio/syy032. eprint: http://oup.prod.sis.lan/sysbio/article-pdf/67/5/901/25517397/syy032.pdf.

———. 2018b. “Posterior summarization in Bayesian phylogenetics using Tracer 1.7.” *Systematic biology* 67 (5): 901.

Ranjeva, Sylvia, Rahul Subramanian, Vicky J Fang, Gabriel M Leung, Dennis KM Ip, Ranawaka APM Perera, JS Malik Peiris, Benjamin J Cowling, and Sarah Cobey. 2019. “Age-specific differences in the dynamics of protective immunity to influenza.” *Nature communications* 10 (1): 1660.

Reisen, William K. 2013. “Ecology of west nile virus in North America.” *Viruses* 5 (9): 2079–2105.

Reisen, William, Hugh Lothrop, Robert Chiles, Minoo Madon, Cynthia Cossen, Leslie Woods, Stan Husted, Vicki Kramer, and John Edman. 2004. “West nile virus in california.” *Emerging infectious diseases* 10 (8): 1369.

Rizzoli, Annapaola, Luca Bolzoni, Elizabeth A Chadwick, Gioia Capelli, Fabrizio Montarsi, Michela Grisenti, Josue Martınez de la Puente, Joaquin Muñoz, Jordi Figuerola, Ramon Soriguer, et al. 2015. “Understanding West Nile virus ecology in Europe: Culex pipiens host feeding preference in a hotspot of virus emergence.” *Parasites & vectors* 8 (1): 213.

Sayers, Eric W, Jeff Beck, J Rodney Brister, Evan E Bolton, Kathi Canese, Donald C Comeau, Kathryn Funk, Anne Ketter, Sunghwan Kim, Avi Kimchi, et al. 2020. “Database resources of the national center for biotechnology information.” *Nucleic acids research* 48 (D1): D9.

Schanzer, Dena, Julie Vachon, and Louise Pelletier. 2011. “Age-specific differences in influenza A epidemic curves: do children drive the spread of influenza epidemics?” *American journal of Epidemiology* 174 (1): 109–117.

Scotch, Matthew, Arjun Magge, and Matteo Vaiente. 2019. “ZooPhy: A bioinformatics pipeline for virus phylogeography and surveillance.” *Online Journal of Public Health Informatics* 11 (1).

Scotch, Matthew, Changjiang Mei, Cynthia Brandt, Indra Neil Sarkar, and Kei Cheung. 2010. “At the intersection of public-health informatics and bioinformatics: using advanced Web technologies for phylogeography.” *Epidemiology (Cambridge, Mass.)* 21 (6): 764.

Scotch, Matthew, Indra Neil Sarkar, Changjiang Mei, Robert Leaman, Kei-Hoi Cheung, Pierina Ortiz, Ashutosh Singraur, and Graciela Gonzalez. 2011. “Enhancing phylogeography by improving geographical information from GenBank.” *Journal of biomedical informatics* 44:S44–S47.

Scotch, Matthew, Tasnia Tahsin, Davy Weissenbacher, Karen O’Connor, Arjun Magge, Matteo Vaiente, Marc A Suchard, and Graciela Gonzalez-Hernandez. 2019. “Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography.” Vey043, *Virus Evolution* 5, no. 1 (February). doi:10.1093/ve/vey043. eprint: https://academic.oup.com/ve/article-pdf/5/1/vey043/27995874/vey043.pdf.

Simpson, Jennifer E, Paul J Hurtado, Jan Medlock, Goudarz Molaei, Theodore G Andreadis, Alison P Galvani, and Maria A Diuk-Wasser. 2011. “Vector host-feeding preferences drive transmission of multi-host pathogens: West Nile virus as a model system.” *Proceedings of the Royal Society B: Biological Sciences* 279 (1730): 925–933.

Smith, Derek J, Stephanie Forrest, David H Ackley, and Alan S Perelson. 1999. “Variable efficacy of repeated annual influenza vaccination.” *Proceedings of the National Academy of Sciences* 96 (24): 14001–14006.

Spellerberg, Ian F, and Peter J Fedor. 2003. “A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the ‘Shannon–Wiener’Index.” *Global ecology and biogeography* 12 (3): 177–179.

Stadler, Tanja. 2011. “Simulating trees with a fixed number of extant species.” *Systematic biology* 60 (5): 676–684.

Stadler, Tanja, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. 2013. “Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV).” *Proceedings of the National Academy of Sciences* 110 (1): 228–233.

Suchard, Marc A, Philippe Lemey, Guy Baele, Daniel L Ayres, Alexei J Drummond, and Andrew Rambaut. 2018. “Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.” *Virus Evolution* 4 (1): vey016.

Sullivan, Brian L, Christopher L Wood, Marshall J Iliff, Rick E Bonney, Daniel Fink, and Steve Kelling. 2009. “eBird: A citizen-based bird observation network in the biological sciences.” *Biological conservation* 142 (10): 2282–2292.

Swaddle, John P, and Stavros E Calos. 2008. “Increased avian diversity is associated with lower incidence of human West Nile infection: observation of the dilution effect.” *PloS one* 3 (6).

Swetnam, Daniele, Steven G Widen, Thomas G Wood, Martin Reyna, Lauren Wilkerson, Mustapha Debboun, Dreda A Symonds, Daniel G Mead, Barry J Beaty, Hilda Guzman, et al. 2018. “Terrestrial bird migration and West Nile virus circulation, United States.” *Emerging infectious diseases* 24 (12): 2184.

Tahsin, Tasnia, Rachel Beard, Robert Rivera, Rob Lauder, Garrick Wallstrom, Matthew Scotch, and Graciela Gonzalez. 2014. “Natural language processing methods for enhancing geographic metadata for phylogeography of zoonotic viruses.” *AMIA Summits on Translational Science Proceedings* 2014:102.

Tahsin, Tasnia, Davy Weissenbacher, Karen O’Connor, Arjun Magge, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2017. “GeoBoost: accelerating research involving the geospatial metadata of virus GenBank records.” *Bioinformatics* 34 (9): 1606–1608.

Towers, S, and Z Feng. 2012. “Social contact patterns and control strategies for influenza in the elderly.” *Mathematical biosciences* 240 (2): 241–249.

Tramer, Elliot J. 1969. “Bird species diversity: components of Shannon’s formula.” *Ecology* 50 (5): 927–929.

Trovão, Nıdia Sequeira, Marc A Suchard, Guy Baele, Marius Gilbert, and Philippe Lemey. 2015. “Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1.” *Molecular biology and evolution* 32 (12): 3264–3275.

Vatant, Bernard, and Marc Wick. 2012. “Geonames ontology.” *Dostupné online:¡ http://www. geonames. org/ontology/ontology v3* 1.

Wallinga, Jacco, Peter Teunis, and Mirjam Kretzschmar. 2006. “Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents.” *American journal of epidemiology* 164 (10): 936–944.

Wheeler, Sarah S, Stanley A Langevin, Aaron C Brault, Leslie Woods, Brian D Carroll, and William K Reisen. 2012. “Detection of persistent West Nile virus RNA in experimentally and naturally infected avian hosts.” *The American journal of tropical medicine and hygiene* 87 (3): 559–564.

Wheeler, Sarah S, Meighan P Vineyard, Leslie W Woods, and William K Reisen. 2012. “Dynamics of West Nile virus persistence in house sparrows (Passer domesticus).” *PLoS neglected tropical diseases* 6 (10).

Worby, Colin J, Sandra S Chaves, Jacco Wallinga, Marc Lipsitch, Lyn Finelli, and Edward Goldstein. 2015. “On the relative role of different age groups in influenza epidemics.” *Epidemics* 13:10–16.

Yang, Jing, Dong Xie, Zhen Nie, Bing Xu, and Alexei J Drummond. 2019. “Inferring host roles in bayesian phylodynamics of global avian influenza A virus H9N2.” *Virology* 538:86–96.

APPENDIX A

STATEMENTS FROM CO-AUTHORS IN PUBLISHED WORK

Chapter 4 of this document has been published in a peer-reviewed journal. Citations for these chapters are listed below. I have obtained permission to use this work from all co-authors: Matthew Scotch.

A.1 Chapter 4

Vaiente MA, Scotch M. Going back to the roots: Evaluating Bayesian phylogeographic models with discrete trait uncertainty. Infection, Genetics and Evolution. 2020 Aug 13:104501. doi:10.1016/j.meegid.2020.104501

APPENDIX B

SEQUENCE METADATA FOR CHAPTERS 1 & 2

In Chapter 1, I studied the age associated diffusion of influenza A/H3N2 using hemagglutinin molecular sequences annotated with patient age obtained from GenBank genbank2012. In Chapter 2, I studied the geographic diffusion of WNV in the US using whole genome sequences annotated with host species and location of sampling. Sequence accessions and relevant epidemiological are available via GitHub (https://www.github.com/matteo-V/appendices).