Generalized Linear Models

in Bayesian Phylogeography

by

Daniel Magee

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

Approved March 2017 by the
Graduate Supervisory Committee:

Matthew Scotch, Chair
Graciela Gonzalez
Jesse Taylor

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

Bayesian phylogeography is a framework that has enabled researchers to model the spatiotemporal diffusion of pathogens. In general, the framework assumes that discrete geographic sampling traits follow a continuous-time Markov chain process along the branches of an unknown phylogeny that is informed through nucleotide sequence data. Recently, this framework has been extended to model the transition rate matrix between discrete states as a generalized linear model (GLM) of predictors of interest to the pathogen. In this dissertation, I focus on these GLMs and describe their capabilities, limitations, and introduce a pipeline that may enable more researchers to utilize this framework.

I first demonstrate how a GLM can be employed and how the support for the predictors can be measured using influenza A/H5N1 in Egypt as an example. Secondly, I compare the GLM framework to two alternative frameworks of Bayesian phylogeography: one that uses an advanced computational technique and one that does not. For this assessment, I model the diffusion of influenza A/H3N2 in the United States during the 2014-15 flu season with five methods encapsulated by the three frameworks. I summarize metrics of the phylogenies created by each and demonstrate their reproducibility by performing analyses on several random sequence samples under a variety of population growth scenarios. Next, I demonstrate how discretization of the location trait for a given sequence set can influence phylogenies and support for predictors. That is, I perform several GLM analyses on a set of sequences and change how the sequences are pooled, then show how aggregating predictors at four levels of spatial resolution will alter posterior support. Finally, I provide a solution for researchers

that wish to use the GLM framework but may be deterred by the tedious file-manipulation requirements that must be completed to do so. My pipeline, which is publicly available, should alleviate concerns pertaining to the difficulty and time-consuming nature of creating the files necessary to perform GLM analyses. This dissertation expands the knowledge of Bayesian phylogeographic GLMs and will facilitate the use of this framework, which may ultimately reveal the variables that drive the spread of pathogens.

DEDICATION

I dedicate this work to all my friends and family that, at the very least, pretend to be interested when I explain what exactly it is that I've been doing since I arrived at ASU. This includes my fiancé, Hansa, my immediate family, Bob, Kathy, Bill, Andy, and Kate Magee, and my grandparents, Jim and Jean Magee and Diane Kasych.

ACKNOWLEDGMENTS

I would like to thank my Graduate Supervisory Committee, Drs. Matthew Scotch, Graciela Gonzalez, and Jay Taylor for their guidance in the completion of my dissertation. I thank all other individuals that scientifically contributed to this work, including Rachel Beard, Dr. Philippe Lemey, and Dr. Marc A. Suchard. This work would not have been possible without various assistance provided by Dr. Abdelsatar Arafa, Dr. Peter Beerli, Sahithya Dhamodharan, Dr. Tony Goldberg, Dr. Andriyan Grinev, Dr. Laura Kramer, Demetri Shargani, Dr. Steve Zink, and the authors, originating and submitting laboratories of the sequences obtained from GISAID's EpiFlu Database. I would also like to thank those individuals that provided academic and other support to me, including Dr. Rolf Halden, Maria Hanlin, Laura Kaufman, Lauren Madjidi, Dr. Anita Murcko, Dr. George Runger, and Marcia Spurlock. I thank the various sources of funding that I have received that allowed me to complete my dissertation: the ARCS Foundation, especially my generous donors, Ellie and Michael Ziegler, the ASU Department of Biomedical Informatics, the National Institutes of Health, and the PLuS Alliance. I thank those that have provided various feedback pertaining to my work over the last four years, including members of the Biodesign Center for Environmental Security and the Department of Biomedical Informatics. Finally, I would like to thank those that allowed me to gain research experience which ultimately led to my admittance into the ASU Biomedical Informatics Ph.D. program, including Larissa Topeka, Drs. Kay Huebner, Josh Saldivar, Matthew During, Lei Cao, and Deborah Lin.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

COMBINING PHYLOGEOGRAPHY AND SPATIAL EPIDEMIOLOGY TO UNCOVER PREDICTORS OF INFLUENZA A/H5N1 VIRUS DIFFUSION IN EGYPT

Introduction

Currently emerging and re-emerging infectious diseases of zoonotic origin such as highly pathogenic avian influenza A pose a significant threat to human and animal health due to their elevated transmissibility (Chen, Liu, Cai, Du, & Li, 2013; Krauss, 2003). Predicting the spread of these viruses is challenging because many of the drivers of disease are not easily identifiable. These drivers can be of an environmental, geographic, demographic, genetic, or other nature. For example, diffusion could be caused by climate, human and avian population density, and other key demographic profiles (Herrick, Huettmann, & Lindgren, 2013). Several techniques exist to help identify these drivers including bioinformatics, phylogeography, and spatial epidemiology but these methods are generally evaluated separately and do not consider the natural complementary principles of each other.

Successful analysis of spatial epidemiological factors have identified air travel and global mobility as key drivers of influenza (Viboud, Bjornstad, et al., 2006) but do not consider the key elements of molecular sequence analysis such as gene flow, cross-species transmission (CST), and viral mutations to support and complement their work. Similarly, bioinformatics and phylogeographic techniques which thoroughly analyze sequence data often ignore climate and demographic factors. Here I will adopt an approach which integrates these separate techniques and helps identify the most important drivers of disease spread. A more comprehensive model of viral diffusion will

be useful for public health and other agencies to develop strategies for curbing spread of these devastating diseases. Knowing the factors that are most relevant in predicting the diffusion will allow for an accurate and continuous threat assessment and prevention. Two previous studies on various influenza subtypes have identified several potential environmental and demographic drivers of viral diffusion including precipitation, humidity, and temperature (Tamerius et al., 2013), human, duck, and chicken density (Van Boeckel et al., 2012) but fail to account for genetic variables. Conversely a study by Lam *et al*. (Lam et al., 2012) showed that H5N1 in Indonesia began by an introduction of the virus in East Java in 2002 and was followed by east and westward migration to cover the entire country. This work highlights that phylogeographic and bioinformatics techniques can pinpoint locations and demonstrate migratory patterns of viral diffusion. Unfortunately, this study lacks demographic and epidemiological factors which also could have contributed to the diffusion, demonstrating a lack of coordination between the methodologies.

Ypma *et al*. (Ypma et al., 2012) presented an integration of these techniques by estimating the migratory patterns of influenza A H7N7 transmission between farms in the Netherlands using genetic data as well as spatiotemporal elements. The authors were able to demonstrate that geography alone is not a reliable indicator of transmission routes but that it does improve the accuracy of the routes when combined with both genetic and temporal data. A different study by Ypma *et al*. (Ypma, van Ballegooijen, & Wallinga, 2013) then utilized within-host dynamics and genetic data to create phylogenetic trees to estimate transmission routes and connect estimating variables. Their separate evaluation of space-time and genetic contributors was a unique innovation to the performance

evaluation of transmission trees. Studies like these have shown how phylogeography, bioinformatics, and epidemiology approaches can be integrated to provide more accurate modeling of disease outbreaks.

The diffusion of H5N1 in Egypt is an excellent candidate for testing such an approach. Egypt has emerged as an epicenter for H5N1 with 173 confirmed human cases as of January 2014, the most of any country outside of Southeast Asia (WHO, 2013). The cultural preference of Egyptian citizens is to utilize live bird markets to obtain their poultry which results in 70% of all poultry trade occurring in this manner (Abdelwhab & Hafez, 2011). The environment of these markets yields a high possibility of infection and spread of H5N1, and in 2009 Abdelwhab *et al.* (Abdelwhab et al., 2010) determined that over 12.4% of tested markets contained infected avian species. These markets thus become a major source of avian-to-human transmission (Abdelwhab & Hafez, 2011). While this can help explain the primary route by which humans are infected by avian species, there is uncertainty as to their connection to human and animal infection across the entire Egyptian landscape.

In this paper, I evaluate the spread of H5N1 in Egypt by reconstructing its phylogenetic history while simultaneously determining the impact of the certain environmental, geographic, demographic, and genetic drivers. This model will help pinpoint the variables most responsible for the diffusion as well as eliminate unsupported characteristics from model consideration. I focus on a variant H5N1 subclade 2.2.1.1., which is one of 10 currently defined subclades within Egypt (WHO, 2012). This particular clade is appropriate because it is found almost exclusively within Egypt and therefore all features of the landscape, culture, and climate are potentially directly

relevant for its diffusion dynamics. I expand on preliminary work by Beard *et al*. (Beard, Magee, Suchard, Lemey, & Scotch, 2014) by including additional predictors of diffusion as well as new techniques for analysis of viral sequences.

Results

In Tables 1.1 and 1.2, I provide the posterior inclusion probabilities and BFs for each predictor, stratified by governorate of origin and destination. The two most strongly supported predictors are avian counts from governorate of destination (BF>20,000) followed by avian counts from governorate of origin (BF=80.28). Although these BFs are in the "very strong" and "strong" categories of Kass and Raftery (Kass & Raftery, 1995), respectively, these likely arise from sampling differentiation between locations. While these predictors are not of direct scientific interest, their inclusion does enable the GLM to help control for differential sampling bias in estimates for the remaining predictors. The following predictors, in order, constitute the remaining factors which reached the BF threshold of 3.0, all coming from the governorate of origin: avian density, pigeon density, longitude, goose density, proportion of avian viral genomes without the genetic motif, chicken density, human density, elevation, precipitation, duck density, human counts, latitude, humidity, temperature, and duck density. There were no supported predictors from the governorate of destination, apart from the avian counts.

Table 1.1

Inclusion support statistics for governorate of origin

Predictor	Posterior Inclusion Probability	Bayes Factor
Avian Counts	0.63	80.28
Avian Density	0.32	22.87
Pigeon Density	0.31	21.45
Longitude	0.30	20.35
Goose Density	0.30	20.24
No Motif Density	0.26	16.78
Chicken Density	0.25	15.63
Human Density	0.24	15.08
Elevation	0.24	14.99
Precipitation	0.22	13.64
Duck Density	0.22	13.20
Human Counts	0.21	12.69
Latitude	0.17	9.51
Humidity	0.16	9.21
Temperature	0.13	7.13
Turkey Density	0.10	5.50
Distance	0.01	0.46

Table 1.2

Inclusion support statistics for governorate of destination

Predictor	Posterior Inclusion Probability	Bayes Factor
Avian Counts	1.00	28058.39
Goose Density	0.01	0.73
No Motif Density	0.01	0.67
Avian Density	0.01	0.62
Pigeon Density	0.01	0.59
Chicken Density	0.01	0.51
Distance	0.01	0.46
Duck Density	0.01	0.46
Human Density	0.01	0.37
Elevation	0.01	0.29
Human Counts	<0.01	0.16
Latitude	<0.01	0.13
Temperature	<0.01	0.13
Humidity	<0.01	0.13
Turkey Density	<0.01	0.11
Longitude	<0.01	0.08
Precipitation	<0.01	0.08

Of the predictors which reached the BF threshold of 3.0, avian density, pigeon density, longitude, and goose density each had a BF in excess of 20.0, which is the threshold marker of a "strong" predictor (Kass & Raftery, 1995). In Figure 1.1, we show the posterior inclusions probability of the 15 supported predictors, BF markers, and the β-coefficient complete with the 95% Bayesian credible interval to visualize uncertainty. The wide range of the 95% credible intervals for each β-coefficient make interpretation of their relative contribution difficult; however the size of the BF for each predictor provides confidence that these variables are in fact playing a role in the spread of H5N1.

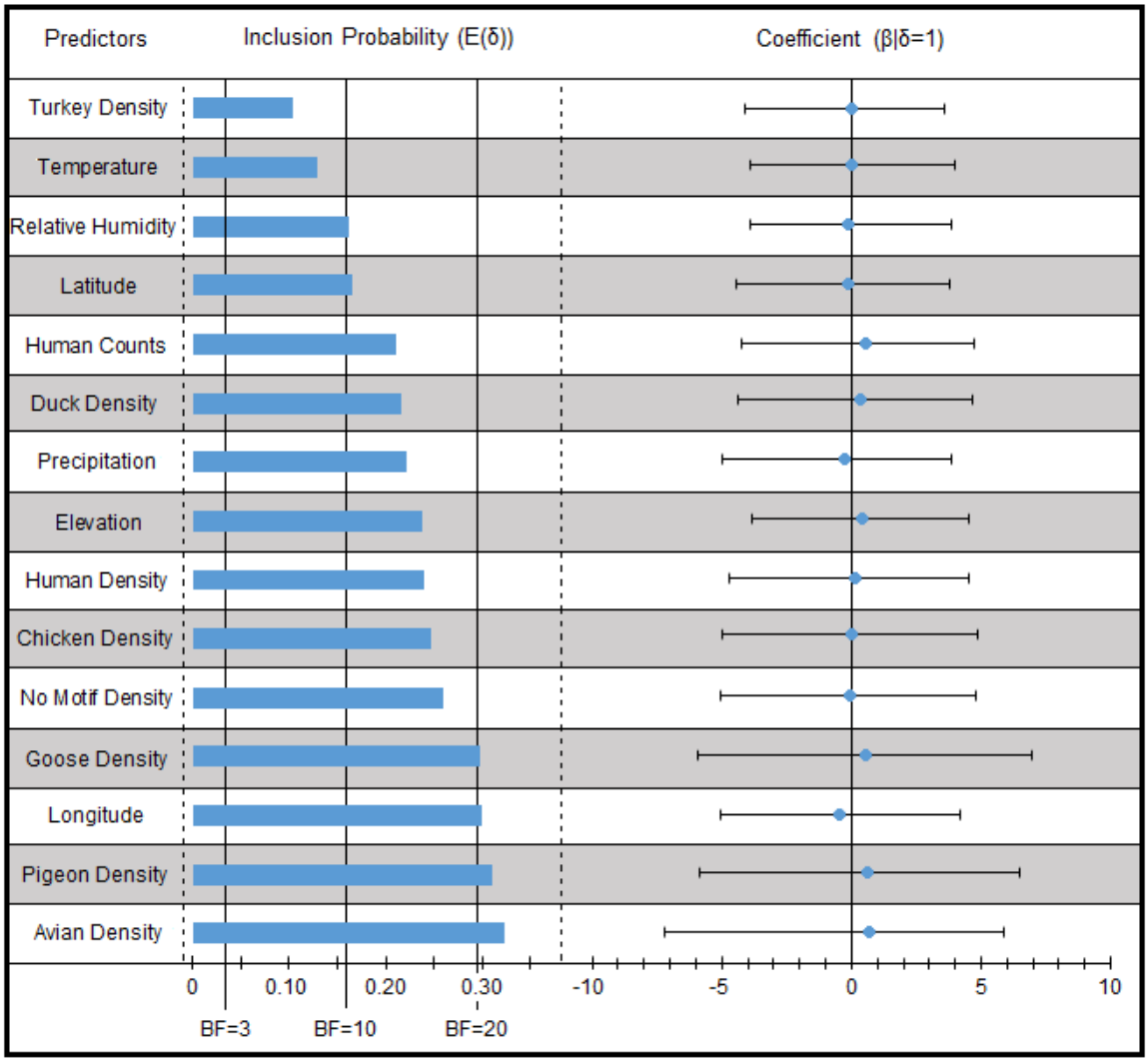

Figure 1.1. Posterior support metrics for the 15 relevant predictors that achieved BF > 3. Inclusion probabilities are represented by the blue bar, and several BF values are annotated as vertical black lines. Also included is the mean posterior regression coefficient, represented by the blue dot, and 95% confidence interval of the GLM test coefficient.

Since the GLM shows a lack of support for any predictor dependent upon governorate of destination it can be concluded that origin-based predictors are primarily responsible for viral spread. Fixed variables such as latitude, longitude, and elevation had

similar support scores as naturally occurring factors like precipitation, relative humidity, and temperature as well as variable agricultural quantities like the densities of specific avian birds and humans. The support of the density of avian birds without the motif indicates that the mutation identified by Yoon *et al*. (Yoon et al., 2013) indeed plays a role in the diffusion process and confirms the role of at least one demographic, geographic, environmental, and genetic feature for the complex spatiotemporal spread of H5N1 influenza in Egypt.

In Table 1.3, I provide the CST results, which indicate that transmission to humans is generally caused by ducks, turkeys, and geese. This is surprising given that the overall population density of chickens in Egypt is far larger than any of the other avian species analyzed. Humans were also calculated to have a high transmissibility to turkeys, geese, and ducks but not toward chickens and had the highest mean of per-capita transmission to all species. By these same calculations, turkeys were second most transmissible, followed closely by ducks and geese while chickens were least-transmissible among species measured. The mean per capita CST values from largest to smallest is: human, turkey, duck, geese, and chicken. Mean duck and geese CST values are very similar as well at 2.37 and 2.31, respectively.

Table 1.3

Calculated cross-species transmission values from Migrate-n

		Species Transmitted To					
		Human	Chicken	Duck	Goose	Turkey	Mean
Species Transmitted From	Human		1.02	3.42	4.61	5.23	3.57
	Chicken	1.40		0.85	2.23	2.10	1.65
	Duck	3.58	0.70		3.01	2.18	2.37
	Goose	3.08	0.70	2.97		2.49	2.31
	Turkey	3.34	0.99	2.53	3.30		2.54
	Mean	2.85	0.85	2.44	3.29	3.00	

Discussion

In this work, I modeled H5N1 viral spread in Egypt while simultaneously testing the role of various environmental, geographic, demographic, and genetic predictors. The posterior inclusion probabilities and calculated BF values show support for 15 variables of direct scientific interest. While these 15 variables have relatively low probabilities ($E(\delta) < 0.35$) this should not be taken to mean that the variables are not relevant to the diffusion process. If we have $E(\delta)=0.30$ for a given predictor, this means that 30% of all possible linear models, including or excluding that and all other predictors, support its inclusion with a high probability. Furthermore, the BF values indicate how much more likely it is that the predictor should be included than the defined posterior probability that there was a 50% chance of no predictor being included. This conservative prior probability allows us to state the strength of support for each predictor with high confidence, even if the posterior inclusion probability remains low.

Among avian species I found that densities of ducks, geese/guinea fowl, turkey, pigeons/other birds and chickens are all supported for inclusion within the phylogeographic GLM, all with similar BFs while human density has an inclusion

probability ranking in between that of the various avian species. The support for geese/guinea fowl and pigeons/other birds is likely a result of collinearity with the chicken, duck, and turkey predictors based on the way their point estimates were obtained. Pearson's r between the overall avian density predictors and the pigeon/other bird and geese/guinea fowl predictors exceeded 0.99, although the overall predictor design matrix did achieve full rank. This emphasizes the need for health agencies to consider human and animal census data when determining infectious disease risk while focusing on known viral carriers and reservoir species. This also supports the notion that live bird markets are involved with transmission due to high density and close contact with humans. Real-time monitoring of live bird market inventory would provide public health agencies with very accurate numbers of poultry and enable them to have detailed information in specific locations. This could be done simply by requiring all market vendors to report their stocks each day and the market could submit a compiled dataset on a weekly basis. Active data collection such as this would be effective in determining whether specific species are directly linked with trends in the diffusion of various viruses including H5N1.

Our findings that environmental factors are predictors of influenza diffusion are consistent with work by He *et al*. (Herrick et al., 2013) who analyzed virus spread in Canada. Specifically, the authors identified longitude, temperature, and humidity as strong predictors, all of which are supported in our GLM by the BF metric. This reiterates the previous findings that geographic and climate factors impact the diffusion of influenza. In contrast, their model did not identify human population as a significant predictor (Herrick et al., 2013). I used population density rather than raw population and

our result positively indicates human density should be included within the model (BF = 15.08) from the governorate of origin. This discrepancy could be explained by the fact that Egypt's population density is approximately 24-fold that of Canada's (CAPMAS, 2012a; Statistics Canada, 2013) so human-to-human transmission is far more likely. Poultry density and household density were also found to be among ecological determinants of H5N1 spread in Bangladesh (Ahmed et al., 2012). Since our model analyzed the same virus in a country where live bird markets are also prevalent (Dolberg, 2009) these conclusions strongly suggest that both avian and human population sizes are reliable indicators of H5N1 diffusion.

Several of the predictors supported in our model have also been linked to H5N1 risk in various other studies. For example, elevation had previously been identified as a risk factor of other HPAIs including H5N1 in Indonesia (Loth et al., 2011), and Vietnam (Pfeiffer, Minh, Martin, Epprecht, & Otte, 2007) so this predictor should undoubtedly be included in most models and is strongly included in ours. Chicken density has been identified as a risk factor in Vietnam (Pfeiffer et al., 2007) and additionally confirmed in Cambodia, Laos, and Thailand (Gilbert et al., 2008). Furthermore, Gilbert *et al*. (Gilbert et al., 2008) concluded that duck, geese, and human population were correlated as risk factors in southeast Asia, all of which are supported in our model. Precipitation has been shown to be an indicator of outbreak risk of H5N1 in Europe (Si, de Boer, & Gong, 2013) and given the relative ease of tracking and reporting such a value via active World Meteorological Organization stations it should be included in future models. The consistent identification of these variables in Egypt as well as various regions indicates

that these should be carefully monitored by health agencies during surveillance efforts regarding avian influenza.

Lemey *et al*. (Lemey et al., 2014) previously demonstrated the capabilities of a phylogeographic GLM for determining spread of H3N2 using a similar set of predictors. While that study provided a global look, our work focused on one region to identify diffusion drivers specific to Egypt. Our approach has allowed us to identify key variables which contribute to the H5N1 diffusion and provides a rough model that can be tested in other countries and with other viruses. The ability to determine consistent variables relating to viral diffusion would undoubtedly be a huge breakthrough to understanding spatial spread.

This study has several limitations including the inability to include CST values directly within the GLM. The CST values represent rates of transmission between species but are not location-specific, thus could not be incorporated as predictors. I therefore used CST data as a complement to our GLM. I was unable to use transmission path distance between the locations because road access was not available to the centroid location for each governorate. Trends in variable predictors could prove to match up with spikes in reported cases that will further supplement their inclusion within our GLM. In addition, I was unable to obtain the exact location from which the sequences were collected and could therefore only utilize the centroid coordinates for each location. These discrepancies in distance and true location could certainly impact the inclusion of the latitude, longitude, and geographical distance predictors within the GLM. At the time of this writing the most recent World Health Organization update on human case counts within Egypt was January 2014 (WHO, 2013) which provides us with potentially

outdated data for this predictor. Additionally, the number of avian birds by species needed to be estimated for chickens, geese/guinea fowl, and pigeon/other birds because these data were not available per governorate for 2011. Although these were approximations, the BF support values make a compelling case that the estimations were accurate and are consistent with previous findings. Our estimates and the data included within the GLM are under the assumption that there has not been a large overhaul of agricultural land within each governorate since the most recent publication of these population values.

Although this work focused solely on influenza H5N1 in Egypt, this approach remains generalizable to additional locations and viruses and demonstrates the usefulness of combining phylogeographic, bioinformatics, and epidemiological approaches to simultaneously to evaluate the viral spread. These methods can be combined with an established framework of evolutionary and ecological dynamics to explain spatial diffusion (Grenfell et al., 2004). Our future work will include other clades of H5N1, an expansion of environmental predictors, and more genes of interest such as neuraminidase to develop a more comprehensive model. I will also expand our geographic focus to determine if our significant predictors are constant across other countries such as China or Indonesia where H5N1 is persisting. GLMs such as this will undoubtedly aid public health agencies in their ability to predict and prevent outbreaks as well as explore improvements in preventative tactics. Our identification of drivers will be useful for public health agencies to monitor pandemic risk levels, plan protocols for reducing threats, and devise strategies best suited to protect citizens from the consequences of outbreaks.

Materials and Methods

Sequence Data

I utilized the dataset by Scotch *et al*. (Scotch et al., 2013) which contains 226 sequences of the hemagglutinin gene of H5N1 influenza variant subclade 2.2.1.1. The dataset includes sequences from 20 of the 27 governorates (Figure 1.2) that were isolated from 2007-2012 from both human and avian hosts. The host species and number of sequences is as follows: chicken (156), duck (43), human (14), goose (6), turkey (4), environment (2), and quail (1). I refer the reader to Scotch *et al.* (Scotch et al., 2013) for details on classification of the sequences into subclade 2.2.1.1. and analysis of phylogeographic trees. I provide the GenBank accession, governorate of isolation, host, and year of isolation of each sequence in Appendix A.

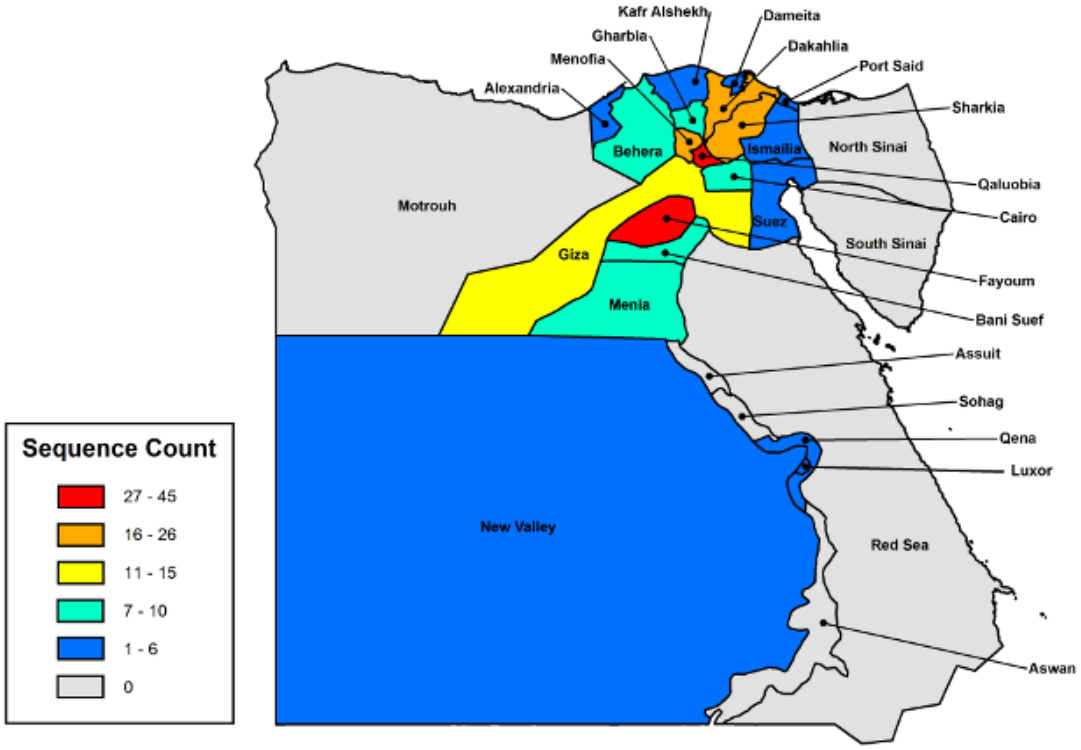

Figure 1.2. Map of Egypt and the governorates included. The 226 sequences used in this study span 20 of the 27 governorates.

Generalized Linear Model

I adopted a Bayesian phylogeographic generalized linear model (GLM) approach by Lemey *et al*. (Lemey et al., 2014) to reconstruct spatiotemporal patterns of viral spread while simultaneously assessing the impact of our predictors. In this approach, I discretize geographic locations and model diffusion between locations through a continuous-time Markov chain (CTMC) process in which I parameterize the instantaneous rates via a GLM. Specifically, I used a non-reversible CTMC process expressed as a K x K infinitesimal rate matrix of location change (Λ) among K discrete locations (Lemey et al., 2014). I parameterize the instantaneous rate Λ_{ij} by utilizing a

linearized log function to incorporate all potential pairwise predictors $p_1, \ldots, p_n$ and evaluated them on a log-scale, per the following equation:

$$log\Lambda_{ij} = \beta_1\delta_1 \log\left(p_{1\{ij\}}\right) + \beta_2\delta_2 \log\left(p_{2\{ij\}}\right) + \cdots + \beta_n\delta_n \log\left(p_{n\{ij\}}\right)$$

Here, β_i indicates the relative contribution of predictor p_i to the whole GLM and δ is a binary indicator which determines whether an individual predictor is included in the model for evaluation (Kuo & Mallick, 1998). The indicator enables a Bayesian stochastic search variable selection (Chipman, George, & McCulloch, 2010; Kuo & Mallick, 1998) such that all posterior probabilities of each possible model, including or excluding every predictor, are estimated. I used a Bernoulli prior probability distribution to place an equal probability for inclusion or exclusion of each predictor (Lemey et al., 2014), and set the prior success probability of the Bernoulli distribution such that there was a 50% prior probability that the model does not include any predictor. I log-transformed and standardized all predictor values, specified a constant size coalescent prior and general time reversible (GTR) substitution model and implemented the GLM within Bayesian Evolutionary Analysis by Sampling Trees (Drummond, Suchard, Xie, & Rambaut, 2012) (BEAST) v1.8.0 with the Broad-platform Evolutionary Analysis General Likelihood Evaluator (BEAGLE) 2.1 (Ayres et al., 2012) library implementation. The model was evaluated with a chain length of 20 M, logging estimates every 10,000 steps and predictor covariates were evaluated for convergence (e.g. effective sample sizes of regression coefficients exceeded 200 for each predictor) using Tracer v1.5 after discarding the first 10% of logged estimates as burnin. The nature of the log-linear function requires each value to be positive so any data points that were missing or zero were transformed to avoid this error. Specific instances are detailed below.

Environmental, Geographic, Demographic, and Genetic Predictors

I selected the following potential predictors with the aid of experts studying H5N1 in Egypt. For our nonreversible diffusion process A→B, I evaluated each predictor from the governorate of origin as well as the governorate of destination. In Table 1.4, I provide descriptive statistics for the predictors.

Table 1.4

Descriptive statistics of each predictor for the 20 governorates

Predictor	Units	Mean	Median	SD	IQR
Distance	Kilometers	265	184	206	296
Latitude	Degrees	29.66	30.39	1.94	1.42
Longitude	Degrees	31.31	31.25	0.98	1.03
Avian Counts	Cases / year	17.6	12.9	15.9	25.8
Human Counts	Cases / year	1.1	1.1	0.8	1.3
Human Density	Heads / km^2	1056	536	1094	1197
Avian Density	Heads / km^2	1290	459	1465	1992
Chicken Density	Heads / km^2	998	379	1065	1698
Turkey Density	Heads / km^2	14	3	24	20
Duck Density	Heads / km^2	120	23	304	35
Goose Density	Heads / km^2	55	20	63	84
Pigeon Density	Heads / km^2	103	37	118	159
No-Motif Density	Heads / km^2	1090	428	1153	1911
Elevation	Meters	88.6	59.0	72.7	60.7
Precipitation	mm / year	41.9	30.0	45.5	53.0
Temperature	Celsius	21.6	21.3	1.4	1.4
Relative Humidity	Percent	56.1	54.5	10.4	15.5

Latitude, Longitude, and Elevation. I obtained geographic coordinates for the centroid of each governorate using geonames.org. While these coordinates likely do not reflect the exact location of the host, we chose the centroid to create uniformity in the model. I used Google Earth to obtain the elevation of each centroid.

Distance. I used Google Maps to calculate the raw linear distance between the centroid of each governorate. Although road or travel distances would likely be more accurate in terms of true transmission paths, the isolated location of some of the centroid locations made this impossible to calculate.

Human and Avian Population Density. Currently, the most recent data for human populations per governorate is a 2012 estimate by the Egyptian Central Agency for Public Mobilization and Statistics (CAPMAS, 2012b). I used two databases provided by the Food and Agricultural Organization of the United Nations (FAO) to obtain the avian populations: FAOSTAT (FAO, 2014a) and the Global Livestock Production and Health Atlas (GLiPHA) (FAO, 2014b). The specific categories of avian populations provided by these resources are chickens, turkeys, ducks, geese/guinea fowl, and pigeons/other birds. I was unable to use 2012 data for the avian populations because there is no breakdown of populations per species for each governorate available for that year. The number of ducks and turkeys were available for each governorate for 2011 and were available for chickens for 2005 via GLiPHA. I estimated the chicken populations for 2011 by prorating the 2005 value per governorate to the total FAOSTAT value for 2011. There was no data available per governorate for geese/guinea fowl or pigeons/other birds for any year so I estimated these values to be the percentage of total geese/guinea fowl or pigeons/other birds from FAOSTAT equal to the percentage of chickens, ducks, and turkeys relative to the total amount in Egypt for 2011 per governorate. To meet the requirements for the log-linear model, any missing value was imputed via mean imputation. Total avian populations reflect the sum of the five avian categories

previously described. For avian and human density, I divided total population by the land area of each governorate to obtain a density of heads per km^2.

Viral Genomes Lacking a Genetic Motif. According to Yoon *et al*. (Yoon et al., 2013) the pathogenicity of H5N1 depends on the number of basic amino acids at the HA cleavage site. This includes a mutation PQGE**R**RRK/RKR*GLF to PQGE**G**RRK/RKR*GLF. The presence of this motif results in a reduced pathogenicity of the virus and I used Geneious Pro 5.0.3 (Biomatters Ltd., Auckland, New Zealand) to locate the presence of this mutation in our HA sequences. I calculated the expected number of total avian influenza sequences per governorate which lack the motif by the following equation:

$$N_j = T_j * (A_j - M_j) / A_j$$

In this equation N_j is the expected number of avian influenza sequences that lack the genetic mutation, T_j is the total avian population for 2011, A_j is the number of avian influenza sequences obtained from the governorate, and M_j is the number of sequences which contain the motif. The resulting value was divided by the land area in order to obtain a density in heads per km^2.

Precipitation, Temperature, and Relative Humidity. I obtained the data for average annual rainfall, temperature, and relative humidity from the National Climatic Data Center as part of the National Oceanic and Atmospheric Administration (NOAA, 2014). I obtained data for each governorate from the climate station nearest to the centroid. The values represent 30-year averages for the window of January 1, 1961 through December 31, 1990. Although this range does not cover the time period from which our sequences were obtained, the World Meteorological Organization has defined

this period as the current climate normal (WMO, 2013) and likely represents an accurate depiction of typical weather over the timespan of our study.

Case Counts. I obtained the number of confirmed human and estimated avian cases from the Dr. Abdelsatar Arafa at the FAO spanning the years 2007-2013. In total, 2,460 avian cases and 158 human cases covered the 20 governorates in the study and data imputed in the GLM reflects the average number of cases per year for each governorate. Two governorates, New Valley and Port Said, did not have any recorded human cases over the time period so each was fixed with one case to avoid an undefined value for log-transformation. These imputations should not create a sampling bias due to their minimal increase in the sample size.

Cross Species Transmission

I used the program Migrate-n v3.6 (Beerli & Felsenstein, 2001) in order to analyze the relationship between sequences obtained from different species. To maximize the amount of sequences that could be analyzed, I fitted sequences of a unique length with up to 3 "wild-card" nucleotides at the c-terminus to be added in with the nearest population of sequences. I ran the program under the default settings with all sequences fitting these criteria including chicken, duck, turkey, goose, and human hosts. This accounted for 219 of the 226 original sequences in our dataset and resulted in the loss of our only quail sequence. The calculation and description of CST values were described by Streicker *et al*. (Streicker et al., 2010) and I used the following equation to incorporate the Migrate-n output (Faria, Suchard, Rambaut, Streicker, & Lemey, 2013):

$$R_{ij} = \beta_{ij} * \theta_j * \tau^{-1}$$

Here, R_{ij} represents the per capita CST from species *i* to species *j*, β_{ij} represents the unidirectional migration rate obtained by Migrate-n from species *i* to species *j*, θ_j represents the estimate of genetic diversity for species *j* obtained from Migrate-n, and τ represents the generation time of H5N1. τ is defined as the sum of the incubation and infectious periods for H5N1 which is approximately 2.48 days (Bouma et al., 2009). The CST can be interpreted as the expected number of infections in species *i* resulting from just one infected individual of species *j*, although these data may not necessarily reflect the sampling distribution of the host species of our virus sequences. That is, I cannot be certain whether the hosts would maintain a constant CST value per discrete state, and cannot perform additional Migrate-n analyses as not every host was sampled in every discrete state.

Evaluation of Predictor Inclusion

I obtained posterior inclusion probabilities for each individual predictor via BEAST and used Bayes factors (BFs) to determine support of each predictor within the model (Suchard, Weiss, & Sinsheimer, 2005). The inclusion probability is the indicator expectation, $E(\delta)$, which is defined as the probability that the individual predictor is included in the model and is a raw support statistic (Lemey et al., 2014). The greater the inclusion probability the more likely it is that the predictor is contributing to the diffusion process. To compare these probabilities with a baseline, I calculated BFs via posterior odds of predictor inclusion divided by prior odds as demonstrated by the following equation (Lemey et al., 2014):

$$BF = \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$$

Here p_i is the posterior probability of predictor inclusion, or $\delta=1$, while q_i is the prior probability that $\delta=1$. In this model q_i is the binomial prior on the total number of successes ($\delta=1$) that prefers a 50% likelihood of no predictor being included in the model and is calculated using the binomial distribution probability mass function. The BF quantifies the relative support of two competing hypotheses, p_i and q_i, given the observed data (Suchard et al., 2005) and shows which of the two hypotheses is more likely given the data. The cutoff BF for support within the model was set at 3.0 as is consistent with previous work (Philippe Lemey, Rambaut, Drummond, & Suchard, 2009), for establishing a threshold for positive evidence against the null hypothesis, q_i (Kass & Raftery, 1995). This allowed us to account for the possibility of high correlation between predictors. For example, a BF score of 3.0 indicates that the model including that covariate is 3-fold more likely than the model not including it.. The GLM also produces a β-coefficient for each predictor which is the contribution of the predictor to the model as seen in the equation for the log-linear GLM. I used a bit flip operator to evaluate δ similar to Drummond *et al*. (Drummond & Suchard, 2010) in order to complete the calculations.

CHAPTER 2

BAYESIAN PHYLOGEOGRAPHY OF INFLUENZA A/H3N2 FOR THE 2014-15 SEASON IN THE UNITED STATES USING THREE FRAMEWORKS OF ANCESTRAL STATE RECONSTRUCTION

Introduction

Bayesian phylogeography has emerged as a powerful approach to analyzing virus spread. It utilizes sequence data to perform ancestral reconstruction and estimate the most likely lineages of the viruses in rooted, time-measured phylogenies (Lemey et al., 2009) using nucleotide substitution models, molecular clocks, and coalescent priors under a probabilistic Bayesian framework known as Bayesian stochastic search variable selection (BSSVS) (Chipman et al., 2001; Kuo & Mallick, 1998; Lemey et al., 2009). This framework has improved ancestral state reconstruction and has recently been used to analyze human and animal influenza viruses both globally (Bedford et al., 2015; Nelson et al., 2015) and nationally (Pollett et al., 2015; Scotch et al., 2013). By identifying the relationship between geospatial origins and genetic lineages, much can be learned about the complex process in which these viruses spread. Phylodynamic analyses that aim to combine immunological, epidemiological, and evolutionary biology techniques (Grenfell et al., 2004) also enhance our understanding of virus transmission dynamics and their relationship to a phylogeny. These studies have unveiled novel properties of several influenza viruses, including pdm09 (Su et al., 2015), H3N2 (Koelle & Rasmussen, 2015) and highly pathogenic avian influenza H5N1 (Arafa et al., 2016). Building upon the benefits of a BSSVS framework, recent work by Lemey *et al.* (Lemey et al., 2014) utilized a phylogeographic generalized linear model (GLM) approach to identify

environmental, genetic, demographic, and geographic predictors that contributed to the global spread of H3N2 influenza viruses. In the GLM, the BSSVS on the discrete location variable is also used to estimate the posterior inclusion probability of potential predictors in a log-linear combination to model the transition rate matrix. Similarly, studies have followed this approach to uncover the predictors associated with the spread of H5N1 in Egypt (Magee, Beard, Suchard, Lemey, & Scotch, 2015) and for HIV in Brazil (Graf et al., 2015). Such studies have demonstrated the utility of combining genetic and geospatial inferences from phylogeography with surveillance data in epidemiological studies like Yang *et al.* (Yang, Lipsitch, & Shaman, 2015). These analyses may enable actionable solutions for public health officials once consistent identification of contributing predictors is achieved.

Although the GLM appears to show promise with its simultaneous ability to perform ancestral state reconstruction and also assess the contribution of predictor variables of interest, there has yet to be an assessment of how a standard BSSVS approach and a GLM approach compare in their phylogeographical reconstructions. Specifically, no study has yet compared root state probabilities in a phylogeny constructed via BSSVS to the same probabilities using the GLM approach. Such information may inform researchers of differences in phylogeographic trends that may be experienced by choosing one framework over the other. In this work I analyze the 2014-15 H3N2 flu season within the U.S. by performing ancestral state reconstruction of a discrete location variable via the following three frameworks: an asymmetric substitution model without BSSVS (–BSSVS), an asymmetric substitution model with BSSVS (+BSSVS) (Lemey et al., 2009), and a GLM (Lemey et al., 2014). For the BSSVS

framework, I analyze separate versions that place either a Poisson distribution (+BSSVS(P)) or a prior uniform distribution (+BSSVS(U)) on the number of positive rate parameters to determine the influence of location priors. For the GLM framework, I analyze separate versions that include and do not include sample size predictors, which I denote as GLM(+SS) and GLM(–SS), respectively, to directly quantify the effect of sampling bias on GLM-constructed rate matrices and potential suppression of the signal of other predictors. This brings us to a total of five methods that encompass the three frameworks. I refer readers to *Materials and Methods* for full details on the methods. These selections allow us to empirically evaluate differences in the phylogenies obtained via each method and to determine whether one framework provides more accurate posterior estimates given a fixed set of data. I demonstrate these trends using multiple random samples from a large collection of flu sequences to show reproducibility as well as analyze several coalescent tree priors to show consistency among the reconstruction methods across varying parameters. Finally, I show that support for GLM predictors can change given the tree priors and sequence sets, but that trends among specific predictors will emerge to allow accurate determination of their impact on viral diffusion.

Results

In Figure 2.1A, I show mean log marginal likelihood estimates among the six samples obtained by path sampling (PS) and stepping stone sampling (SSS) for each prior and reconstruction method. For PS, the two methods that obtain the highest mean log marginal likelihoods are the GLM(+SS) and GLM(-SS), respectively, under each prior. The mean +BSSVS(U) finds greater log marginal likelihoods than the mean +BSSVS(P)

under each prior as well, although the mean –BSSVS exceeds both under the constant and exponential priors. For SSS, the log marginal likelihood increases in a near-linear manner for the +BSSVS(P), +BSSVS(U), GLM(–SS), and GLM(+SS) methods. The –BSSVS method, however, finds the largest posterior support under the constant, expansion, exponential, logistic, and Skyline priors.

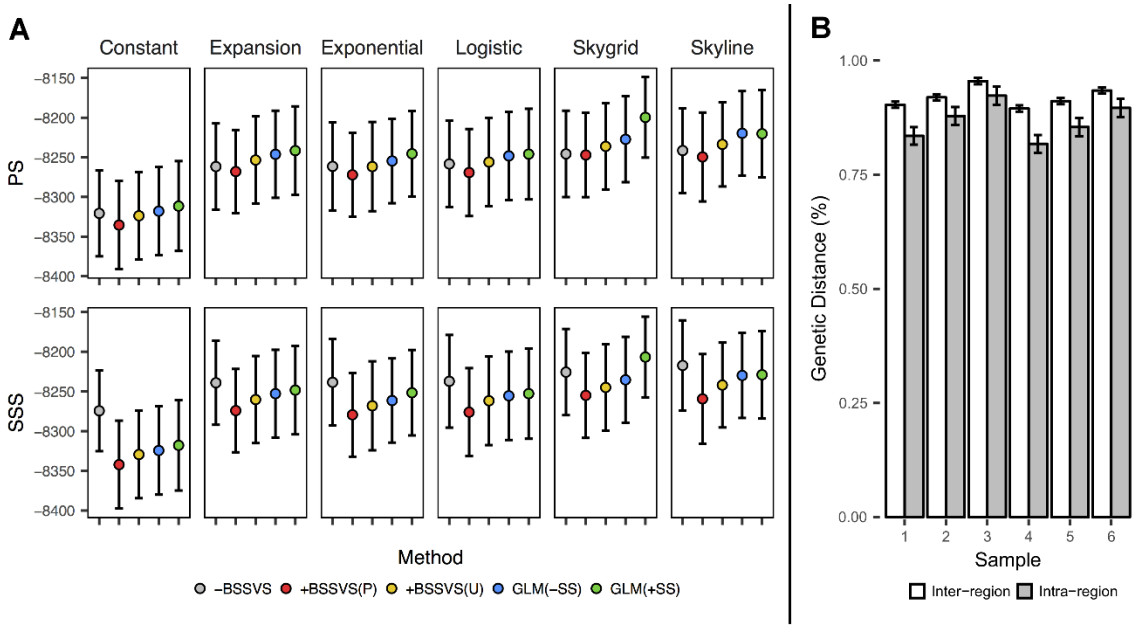

Figure 2.1. Model comparison statistics and location-specific genetic diversity. (A) Model comparisons obtained via path sampling (PS) and stepping stone sampling (SSS) for the six coalescent priors and five methods. (B) Average genetic distances between all pairwise intra-region and inter-region sequences for the six samples, expressed as a percent, with 95% confidence intervals shown as error bars.

In Figure 2.2, I present log marginal likelihood estimates for each individual model. From Figure 2.2, I show that each GLM(+SS) and GLM(–SS) unanimously finds more posterior support than their corresponding +BSSVS(P) for both PS and SSS. The

+BSSVS(P) method demonstrates consistently poor performance, as its posterior estimates are the worst of the five methods in 25 of 36 PS analyses and 32 of 36 PS analyses (79% overall) across all priors, while no GLM(+SS) or GLM(–SS) yields the lowest posterior estimate of model support among the three methods for either PS or SSS under any prior, although no pairwise t-test shows a significant difference.

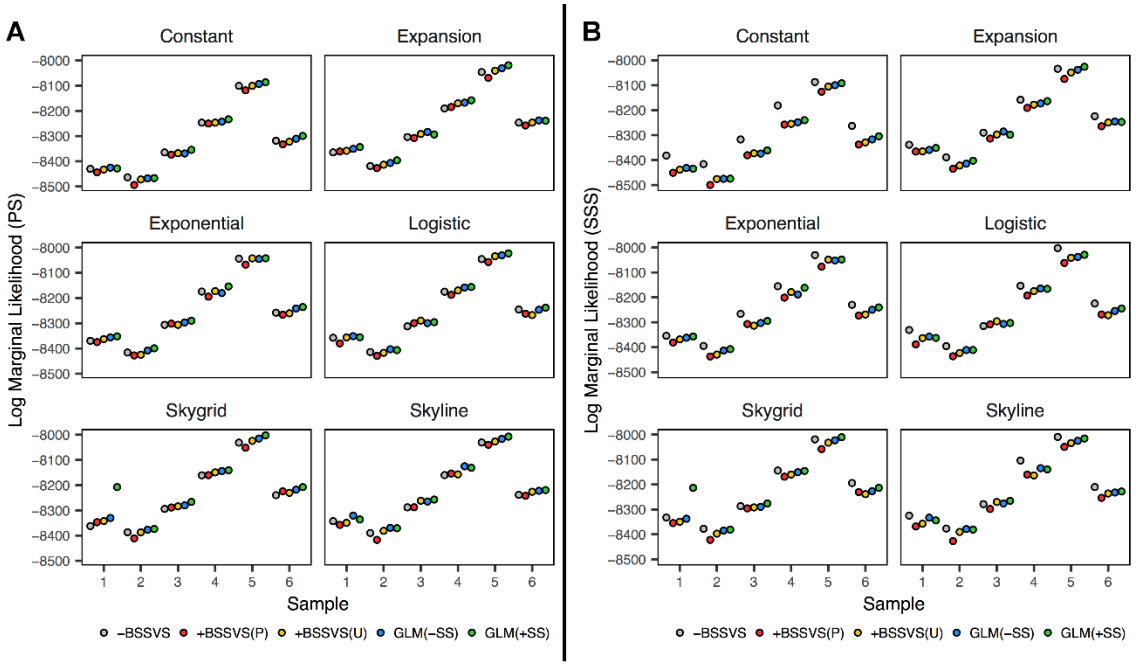

Figure 2.2. Model comparisons for the 180 analyses. (A) Log marginal likelihood obtained via path sampling (PS). (B) Log marginal likelihood obtained via stepping-stone sampling (SSS). Metrics are shown for each sample, prior, and method.

Each of the 180 models show statistically significant differences between the null and observed means for the association index (Figure 2.3). These data suggest stronger support for the phylogeny-trait association (Parker, Rambaut, & Pybus, 2008) and, as all $p < 0.01$, suggest the evolution of influenza during this flu season was structured by geography. The support of the sampling location-phylogeny associations observed in

Figure 2.3 can be explained, in part, by the amount of genetic diversity observed within and across each region. In Figure 2.1B I show the average genetic distances between intra-region and inter-region sequences. Here, I calculated the genetic distances among all 40,470 pairwise sequences and present the mean distance of sequences sampled in the same region (*e.g.* Region 1-Region 1) to those sampled in different regions (*e.g.* Region 1-Region 2). From Figure 2.1B, the pairwise intra-region sequences (n=4,496 per sample) have a lesser amount of genetic diversity than the pairwise inter-region sequences (n=35,974 per sample) in each our six sequence sets. A two-tailed t-test shows $p < 0.01$ for each sample, indicating that sequences from within the same region demonstrate significantly lower amounts of genetic diversity than those from external regions. The average intra- and inter-region distances in the full set of 1,163 sequences are 0.872% (95% CI = [0.867, 0.878]), and 0.929% (95% CI = [0.926, 0.932]), respectively ($p < 0.0001$). These data demonstrate that our method of downsampling maintained representative levels of genetic diversity across the six samples.

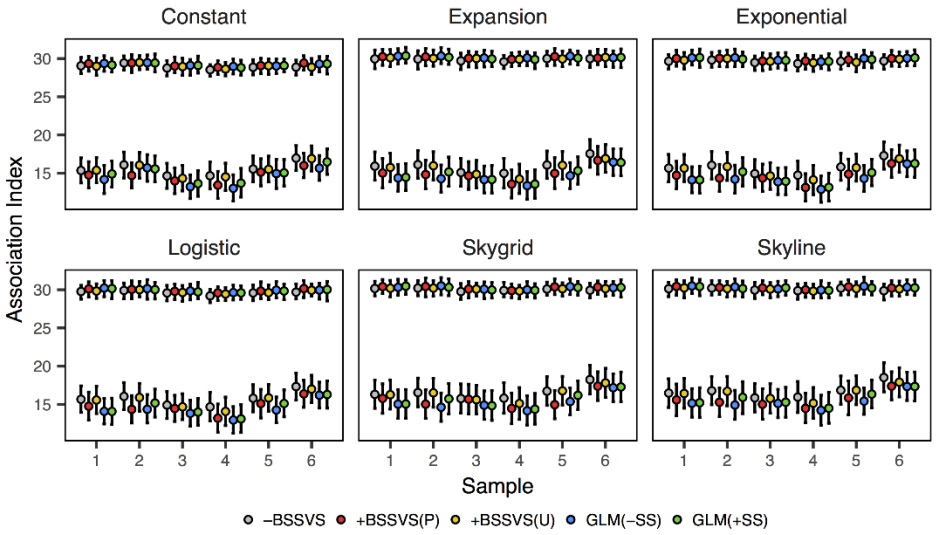

Figure 2.3. Association index scores obtained via BaTS. For each model, I show the null mean (larger value) and observed mean (smaller value) and their respective 95% confidence intervals. For each model, I observe $p < 0.0001$ between the null and observed means.

In Figure 2.4, I show four root state metrics obtained from the maximum clade credibility (MCC) trees of each of the 180 models. In Figure 2.4A, I show the mean root state posterior probability (RSPP). Aside from the constant coalescent prior, the mean GLM(–SS) and GLM(+SS) methods consistently show the largest mean RSPP of the five methods. The mean GLM(–SS) finds significantly greater RSPPs under each coalescent prior than the mean –BSSVS ($p < 0.03$ for each coalescent prior) and significantly greater RSPPs than both the mean +BSSVS(P) and +BSSVS(U) for the expansion and exponential coalescent priors. Similarly, the GLM(+SS) shows a mean RSPP significantly greater than the –BSSVS and +BSSVS(U) methods for all coalescent priors except constant, and significantly greater RSPP than the +BSSVS(P) for the constant, expansion, Skygrid, and Skyline coalescent priors. Across all coalescent priors, the mean

RSPP for the –BSSVS, +BSSVS(P), +BSSVS(U), GLM(–SS), and GLM(+SS) methods are 0.48, 0.56, 0.49, 0.81, and 0.74 respectively. These differences per method could be influenced by the sample size per discrete state, so I show the Pearson's r correlation coefficient between the sample size at each discrete state and its corresponding posterior probability at the root in Figure 2.4B. Here I observe that the +BSSVS(P) shows a correlation coefficient less than 0.4 for the constant, expansion, Skygrid, and Skyline coalescent priors but for the exponential and logistic coalescent priors the coefficient is nearly doubled. Meanwhile, the +BSSVS(U), –BSSVS, GLM(–SS), and GLM(+SS) methods are generally consistent under all priors. The mean +BSSVS(P) shows significantly less correlation than each of the other four methods for the constant, expansion, and Skyline coalescent priors ($p < 0.02$ for each) while the +BSSVS(U), –BSSVS, and GLM methods do not show any significant differences under any coalescent prior.

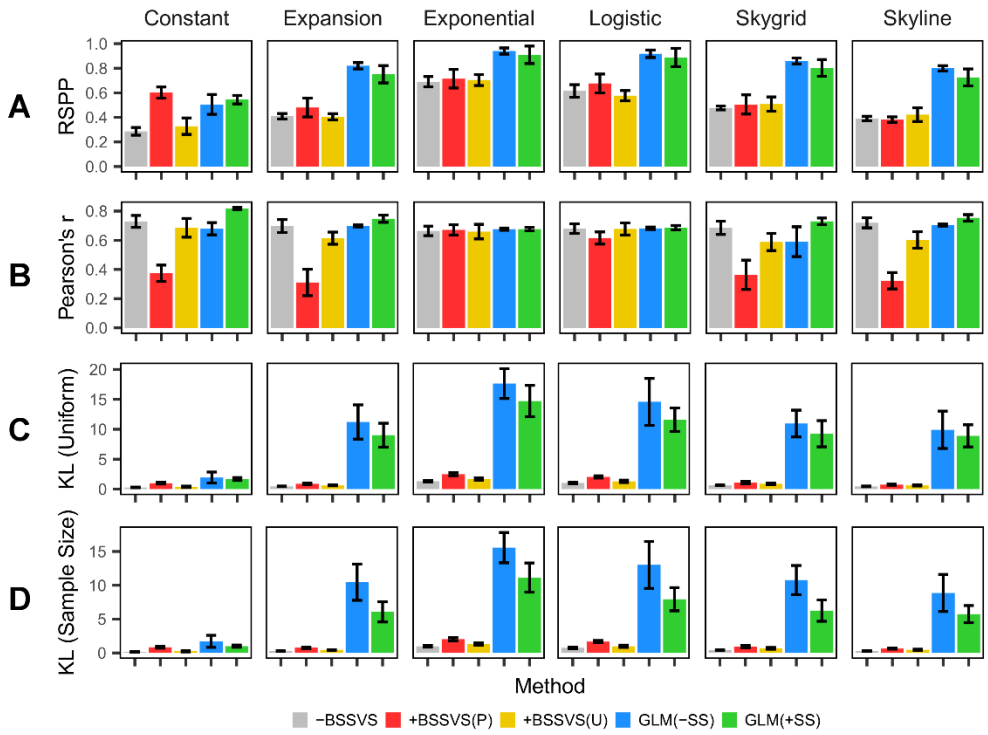

Figure 2.4. Mean posterior metrics of the MCC phylogenies. Values represent the mean indicated statistic from the six samples under each coalescent prior and method with error bars representing the standard error. (A) Root state posterior probability. (B) Pearson's correlation coefficient for the number of sequences per discrete state and the root state posterior probability for each discrete state in each model. (C) Kullback-Leibler divergence calculated assuming a uniform prior probability per discrete state. (D) Kullback-Leibler divergence calculated assuming a prior probability proportional to the number of sequences per discrete state.

Figures 2.4C and 2.4D show the Kullback-Leibler (KL) divergence between the prior and posterior probabilities at the root states calculated using two different prior assumptions (see *Materials and Methods* for details). KL values indicate the extent to which a model is able to generate posterior probabilities at the root state that differ from

the prior probabilities at the root state. That is, high KL values indicate strong divergence from the prior probabilities and, thus, strong posterior information gain, while low KL values indicate the opposite. From Figures 2.4C and 2.4D, the mean GLM(–SS) and GLM(+SS) KL divergences demonstrate a marked increase over the –BSSVS, +BSSVS(P), and +BSSVS(U) methods under the expansion, exponential, logistic, Skygrid, and Skyline coalescent priors ($p < 0.02$ for all two-tailed t-tests. Under the constant coalescent prior, both the mean GLM(–SS) and GLM(+SS) KL divergences exceed the mean KL under both assumptions of the –BSSVS, +BSSVS(P), and +BSSVS(U) methods, but none of these values are significant. The +BSSVS(P) method, in turn, shows significantly greater KL divergences under both assumptions than the –BSSVS method under all coalescent priors and significantly greater than the +BSSVS(U) method under the constant, exponential, and logistic coalescent priors. I show data for each of the four metrics in Figure 2.4 by individual model in Figures 2.5 and 2.6.

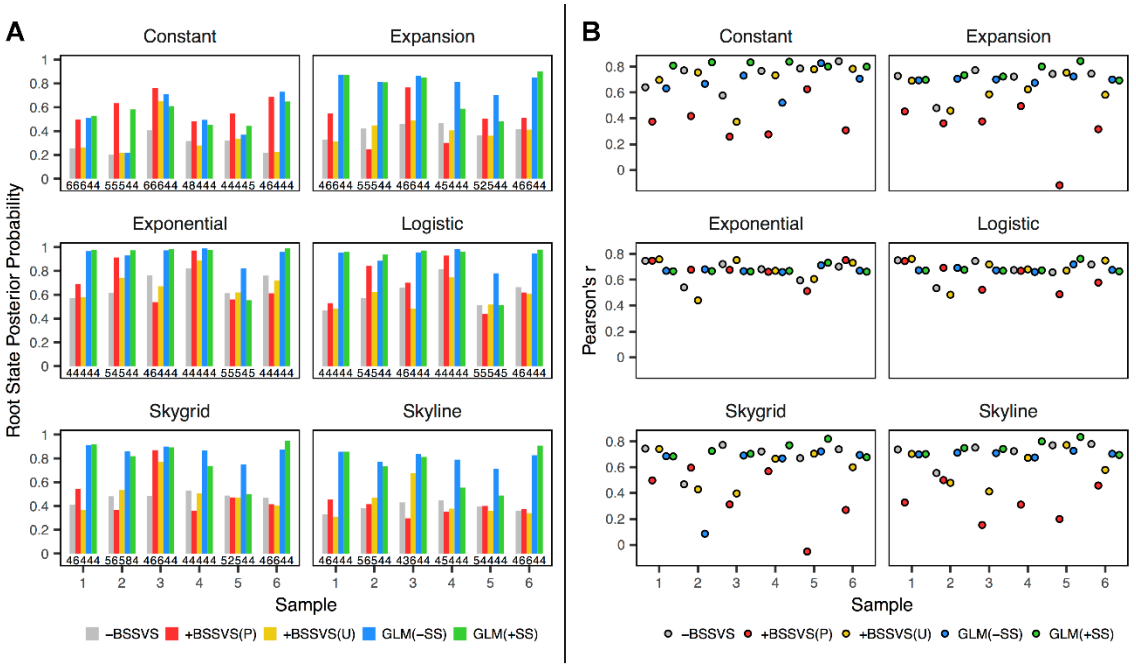

Figure 2.5. Individual root state posterior probabilities and potential sampling bias analyses. (A) Root state posterior probability from the MCC tree of each model. The corresponding root state is shown below each bar. See Figure 2.8B for the locations of these root states. (B) Pearson's r correlation coefficient between the number of sequences per discrete state and the RSPP for each discrete state in each model.

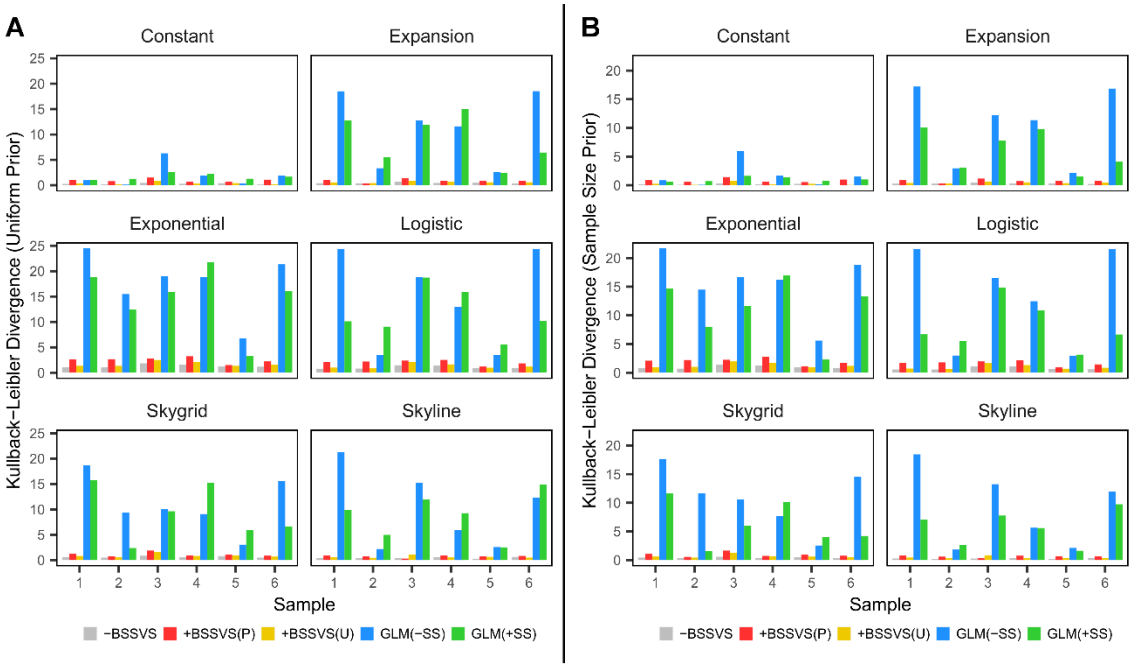

Figure 2.6. Individual Kullback-Leibler divergence statistics of the root state prior and posterior probabilities for each model. (A) Values are calculated assuming a uniform prior probability per discrete state. (B) Values are calculated assuming a prior probability proportional to the number of sequences per discrete state.

I summarize the identified root states of the four methods in Table 2.1. Here, the –BSSVS method identified three different regions, with the majority occurring in Region 4, while Region 5 is identified in over 30% of –BSSVS models. The +BSSVS(P) method identified six different regions as the root state, with Regions 6 and 4 representing the most frequently-identified. The +BSSVS(U) method identified Region 4 in nearly half of the models while Regions 5 and 6 account for the remainder of models. Comparatively, 35 of the 36 GLM(–SS) runs identified Region 4 as the root state, with the lone exception being Sample 2 using the Skygrid coalescent prior, which identified Region 8. For the GLM(+SS) analyses, Region 4 is identified as the root state in 33 of 36 models while Region 5 accounts for the remaining three. The root heights and corresponding Bayesian

credible intervals are similar between the three methods for each sample and each coalescent prior (Figure 2.7).

Table 2.1

Frequencies of the root states identified in the MCC tree under each reconstruction method

	Root State									
Method	1	2	3	4	5	6	7	8	9	10
–BSSVS	–	–	–	23	11	2	–	–	–	–
+BSSVS(P)	–	2	1	10	6	16	–	1	–	–
+BSSVS(U)	–	–	–	17	10	9	–	–	–	–
GLM(–SS)	–	–	–	35	–	–	–	1	–	–
GLM(+SS)	–	–	–	33	3	–	–	–	–	–

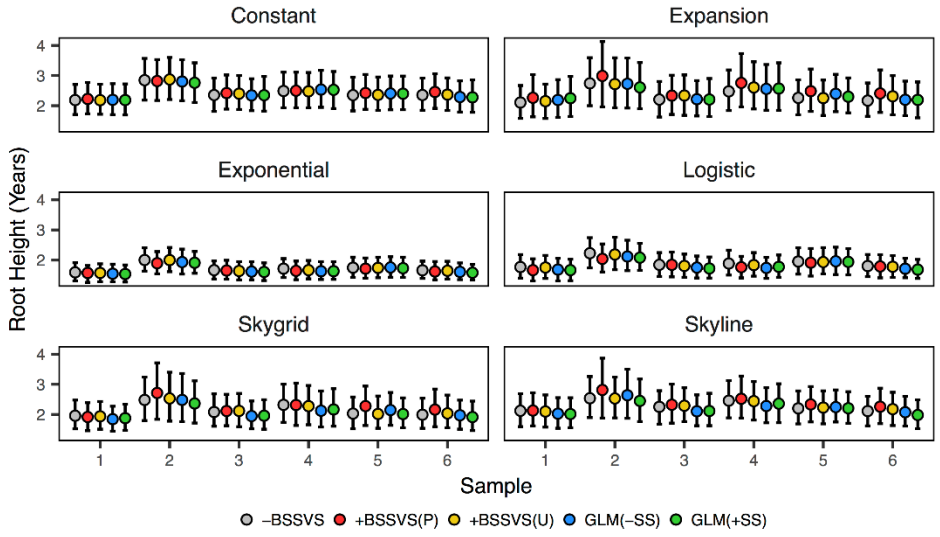

Figure 2.7. Root heights for the MCC phylogenies. Mean heights are represented by the colored circles with 95% Bayesian credible intervals shown as error bars.

As influenza viruses rarely persist for more than one season, except in tropical areas (Rambaut et al., 2008; Viboud, Alonso, & Simonsen, 2006), I obtained the geographic distribution of the number of internal nodes with a height of at least one year

(NH1s) from the MCC tree of each model and show these data in Figure 2.8A. From Figure 2.8A, the –BSSVS method indicates that Region 4 contains the greatest number of NH1s under each prior, while Region 5 contains the second-largest volume of NH1s. The +BSSVS(P) method shows Region 4 containing the most NH1s for the exponential, logistic, Skyline, and Skygrid coalescent priors, with Region 6 accounting for the next largest volume in the latter three priors. Under the constant coalescent prior, a nearly equal amount of NH1s are observed in Regions 4, 6, and 8, while the expansion prior shows Region 5 containing the largest number of NH1s. For the +BSSVS(U) method, the NH1s are most commonly observed in Region 4 under each coalescent prior, with Regions 5 and 6 primarily accounting for the remaining nodes. The frequency of NH1s in Region 8 are low under this method, but do occur under the constant, expansion, and Skygrid coalescent priors. Finally, the NH1s are largely concentrated in Region 4 for both the GLM(–SS) and GLM(+SS) methods under each coalescent prior.

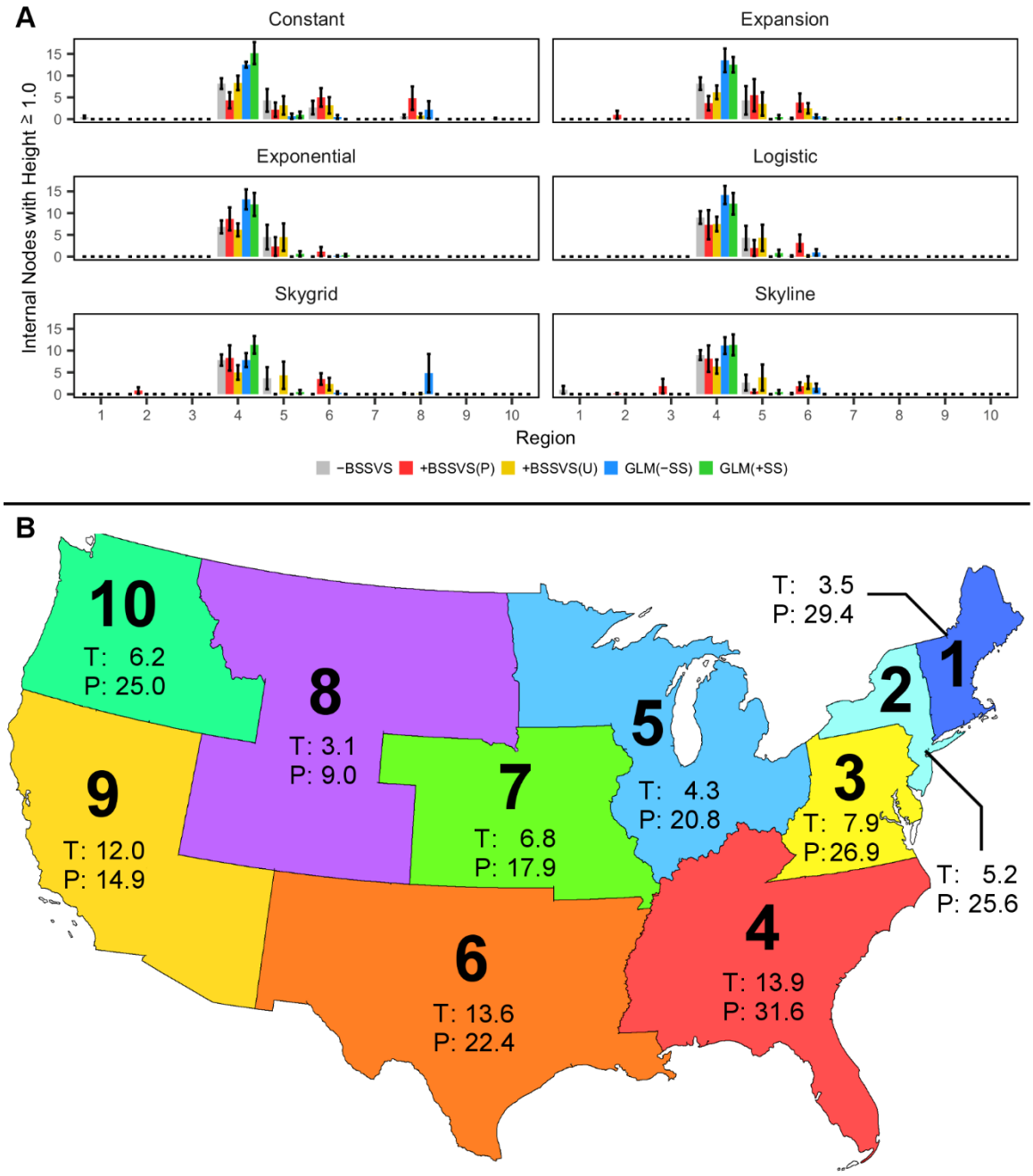

Figure 2.8. Geographic trends in coalescent events. (A) The number of internal nodes with a height of at least one year in age (NH1s) under each method and for each coalescent prior. Bars represent the total number of such nodes across all six samples. (B) Map of the contiguous U.S., colored by the ten discrete states used in this study. Each region is annotated with its average temperature (T, in ˚C) and precipitation (P, in cm)

during the September – May months. Temperature and precipitation data represent the point estimates used in our GLMs for those respective predictors.

The frequent identification of Region 4 as the root state (Table 2.1) and location of NH1 events (Figure 2.8A) indicates that there is likely at least one local variable playing a role in the tree topologies. Given this, from Figure 2.8B I note that Region 4 exhibits both the highest expected temperature and precipitation during a typical flu season as I compare the posterior support of all predictors for both the GLM(–SS) and GLM(+SS) methods in Figure 2.9.

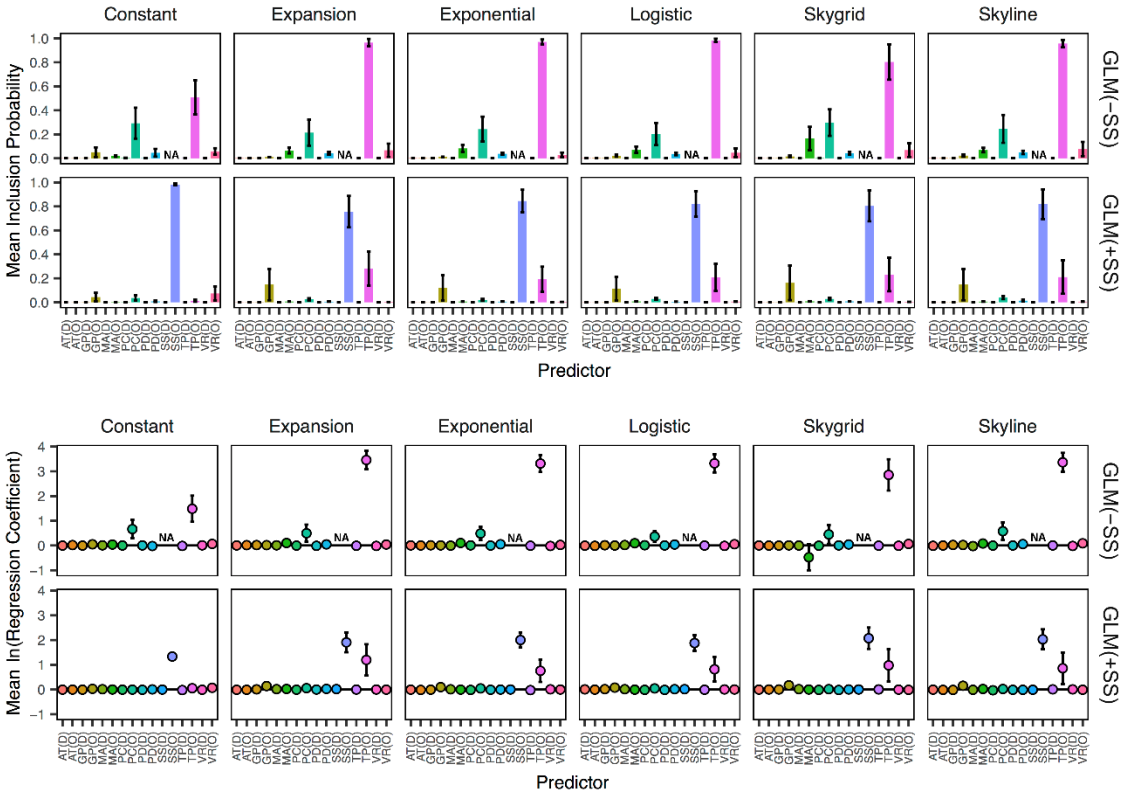

Figure 2.9. Mean posterior estimates of supported predictors. I show the inclusion probabilities and regression coefficients for all predictors for both the GLM(–SS) and

GLM(+SS) analyses. Point estimates represent the mean of each statistic across the six models for each prior, with error bars representing the standard error of these estimates. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR).

From Figure 2.9, sample size at the region of origin (SS(O)) is strongly supported for the GLM(+SS) runs with Bayes factor (BF) > 69 for each coalescent prior and with each corresponding mean regression coefficient greater than 1.33. The predictor with the second largest support for inclusion in the GLM(+SS) runs is temperature at the region of origin (BF > 5 and regression coefficient > 0.75 for each prior except constant size), followed by glycoprotein at the region of origin (3.0 < BF < 4.5 for the expansion, exponential, Skyline, and Skygrid coalescent priors) although the respective mean regression coefficients for glycoprotein remain near zero. For the GLM(–SS) runs, temperature at the region of origin yields the largest mean posterior inclusion probability across all coalescent priors (BF > 20 for each prior, BF > 400 for the expansion, exponential, logistic, and Skyline priors) followed by precipitation at the region of origin (5.0 < BF < 8.5 for all priors). Mean posterior estimates of the corresponding regression coefficients and their standard errors indicate strictly positive values for these two predictors in the GLM(–SS) runs, although the 95% highest posterior density (HPD) of the regression coefficient for precipitation at the region of origin spans zero for each model (Figure 2.10). If the entire HPD lies on the positive side of zero, this suggests that the predictor is driving the diffusion of the virus. Conversely, if the entire HPD lies on the negative side of zero, this suggests that the predictor is preventing the diffusion. Thus,

I show the proportion of GLMs in which the absolute value of the HPD is positive in Table 2.2. The 95% HPDs of temperature at the region of origin are strictly positive in 26 of the 36 GLM(–SS) runs and span zero in the remaining ten. The glycoprotein predictor at the region of origin finds the highest mean support for the constant prior (BF = 1.1), which is a sharp turn from the GLM(+SS) runs. See *Materials and Methods* for more information on metrics of support and interpretations of our predictors. I show the posterior regression coefficients and inclusion probabilities of every predictor from each of the 36 GLM(–SS) runs in Figures 2.10 and 2.11, respectively, and corresponding data for the 36 GLM(+SS) runs in Figures 2.12 and 2.13, respectively.

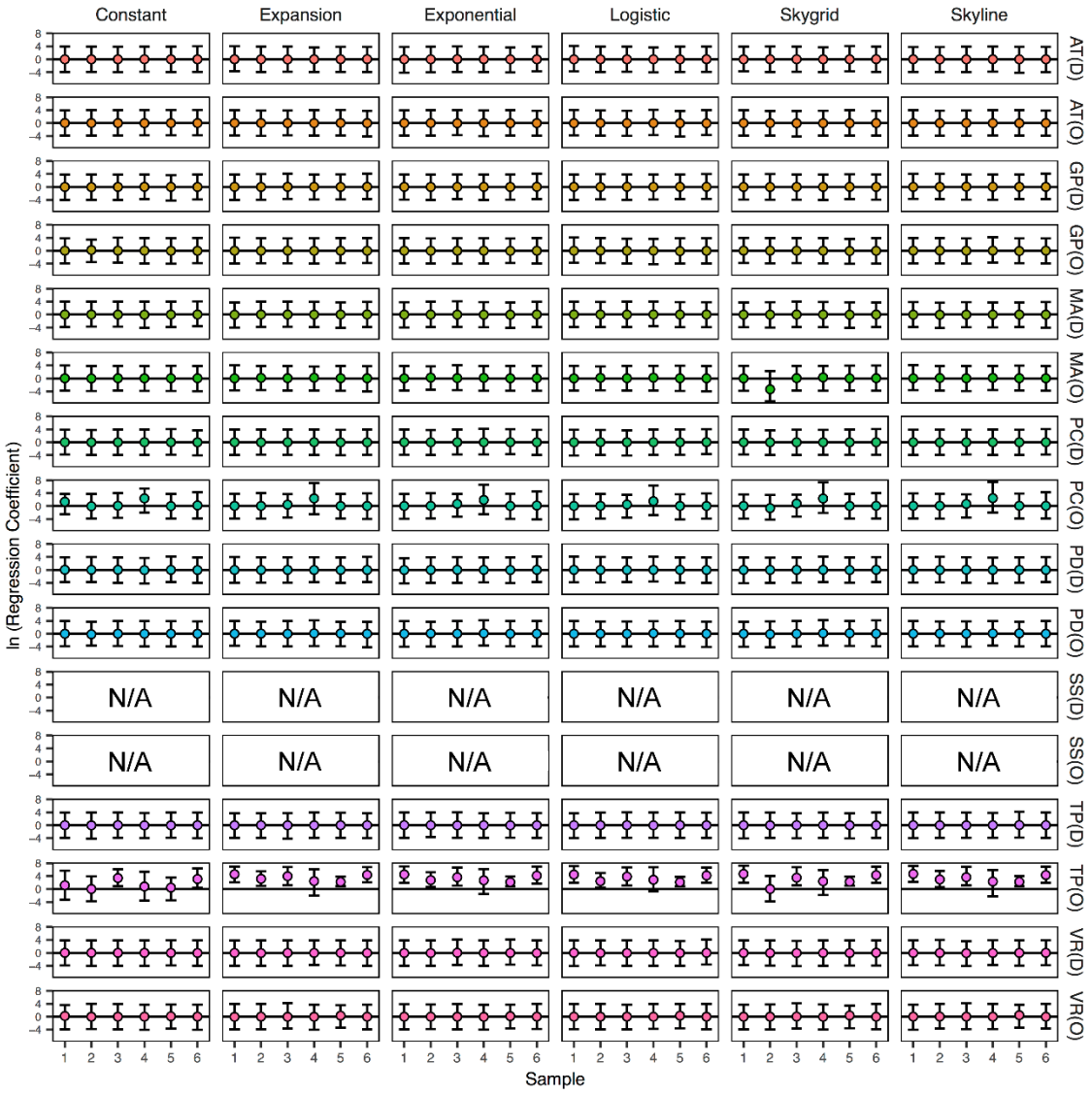

Figure 2.10. Posterior inclusion probabilities of all predictors per sample and prior for the GLM(–SS) runs. I consider predictors with inclusion probabilities exceeding the dotted horizontal line, which corresponds to BF = 3.0, to be supported in that model. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR), each evaluated from both region of origin (O) and region of destination (D).

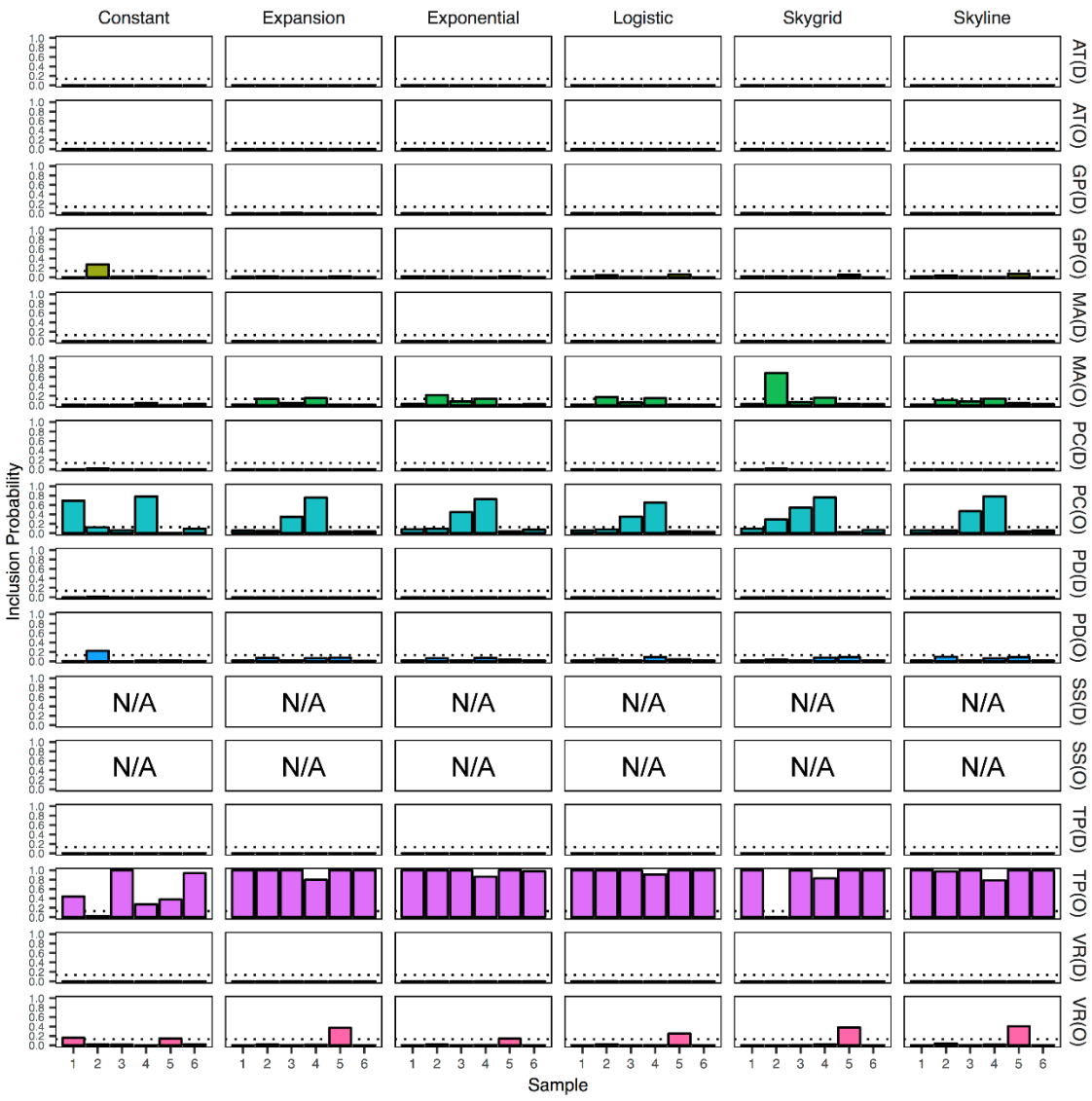

Figure 2.11. Posterior regression coefficients of all predictors per sample and prior for the GLM(–SS) runs. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR), each evaluated from both region of origin (O) and region of destination (D).

Figure 2.12. Posterior inclusion probabilities of all predictors per sample and prior for the GLM(+SS) runs. I consider predictors with inclusion probabilities exceeding the dotted horizontal line, which corresponds to BF = 3.0, to be supported in that model. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR), each evaluated from both region of origin (O) and region of destination (D).

Figure 2.13. Posterior regression coefficients of all predictors per sample and prior for the GLM(+SS) runs. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR), each evaluated from both region of origin (O) and region of destination (D).

Table 2.2

Frequency of GLM predictor support

		Predictor at the Region of Origin							
Method	Criterion	AT	GP	MA	PC	PD	SS	TP	VR
GLM(–SS)	BF ≥ 3	–	3%	25%	36%	3%	NA	94%	19%
GLM(+SS)	BF ≥ 3	–	17%	–	3%	–	97%	36%	3%
GLM(–SS)	|95% HPD (β)| > 0	–	–	–	–	–	NA	72%	–
GLM(+SS)	|95% HPD (β)| > 0	–	3%	–	–	–	61%	8%	–

Notes: Values represent the percentage of models that show BF support for a predictor and the percentage of 95% HPD intervals of the regression coefficient that do not span zero. Predictor abbreviations are: air travel (AT), glycoprotein content (GP), median age (MA), precipitation (PC), population density (PD), sample size (SS), temperature (TP) and vaccination rate (VR).

Discussion

In this paper, I compared three ancestral state reconstruction frameworks and five total methods using six randomly-drawn sequence samples and six coalescent priors for a total of 180 models while fixing the nucleotide substitution process for each. I compared each of our analyses with established model selection techniques (Baele et al., 2012; Baele, Li, Drummond, Suchard, & Lemey, 2013) and compared features of each model's MCC tree to identify posterior statistical support and discrepancies in the phylogeographic reconstructions. Regarding model selection, I found that PS shows the most posterior support for either the GLM(–SS) or GLM(+SS) in 34 of 36 runs (with one –BSSVS and one +BSSVS(U) accounting for the remaining two), while SSS shows the most support for 29 of 36 –BSSVS models, five GLM(+SS), one GLM(–SS), and one +BSSVS(U). Each GLM(–SS) and GLM(+SS) outperformed its corresponding +BSSVS(P) under both PS and SSS. Both statistics agree that +BSSVS(P) models offered the poorest posterior support, as 72% of PS analyses and 89% of SSS analyses (81% combined) show the +BSSVS(P) model as the least-supported among the five

frameworks (Figures 2.1A and 2.2), although I note that no framework shows significantly more support than any other framework for PS or SSS via t-tests.

Although the –BSSVS method is highly supported under SSS, the method fails to find strong support regarding both RSPP and KL divergence (Figures 2.4C, 2.4D, 2.5A, and 2.6) compared to the other methods. The RSPPs using the –BSSVS method are significantly lower than those obtained via the GLM(–SS) method ($p = 0.03$ for the constant coalescent prior, $p < 0.001$ for the expansion, exponential, logistic, Skygrid, and Skyline coalescent priors), while the GLM(–SS) also show a significant increase for KL divergence for both the uniform and sample size assumptions over the –BSSVS models under each coalescent prior except for constant size. Similarly, the GLM(+SS) method shows significantly greater RSPPs and both KL divergences than the –BSSVS models ($p < 0.03$ for all coalescent priors except constant). Meanwhile, the +BSSVS(P) method finds significantly greater RSPPs than the –BSSVS method only under the constant coalescent prior ($p < 0.001$) and significantly greater KL divergences over the –BSSVS method under each coalescent prior, each with $p < 0.03$. The +BSSVS(P) method also found significantly greater KL divergences for the constant, exponential, and logistic coalescent priors. The +BSSVS(U) method only found significantly greater support over the –BSSVS method via KL with the sample size assumption for the expansion coalescent prior. While these results show that the –BSSVS method finds poor statistical support at the identified root state, I also found that both the GLM(–SS) and GLM(+SS) methods in turn significantly outperformed both the +BSSVS(P) and +BSSVS(U) models for KL divergence under both prior assumptions under five of the six coalescent priors (excluding constant). The GLM(–SS) runs also found significantly greater RSPPs than

the +BSSVS(P) and +BSSVS(U) under each coalescent prior except constant, while the GLM(+SS) runs found significantly greater RSPPs than the +BSSVS(P) and +BSSVS(U) methods for the expansion, Skygrid, and Skyline priors.

The association index of each model obtained via BaTS (Figure 2.3) demonstrate a strong association between sampling location and the phylogeny for each of the 180 models, which suggests that the diffusion was spatially-structured. Some of the phylogeny-location association can be attributed to the smaller amount of genetic diversity in sequences from the same region (Figure 2.1B), however the statistical significance of the intra- and inter-region genetic distances could not fully account for the differences in RSPP and KL divergence, regardless of the coalescent prior. Furthermore, Region 4 was the most frequently-identified root state for the –BSSVS, +BSSVS(U), GLM(–SS), and GLM(+SS) methods, the second most frequently identified root state for +BSSVS(P) method (Table 2.1), and was also the location of the most NH1s (Figure 2.8A). These NH1s are biologically important for seasonal influenza, as these viruses typically experience bottlenecking at this height as part of a sink-source ecological dynamic (Bahl et al., 2011; Rambaut et al., 2008; Viboud, Bjornstad, et al., 2006). As Region 4 experiences the highest temperature and most precipitation during flu season, at 6.9˚C warmer and 10.3 cm wetter, respectively, than the remaining nine regions (Figure 2.8B) I describe it as the most "tropical" in the U.S. during a typical flu season. This provides a well-supported explanation for the observed trends in Region 4, especially under both GLM methods. As the data for the GLM(–SS) and GLM(+SS) runs indicate strong support for temperature at the region of origin (Figure 2.9), our results would

suggest that Region 4 is the most likely origin of each of the six samples using those two methods.

This conclusion, however, is hindered by the strong sampling bias exhibited by the GLM(–SS), and GLM(+SS) methods. These two methods (as well as the –BSSVS and +BSSVS(U)) demonstrate consistently strong, positive Pearson's r correlation coefficients between the root state posterior probability and sample size at each discrete state, regardless of coalescent prior (Figures 2.4B and 2.5B). Furthermore, the inclusion of the sample size predictors in the GLM(+SS) runs shows that sample size at the region of origin is strongly influencing its posterior estimates, with 35 of 36 runs showing BF > 3 and 22 of 36 showing a positive 95% HPD on the regression coefficient (Table 2.2, Figures 2.10 and 2.11). The mean posterior inclusion probability for the sample size predictor at the region of origin corresponds to BFs of 1317.9, 70.0, 122.9, 102.7, 92.6, and 101.8 for the constant, expansion, exponential, logistic, Skygrid, and Skyline priors, respectively. Given the similarities in RSPP, Pearson's r, and KL data between the GLM(–SS) and GLM(+SS) runs (Figures 2.4-2.6), I believe that sample size is influencing the GLM(–SS) runs to a similar degree, although its BF support cannot be measured. Thus, although both GLM methods presented in this paper are providing biologically justifiable and statistically supported evidence regarding the diffusion of this influenza virus over our selected time period, the strong sampling biases give us pause. Instead, the significant decrease in Pearson's r for the +BSSVS(P) models from the other four methods under the constant, expansion, and Skyline coalescent priors provide more confidence in those data, despite its poor performance with respect to log marginal likelihoods via PS and SSS (Figures 2.1A and 2.2).

I compared the –BSSVS, +BSSVS(P), +BSSVS(U), GLM(–SS), and GLM(+SS) methods for modeling a single discrete trait, sampling location, which highlighted differences in diffusion of seasonal influenza in the U.S. Our results collectively indicate that the GLMs provide the strongest posterior support for MCC metrics of the three ancestral state reconstruction frameworks used in this study, however the strong sampling bias exhibited by that method reduces confidence in their reconstructions. As mentioned, the strong support for sample size is consistent with previous studies that used the phylogeographic GLMs (Lemey et al., 2014; Magee et al., 2015). Air travel was previously shown to be a driver of the global diffusion of H3N2 using a GLM (Lemey et al., 2014), but none of the GLM(–SS) or GLM(+SS) runs showed support for this predictor. However, our study was performed within a single country and aggregated all air travel data from each individual state into a matrix of region-to-region passenger flux, which perhaps limits its contribution to these models. Furthermore, the paper by Lemey *et al.* (Lemey et al., 2014) discretized by "air communities" to better reflect trends in air travel, while I partitioned strictly based on pre-defined, arbitrary geographic regions. I also assumed a single introduction into the U.S. and did not include incoming travel from international flights that could certainly have introduced strains with more genetic diversity than those used in this study.

I recognize several limitations with this study including the omission of international air travel. In addition, our assumption of a single introduction into the U.S. could also have limited inference regarding the contribution of air travel and may explain the lack of BF support for that predictor from both region of origin and destination when a previous study has implicated these data as a driver of the diffusion (Lemey et al.,

2014) . Also, the transportation predictor fails to incorporate inter-region travel via ground transportation, which certainly could have implications within a single country. Furthermore, I only analyzed hemagglutinin sequences in this study and did not investigate neuraminidase or any other segments of the influenza genome. I arbitrarily selected 25% of samples from each region for our subsampling to better reflect the observed sampling frequencies, but it is possible that larger subsample sizes or an alternative sampling approach could have resulted in stronger or weaker support for the predictors in the GLM as well as the RSPPs via the three reconstruction approaches. However, my use of Pearson's correlation coefficient between sample size and root state posterior probability (Figures 2.4B, 2.5B) and comparison of GLMs that include and do not include sample size predictors aim to outline the impact of sampling bias within our dataset. I plan to conduct similar research on additional influenza seasons and using alternative sampling methods to further study whether this sampling bias is a systematic function in the GLMs or is limited to the dataset used in this study. Sampling bias is a known issue in phylodynamics (Baele, Suchard, Rambaut, & Lemey, 2016; Frost et al., 2015) and may not be possible to eliminate, although varying approaches may differ in their sensitivity to such biases. Finally, I limited our study to a single influenza season which prevents seasonality comparisons and impacts from local persistence.

Overall, this study aimed to investigate the phylogeography of the H3N2 influenza viruses that circulated in the U.S. during the 2014-15 flu season and to also investigate three established methods of ancestral state reconstruction. While our GLM results provide superior posterior support than either +BSSVS method or the –BSSVS framework, these results appear to be dominated by a strong sampling bias. Although

these results are not necessarily incorrect, the investigation of additional frameworks reveals that the +BSSVS(P) is likely the "best" approach for this dataset to minimize such concerns, depending on the selection of coalescent prior, if given the choice among the five presented in our work for this virus and time frame. Furthermore, I demonstrate that our approach of subsampling to compare multiple models may not only reflect subtle changes to the phylogeny but also to the contribution of the predictor variables in the GLMs. Although I do not believe that the GLM provides an ideal, unbiased reconstruction framework for our dataset, this type of assessment could be valuable for understanding the true nature of the phylogeny-sampling location association in future work. Such studies may also encourage researchers to utilize the GLM framework as a means of obtaining more information-driven variables into their phylogeographic studies and to unlock the potential for more accurate ancestral state reconstructions to better aid epidemiological and public health efforts.

Materials and Methods

Sequence and Model Setup

Nucleotide Sequences. I used the EpiFlu database from the Global Initiative for Sharing All Influenza Data (GISAID) to collect H3N2 hemagglutinin (HA) sequences from the 2014-15 flu season. I obtained our dataset on 2015-10-16 using the following search terms: Host = *Human*, Location = *United States*, Collection Date = *2014-09-29 to 2015-05-17*, Submitting Laboratory = *[United States, Atlanta] Centers for Disease Control and Prevention*, Required Segments = *HA*, Min Length = *1,659*. This search resulted in 1,220 sequences, and I further eliminated sequences from Alaska, Hawaii, and

the District of Columbia and those that did not have a specific state listed to obtain a final set of 1,163 sequences. In order to reduce the size of the transition rate matrix, I discretized the states into the ten U.S. Department of Health and Human Services (HHS) regions (HHS, 2014), which I show in Figure 2.8B.

Ancestral State Reconstruction Methods. Our phylogeographic assessment assumes that geographic sampling traits follow a continuous-time Markov chain (CTMC) process along the branches of an unknown phylogeny that is informed through sequence data. The models I compare differ in how one parameterizes the infinitesimal rates of the among-location CTMC process. Here, I first parametrized the discrete location trait with a basic asymmetric substitution model (–BSSVS). Next, following Lemey *et al.* (Lemey et al., 2009), I retained the asymmetric substitution model but specified a truncated Poisson prior on the number of non-zero rates (+BSSVS(P)). Here, 50% of the prior probability lies on the minimal rate configuration (*i.e.* nine non-zero rates connecting the ten HHS regions). Similarly, I also placed a uniform probability on the location prior to test the effects of the selected location prior on the BSSVS procedure +BSSVS(U). I compare the –BSSVS and +BSSVS(P) methods with recent developments in virus phylogeography that have advanced modeling of among-location transition rates as a log-linear GLM of predictors of interest (Lemey et al., 2014). Here, I followed this framework and parameterized GLMs with seven demographic, environmental, and genetic factors that I take from both region of origin and region of destination for a total of 14 predictors in the GLM(–SS) runs. In the GLM(+SS) runs I also include an additional two sample size predictors for a total of 16 predictors. This approach yields a quantifiable assessment of the inclusion and contribution of each predictor variable to the

overall transition rate matrix between our ten locations by estimating posterior probabilities of all 2^{14} or 2^{16} possible linear models via a BSSVS procedure. I specified a 50% prior probability that no predictor will be included to enable calculation of Bayes factors (BFs) as a metric of support for the inclusion or exclusion of any given predictor. Here, I consider any predictor with BF > 3.0 to be supported for inclusion. For further details on the underlying theory and mathematical definitions of this GLM approach, I refer readers to Lemey *et al.* (Lemey et al., 2014).

Summary of Rate Parameters. For both the –BSSVS and +BSSVS frameworks, there are K(K–1) relative rate parameters where K = 10 discrete states for our dataset [1]. For the –BSSVS framework, these rate parameters are each a priori independently gamma distributed with scale and shape parameters of 1.0, while for the +BSSVS framework these rate parameters are each a priori with a mixture of a point-mass on 1.0 and on the same gamma distribution as the –BSSVS rate parameters. The number of parameters that achieve the point mass on 1.0 for the +BSSVS framework are Poisson distributed with a mean of 9.0 (for the +BSSVS(P) method) and uniformly distributed for the +BSSVS(U) method (i.e. a uniform distribution on [K, K(K-1)] = [9, 90]). For the GLM framework, there are 14 and 16 regression parameters (*i.e.* predictors) for the GLM(–SS) and GLM(+SS) methods, respectively, as outlined below. The regression parameters are each a priori in part a mixture of point-mass on 0 and in part normally distributed with a mean of 0 and a variance of 4.0 (Lemey et al., 2014).

Sequence Subsampling. To investigate the effects of sampling biases, I performed multiple analyses using random samples from our full set of 1,163 sequences. I created six independent sequence samples by selecting 25% of the sequences in each

region at random without replacement and assume that each is representative of the entire flu season. These samples allow us to reveal whether the three frameworks will agree on the root location, root state posterior probability, height, and other trends in the phylogenies as well as show the reproducibility of the support for our GLM predictor variables. I did not identify any duplicate sequences from the same discrete state in any of the six samples. I aligned these six samples, each of which contained 285 sequences, using MAFFT v7.017 in Geneious Pro v.6.1.8 (Biomatters Ltd., Auckland, New Zealand). I treated each alignment as an independent dataset for our phylogeographic reconstructions and report all GISAID accession numbers and discrete state assignments (*i.e.* HHS regions) in Appendix B. The six samples and six coalescent priors result in a total of 180 total models, 36 from each of the –BSSVS +BSSVS(P), +BSSVS(U), GLM(–SS), and GLM(+SS) methods.

GLM Predictors

Human Population and Age. I obtained population estimates and land area per state from the U.S. Census Bureau (USCB) MAF/TIGER® database (https://www.census.gov/). Population data are released annually and represent the population as of 2014-07-01 for the 2014-15 flu season, and I used these values to create a density per region. I also obtained the median age per state from the USCB and used these values as a separate predictor, aggregated by region.

Temperature and Precipitation. For our climate predictors, I obtained data from the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NOAA). I collected temperature and precipitation data for the 30-year climate normal from 1981-2010 for the 9,359 stations in the contiguous 48 states, not

including the District of Columbia. As I am interested in the typical temperatures and precipitations observed during a flu season, I computed the average of all September-October-November, December-January-February, and March-April-May summary datasets from stations in each region. I take these values for temperature (in degrees Celsius) and precipitation (in centimeters) to represent the typical flu season climate for each region.

Influenza Vaccination Rates. I obtained state-level data on the vaccination rates for the 2014-15 flu season from FluVaxView by the Centers for Disease Control and Prevention (CDC) (CDC, 2016a) and aggregated them to a region-wide average. These data represent all individuals at least six months of age that received the annual flu vaccine at any point in time during the season.

Air Travel. To account for travel between the ten regions, we obtained data from the Official Airline Guide, Ltd. as the number of seats on domestic flights between each pair of airports within the contiguous U.S. for the 2012 calendar year. I assumed that the number of seats is proportional to the number of passengers on each flight and that the 2012 travel data is proportional to that of 2014-15. I discretized the data from each individual airport into a total number per HHS region to create a matrix of travel flux. These data do not include flights originating from international locations and thus strictly represent passenger flux among the ten HHS regions used in this study. I held this predictor constant through each of the six samples.

Glycoprotein Content. Influenza vaccines are designed to induce neutralizing antibodies of both the hemagglutinin and neuraminidase viral surface glycoproteins (Cobbin, Verity, Gilbertson, Rockman, & Brown, 2013) in order to protect against future

infections with similar antigenic properties to the vaccinated strain (Couch & Kasel, 1983). The glycoprotein (GP) content of a sampled virus thus provides an indication of the sample's similarity to the strain vaccinated against during that season. Of the 1,163 sequences in our dataset, 533 (46%) contained metadata regarding the GP content of the sample. The authors annotated these sequences with the binary "LOW GP" or "GP" to represent the similarity of the GP to the A/Texas/50/2012 (H3N2)-like virus strain vaccinated against during the 2014-15 flu season (CDC, 2016b). For each sample, I calculated the proportion of sequences with "LOW GP" to the total sequences with known antigenic content per region as a measure of the circulating strain's disparity from the strain vaccinated against. This is the only predictor in which the values are not fixed among the six samples.

Sample Size. Previous phylogeographic studies using GLMs have included and found strong posterior support for sample size at the location of origin and/or the location of destination (Lemey et al., 2014; Magee et al., 2015) so I included both as predictors in the GLM(+SS) runs. The GLM(+SS) runs thus contain 16 predictors while the GLM(–SS) run contain 14 predictors.

Table 2.3

Descriptive statistics of each predictor for the ten discrete states

Predictor	Mean	SD	Median	IQR
Population Density (people/mi^2)	165.9	141.0	143.9	161.3
Median Age (years)	38.0	1.6	37.8	2.0
Vaccination Rate (%)	42.6	3.5	43.2	4.5
Temperature (˚C)	7.7	4.1	6.5	6.5
Precipitation (cm)	22.4	7.0	23.7	8.2
Low GP Content (%, overall)	88.3	3.7	87.8	3.1
Sample Size [a]	28.5	11.5	27.5	16
Air Travel [b]	6.1×10^6	6.0×10^6	4.1×10^6	6.7×10^6

[a] Accession numbers for the samples and location data are provided in Appendix B
[b] Air travel represents the indicated statistic among all 90 pairwise region-to-region combinations

Influenza Phylogeography

Molecular Clock Fitting. I performed a preliminary analysis with Path-O-Gen v1.4 (http://tree.bio.ed.ac.uk/software/pathogen/) which showed that relaxed molecular clocks may have overparameterized our models. I therefore selected a strict molecular clock with a rate of 0.001 substitutions per site per year.

Coalescent Priors and Substitution Model. In addition to the three reconstruction methods and six sequence samples, I also investigated six coalescent priors in this study: constant size (Kingman, 1982), exponential growth (Griffiths & Tavare, 1994), logistic growth (Griffiths & Tavare, 1994), expansion growth (Griffiths & Tavare, 1994), Bayesian Skyline (Drummond, Rambaut, Shapiro, & Pybus, 2005), and Bayesian Skygrid (Gill et al., 2013). Thus, I completed 180 individual ancestral state phylogeographic reconstructions, one for each sample/coalescent prior/reconstruction method combination (*e.g.* Sample 1/constant size/GLM, Sample 1/constant size/+BSSVS(P), Sample 1/constant size/–BSSVS, etc.). I specified an HKY+G

(Hasegawa, Kishino, & Yano, 1985) substitution model following recent phylogenetic studies of H3N2 (Horm et al., 2014; Lemey et al., 2014) and preliminary performance analyses using other substitution models. I used the–BSSVS, +BSSVS(P), +BSSVS(U), GLM(–SS), and GLM(+SS) methods to perform phylogeographic reconstructions under these parameters using the BEAST v1.8.4 software package (Drummond et al., 2012) with a chain length of 100 M, logging estimates every 10,000 steps while specifying a single seed across all models. These methods aim to minimize all sources of variance but the randomly selected sequences, tree priors, and glycoprotein content.

Analysis of Support for Models. I used path sampling (PS) and stepping-stone sampling (SSS) to estimate marginal likelihoods of each model, as this procedure has been shown to be an improvement over harmonic mean estimators (Baele et al., 2012; Baele et al., 2013). Here, I specify a chain length of 1M with 100 path steps, logging every 1,000 steps. For the GLM predictors, I obtained the mean posterior probability of inclusion, BF support values, and the contribution of each predictor to the log-linear rate matrix. To determine the impact of geography on the phylogeny, I utilized Bayesian Tip-association Significance Testing (BaTS) (Parker et al., 2008). This application tests the null hypothesis that other than by chance, adjoining tips are not more likely to share the same discrete traits. Here, I used our ten HHS regions as discrete traits to be tested under this null hypothesis.

Comparison of Phylogenies. I used TreeAnnotator v1.8.4 to construct a maximum clade credibility (MCC) tree for each of the 180 runs after discarding the first 10% of trees as burnin. I viewed and annotated the MCC trees using FigTree v1.4.2 for direct comparison of the ancestral state reconstructions. From each MCC tree, I recorded

the root state, root height and its 95% Bayesian credible interval, root state posterior probability, and the location of all nodes with a height exceeding one year. I also calculated the Kullback-Leibler (KL) divergence at the root state of each model. Here, I assumed two different prior probabilities at each discrete state: a uniform prior probability per discrete state (*i.e.* 0.1 for each of the ten discrete states), and second, a prior probability that is proportional to the number of taxa from that state (*e.g.* as 26 of 285 taxa were sampled in Region 1 I set its prior probability to 26/285 = 0.0912). The latter approach allows us to account for potential sampling bias in the KL calculations. For several GLMs, I found that the posterior probability of at least one root state was zero, which yields a KL divergence of infinity. To present a finite KL value, I assigned these states a posterior probability of 1.0×10^{-16} and subtracted this artificial probability from the most probable root state. As an additional step to investigate possible sampling bias, I calculated the Pearson correlation coefficient (r) between the sample size for each of the ten discrete states and its corresponding root state posterior probability for each individual model.

Data Availability. I have made the XML file and MCC phylogeny for each of our 180 models available for download at https://figshare.com/projects/Magee-Flu-PLoS/16638. I have also made available the six sequence alignments as well as the full set of 1,163 unaligned sequences from which I created our samples.

CHAPTER 3

THE EFFECTS OF SAMPLING LOCATION AND PREDICTOR POINT ESTIMATE CERTAINTY ON POSTERIOR SUPPORT IN BAYESIAN PHYLOGEOGRAPHIC GENERALIZED LINEAR MODELS

Introduction

Ancestral state reconstruction has long been an important topic in phylogenetic research (Slatkin & Maddison, 1989). Recent years have seen a turn to a Bayesian statistical framework to estimate posterior support of ancestral states (Lemey et al., 2009), making use of a Bayesian stochastic search variable selection (BSSVS) procedure (Chipman et al., 2001; Kuo & Mallick, 1998). Although this popular hypothesis testing framework is effective in identifying root locations with high probability, the posterior probability of ancestral states is drawn exclusively from genomic features. While this interpretation certainly holds value in identifying evolutionary relationships, it can be suboptimal when there is interest in characterizing the effects of suspected epidemiological factors on evolution and diffusion.

In addition to the lack of external influence on the phylogenies, studies in a discrete Bayesian phylogeographic setting must account for the nontrivial issue of identifying geographic sampling locations and, on occasion, pooling multiple locations into a single discrete state. A straightforward approach is to combine adjacent administrative divisions (*e.g.* neighboring countries) which are often divided by arbitrary boundaries, such as a parallel latitude, mountain range, or river, but these combinations may lose the value that each location holds individually. Specifically, population demographics, cultural aspects, and medical and agricultural practices may widely differ

in adjacent locations. Furthermore, these features may vary in specific areas of each individual location, so these differences should be accounted for in some capacity. Failure to do so may lead to biased posterior probability estimates along any branch in the phylogeny.

The development and implementation of a generalized linear model (GLM) in Bayesian phylogeography has enabled the modeling of transition rate matrices as a function of biologically relevant predictors (Gill et al., 2013; Lemey et al., 2014). This framework was first used to evaluate the global diffusion of H3N2 influenza (Lemey et al., 2014) and was subsequently used to assess H5N1 influenza in Egypt (Magee et al., 2015) and HIV in Brazil (Graf et al., 2015). These studies can accommodate properties of the discrete states themselves, such as demographic, environmental, and geographic features. Posterior inclusion probability estimates are available for each predictor, and Bayes factors (BFs) can be used to evaluate the support for each predictor's role in the spatiotemporal dynamics of the pathogen. Regression coefficients are also available for each predictor such that its contribution to the overall diffusion process can be quantified. Although these implementations of the phylogeographic GLM may provide advantages in biological interpretation of phylogenies and identify driving forces behind widespread diffusion, the issue of predictor aggregation remains. Namely, each of these studies used point estimates of their predictors at high levels of spatial order, like continent or country-wide averages, often due to a lack of more specific sampling locations or the inability to assign predictor data to a more local level. While these estimates are not inherently inaccurate, the variance of a temperature predictor, for example, may be rather large when considering the local differences in climate across such a large area. This calls

into question whether a point estimate of a predictor over a large geographic area will enable accurate estimates of posterior predictor support.

In this study, I investigate the effects of aggregating predictor data for phylogeographic GLMs at different spatial scales. Specifically, I examine how changes in the accuracy of the predictor point estimates may alter their respective posterior inclusion probabilities and regression coefficients. For example, climate is known to contribute to the global source-sink dynamic of influenza viruses (Rambaut et al., 2008), but temperature and precipitation are certainly not constant throughout the regions used as discrete states in many cases. Here, I hypothesize that as point estimates of the predictors become more representative of the geographic sampling location, posterior variance of the supported predictors will be minimized. This reduction in variance should provide more confidence in ensuing biological interpretations of the pathogen-predictor relationship.

I use West Nile virus (WNV) in the U.S. as a case study to address this question and gain insight for researchers that wish to utilize the GLM framework. WNV is a vector-borne virus that first emerged in the U.S. in 1999 (Mann, McMullen, Swetnam, & Barrett, 2013; Pybus et al., 2012) and has resulted in over 41,000 human infections in the country (ArboNET, 2015). These infections occur primarily through bites of infected mosquitos of the *Culex* genus (Sardelis, Turell, Dohm, & O'Guinn, 2001), although many bird species are natural hosts (WHO, 2011). To our knowledge, there has been no prior study on WNV that has utilized a phylogeographic GLM. Here, I discretized 299 sequences of WNV by U.S. Census Bureau (USCB) regions (CBR), USCB subdivisions (CBS), state, and county of isolation and perform a separate aggregation of predictor data

at each level. I additionally perform an assessment at the county-level for each of the four CBRs. This study will critically evaluate the impact of discretization of predictor data on a phylogeographic GLM, providing researchers with empirical evidence of how variables contributing to the diffusion of viruses can change given differences in discrete state partitioning and the level at which accurate point estimates of predictor values can be obtained.

Results

At the highest level of aggregation, CBR, the GLM's predictor matrix was not of full rank, which is required to run GLM analyses in BEAST. In fact, of the 105 pairwise predictor-predictor combinations, six show a very strong linear correlation at the CBR level, which is the total number of such instances in the remaining seven models. I list highly-correlated predictors (|Pearsons' r| > 0.9) for all models in Table 3.1. From Table 3.1, 12 of the 15 predictors showed a high correlation with another predictor in at least one model, with only *Corvidae* average counts at the location of origin, distance, and unvaccinated horses at the location of destination failing to do so.

Table 3.1

Predictor combinations where |Pearson's r| > 0.9

Model	Predictor 1	Direction 1	Predictor 2	Direction 2	Pearson's r
CBR	CA	Destination	PC	Destination	–0.90
CBR	CC	Destination	PD	Destination	–0.93
CBR	CC	Origin	PD	Origin	–0.95
CBR	PC	Destination	WL	Destination	>0.99
CBR	PC	Origin	WL	Origin	>0.99
CBR	TP	Destination	UH	Destination	>0.99
CBS	PC	Origin	TP	Origin	0.95
CBS	PC	Destination	WL	Destination	0.95
Midwest	TP	Destination	TP	Origin	0.96
South	TP	Destination	TP	Origin	0.99
South	CC	Origin	PD	Origin	0.92
South	CC	Destination	PD	Destination	0.91

Notes. (CA) *Corvidae* counts; (CC) case counts; (PC) precipitation; (PD) population density; (TP) temperature; (UH) unvaccinated horses; (WL) wetlands.

Although the CBS, Midwest, and South models did exhibit some strong correlations between predictors (Table 3.1), each predictor matrix achieved full rank. For the CBS, state, and national county aggregations, each MCC phylogeny exhibits similar posterior statistics, which I summarize in Table 3.2. Specifically, the time to the most recent common ancestor (tMRCA) and its highest posterior density (HPD) places the root of the viral tree in the late 1990s while the location is identified in the Northeastern U.S. Root states for the national models show "New England", "Connecticut", and "Fairfield County, Connecticut" for the CBS, state, and national county aggregations, respectively. The root state posterior probability (RSPP) is highest for the state aggregation ($p = 0.98$) followed by the CBS and county aggregations ($p = 0.94$ and 0.86, respectively). The Kullback-Leibler (KL) divergence increases from the CBS to state to county aggregation in the three national models. For the Midwest, South, and West regional county analyses, each molecular clock rate's HPD range is larger than any of the national analyses. The

oldest sampling dates for the Midwest, Northeast, South, and West counties were 2002, 1999, 2001, and 2003, respectively, and the tMRCAs of the viral samples from these four regional models were estimated to be 2000, 1997, 1998, and 2001, respectively. The RSPPs of the South and West models ($p = 0.63$ and 0.54, respectively) are substantially lower than the those from the Midwest and Northeast models ($p = 0.99$ and 0.95, respectively), and the South model achieves the weakest KL divergence (1.52) of all models. I note a strong linear correlation between the number of discrete states and KL divergence for all seven models (Pearson's $r = 0.99$), and also between the percent identical sites and number of taxa per model (Pearson's $r = -0.99$). I provide the MCC phylogeny for the three national models in Figures 3.1-3.3.

Table 3.2

Posterior statistics of the MCC phylogenies

Model[a]	Clock Rate (95% HPD)	tMRCA (95% HPD)	Root Location	RSPP	KL
CBS	7.4×10^{-4} ($6.1\text{-}8.7 \times 10^{-4}$)	1997.5 (1995.9-1998.5)	New England	0.94	3.48
State	7.2×10^{-4} ($5.9\text{-}8.6 \times 10^{-4}$)	1997.6 (1996.1-1998.6)	Connecticut	0.98	48.27
County	6.8×10^{-4} ($5.7\text{-}7.9 \times 10^{-4}$)	1997.5 (1996.1-1998.6)	Fairfield Cty., Connecticut	0.86	233.18
Midwest	4.3×10^{-4} ($1.7\text{-}7.2 \times 10^{-4}$)	2000.3 (1997.3-2001.4)	Cook Cty., Illinois	0.99	47.70
Northeast	6.7×10^{-4} ($5.2\text{-}8.3 \times 10^{-4}$)	1997.7 (1996.4-1998.7)	Fairfield Cty., Connecticut	0.95	85.67
South	8.1×10^{-4} ($3.2\text{-}12.1 \times 10^{-4}$)	1998.8 (1996.0-2000.7)	Harris Cty., Texas	0.63	1.52
West	5.3×10^{-4} ($2.0\text{-}8.6 \times 10^{-4}$)	2001.6 (1999.0-2002.7)	Park Cty., Colorado	0.54	17.25

[a] Results not available for the CBR phylogeny as its predictor design matrix did not achieve full rank

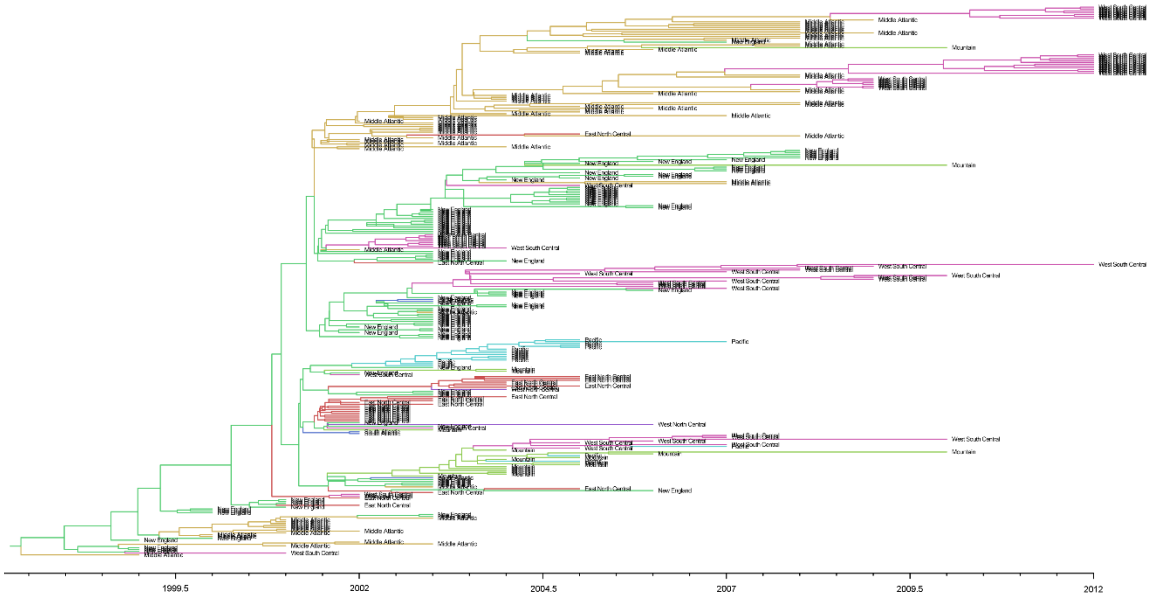

Figure 3.1. MCC phylogeny of the CBS model.

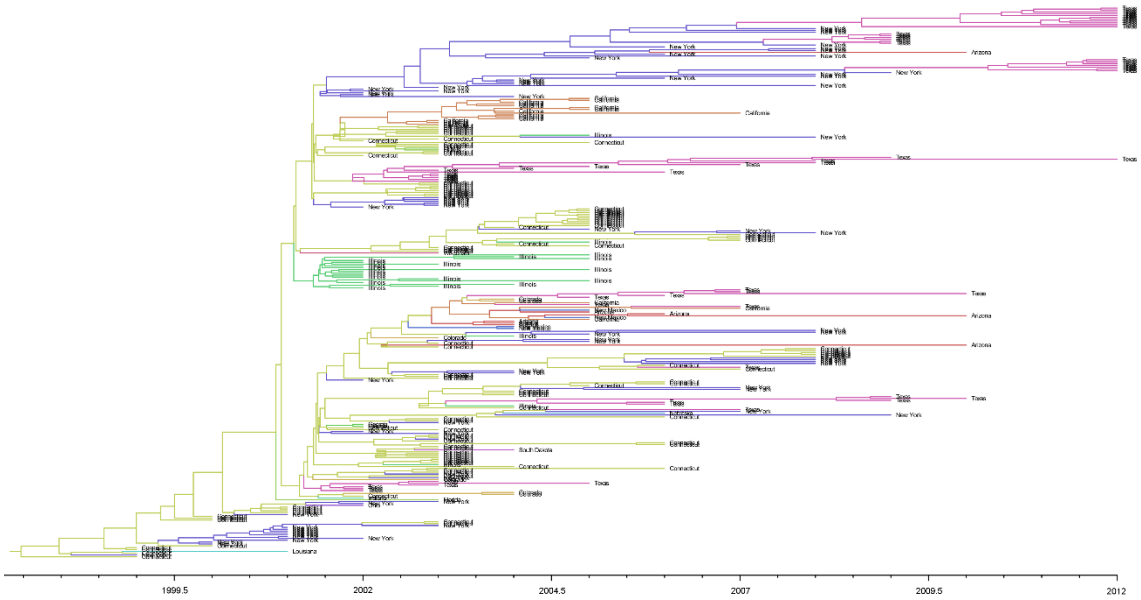

Figure 3.2. MCC phylogeny of the state model.

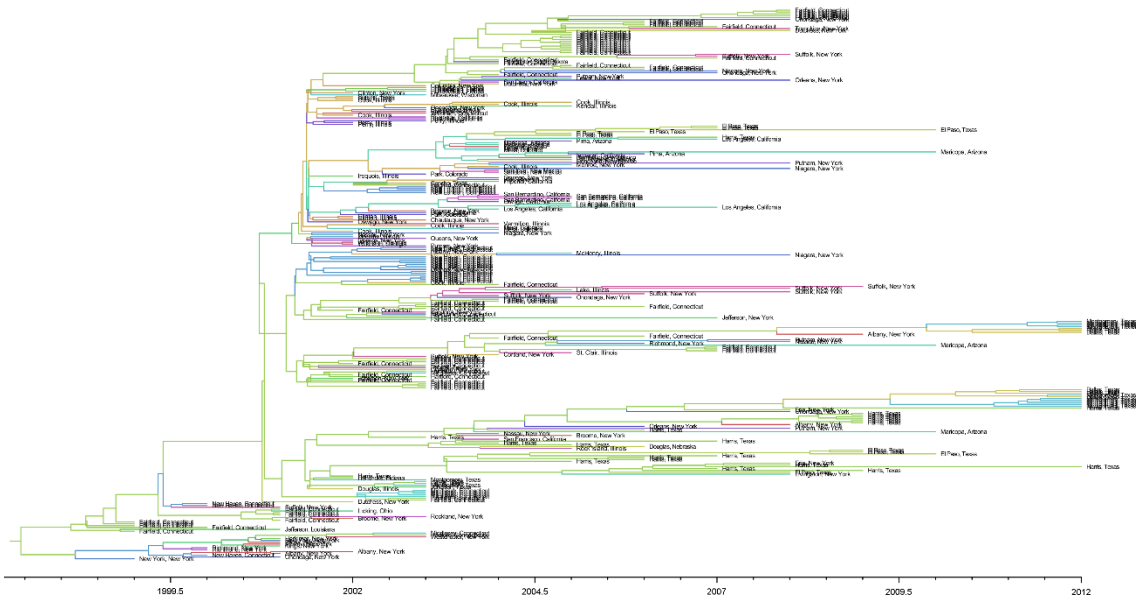

Figure 3.3. MCC phylogeny of the county model.

The three national models demonstrate similar trends in population demographics via Bayesian Skyline plots as well, which I show in Figure 3.4. From Figure 3.4, the genetic diversity shows a sharper decline in the county model than the CBS or state models near the year 2003, but the remainder of the Skylines are nearly identical among the three models. I also show the Bayesian Skyline plots for the four regional models in Figure 3.5. From Figure 3.5, the Skyline plot of the Northeast county-level model appears similar to that of the three national models (Figure 3.4) over its time frame, which is likely an artifact of the density of samples in the Northeast region compared to the other three regions. The Midwest, South, and West models show generally steady levels of diversity across their respective time periods.

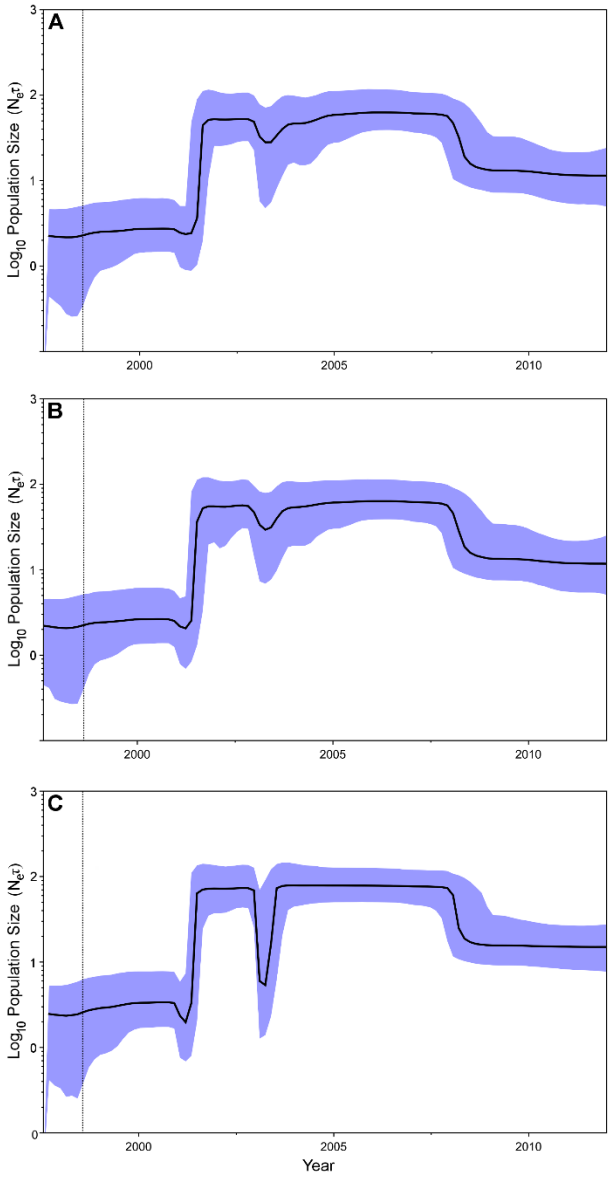

Figure 3.4. Bayesian Skyline plots for the (A) CBS, (B) state, and (C) county models. The y-axis (Neτ) is the effective population size multiplied by the generation length and the x-axis represents the year. The median measure is indicated by the thick black line, with the 95% HPD limits shown as the shaded blue area. The dotted vertical line represents the lower 95% HPD of the root height.

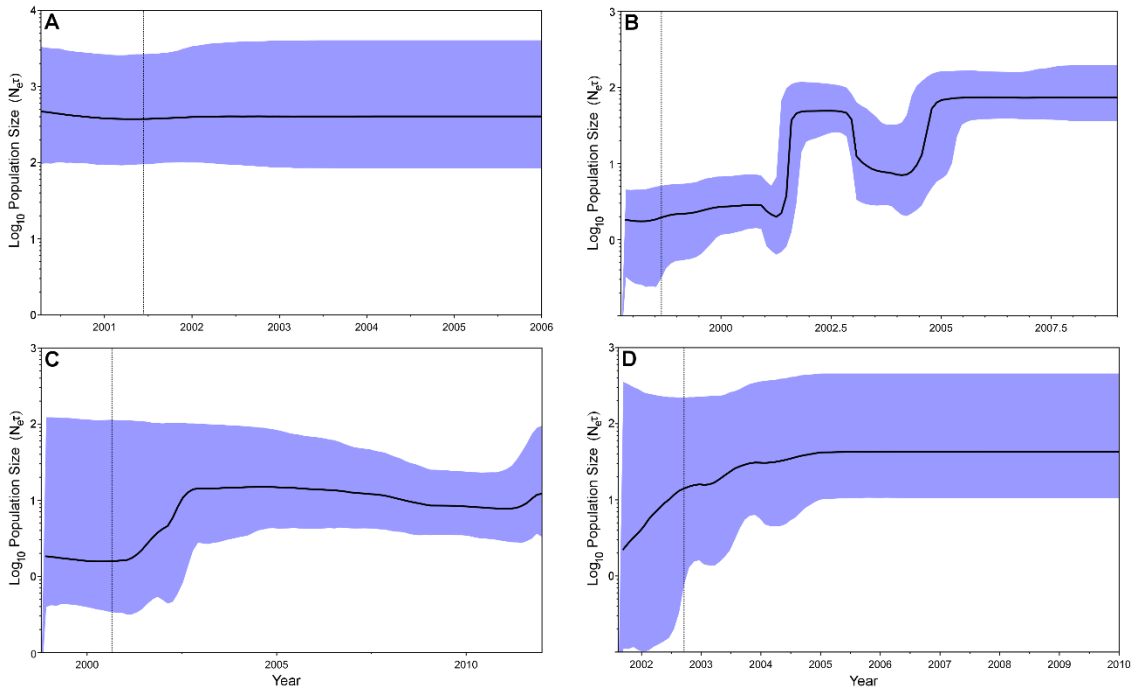

Figure 3.5. Bayesian Skyline plots for the (A) Midwest, (B) Northeast, (C) South, and (D) West regional models, aggregated at the county level. The y-axis (Neτ) is the effective population size multiplied by the generation length and the x-axis represents the year. The median measure is indicated by the thick black line, with the 95% HPD limits shown as the shaded blue area. The dotted vertical line represents the lower 95% HPD of the root height.

While the phylogenies for the CBS, state, and national county models show generally consistent results, this is not true of the predictors included in these three models. I show the posterior inclusion probabilities and corresponding regression coefficients for the CBS, state, and county aggregations in Figure 3.6. From Figure 3.6, the *Corvidae* counts and wetlands predictors fail to achieve a BF > 3 from either location of origin or location of destination in any of the three aggregations. For case counts and precipitation, the CBS and county models yield BF < 3 from both location of origin and

location of destination, while the state model achieves BF support for case counts at the location of destination and precipitation at both the location of origin and destination (BF = 13.1, 14.8, and 6.5, respectively). The distance predictor shows the most scale-dependent behavior, as support increases from the CBS to the state to the county levels (BF = 8.1, 102.4, and 30,185.0, respectively). Furthermore, the 95% HPD of the regression coefficient of the distance predictor decreases at each level of aggregation (95% HPD range = 7.21, 4.15, and 0.42 for the CBS, state, and county aggregations, respectively, in log-space). The entire HPD is negative for the county aggregation, which suggests that distance is preventing the diffusion of WNV in that model. The trend of decreasing posterior variance of the regression coefficient also holds true for population density at the location of origin (95% HPD range = 7.62, 6.53, and 2.89 for the CBS, state, and county aggregations, respectively, in log-space). Here, the entire HPD is positive for this predictor at the state and county aggregations, which suggests that population density at the location of origin is driving the diffusion of WNV in these two models. The CBS aggregation yields BF = 0.03 for this predictor, while the state aggregation yields BF = 227.6. For the county aggregation, this predictor was included in every sample after the 10% burn-in period, which corresponds to an inclusion probability of 1.0 and a Bayes factor that tends to infinity. This is also true for the county aggregation of the unvaccinated horses data at the location of origin, and the entire 95% HPD of the regression coefficient is positive, indicating that this predictor is also driving the diffusion of WNV for the county model. The CBS aggregation shows support for this predictor while the state aggregation does not (BF = 18.4 and 1.9, respectively). The state aggregation does show support for unvaccinated horses at the location of destination (BF

= 10.8) while the CBS and county aggregations do not (BF = 0.02 and 1.20, respectively). Finally, the CBS and county aggregations show similar support for temperature at the location of origin (BF = 21.5 and 25.1, respectively), while the state aggregation does not (BF = 0.4).

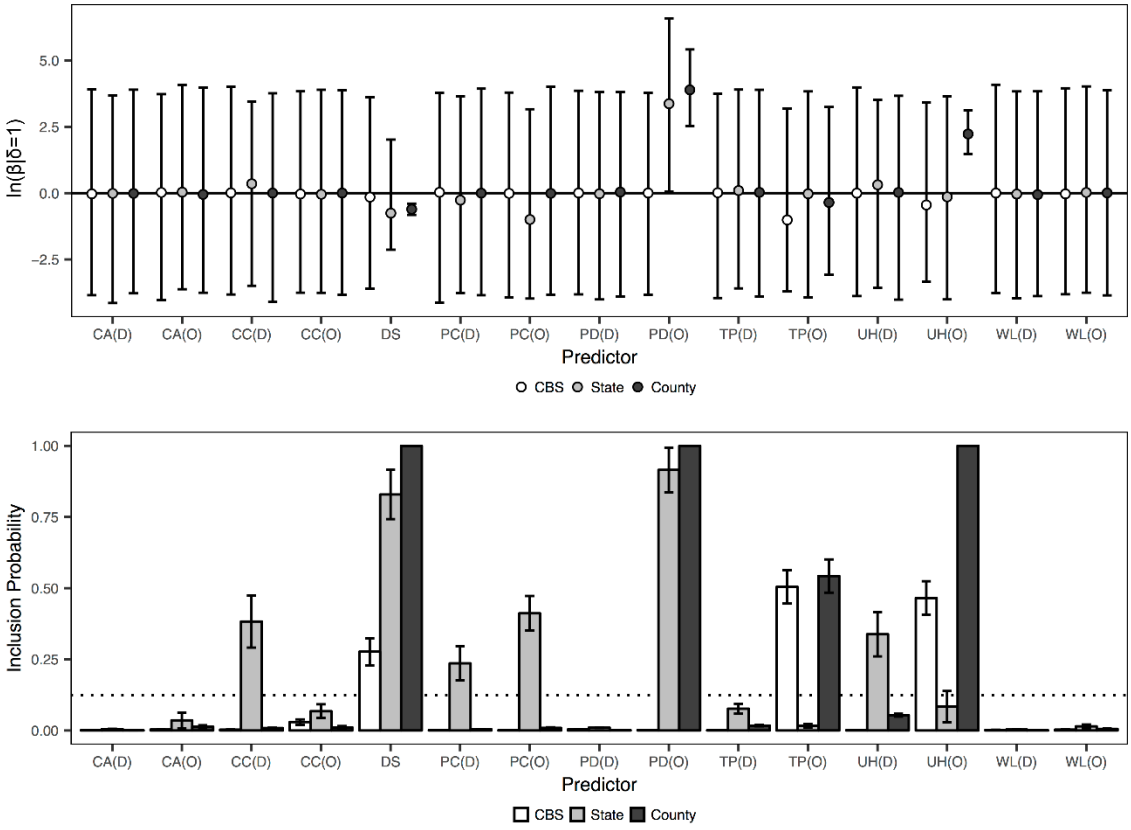

Figure 3.6. Inclusion probabilities and corresponding regression coefficients for the 15 predictors for the CBS, state, and county aggregations. The dotted line corresponds to BF = 3. Error bars represent the standard error for each predictor's inclusion probability and the 95% HPD for each predictor's regression coefficient. Predictor abbreviations are: *Corvidae* counts (CA), case counts (CC), distance (DS), precipitation (PC), population

density (PD), temperature (TP), unvaccinated horses (UH), and wetlands (WL), evaluated both from location of origin (O) and location of destination (D).

I also show the posterior predictor data for the four regional models at the county level in Figure 3.7. Here, I again see that the *Corvidae* counts and wetlands area predictors fail to achieve BF support from either location of origin or location of destination in any of the four regional models. Of the 15 predictors included, only case counts, precipitation, population density, temperature, and unvaccinated horses, each from the location of origin, showed BF > 3 in these models. Only case counts and population density were supported in more than one of the regional models. For the Northeast model, unvaccinated horses at the location of origin yields a positive 95% HPD, which suggests that this predictor was also driving viral propagation in this region. This is consistent with the national county aggregation (Figure 3.6). The BF for this predictor tends to infinity in the Northeast model, and this model also shows support for population density at the location of origin (BF = 39.1). In the Midwest model, case counts and population density are supported (BF = 4.4 and 108.3, respectively). In the South model, case counts and population density at the location of origin are supported (BF = 17.8 and 3.8, respectively). The West model only shows support for temperature and precipitation at the location of origin (BF = 12.4 and 9.0, respectively).

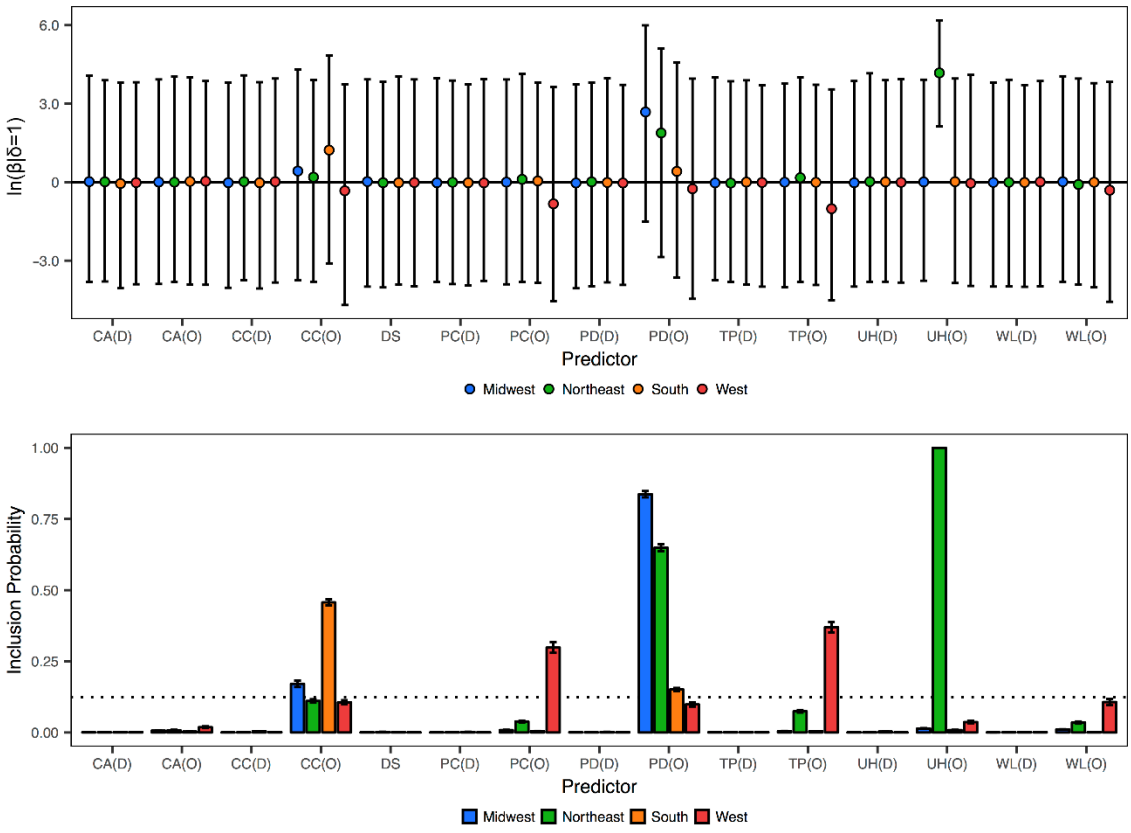

Figure 3.7. Inclusion probabilities and corresponding regression coefficients for the 15 predictors for the regional county-level aggregations. The dotted line corresponds to BF = 3. Error bars represent the standard error for each predictor's inclusion probability and the 95% HPD for each predictor's regression coefficient. Predictor abbreviations are: *Corvidae* counts (CA), case counts (CC), distance (DS), precipitation (PC), population density (PD), temperature (TP), unvaccinated horses (UH), and wetlands (WL), evaluated both from location of origin (O) and location of destination (D).

While Figure 3.6 outlines the variance in predictor support given the level of spatial aggregation and Figure 3.7 shows that local predictor trends are not necessarily consistent with those observed on a national basis, it is also pertinent to analyze possible sources of variance in the posterior estimates. In Figure 3.8, I plot the variance of the

inclusion probabilities and corresponding regression coefficients against the variance of the predictor point estimates for each individual model. That is, I show the posterior estimates as a function of the known variance in predictor point estimates. From Figure 3.8, the 95% confidence intervals fail to encapsulate many of the data points for any of the three statistics. The low R^2 values indicate that the variance in posterior estimates are not linearly correlated with the variance in predictor point estimates.

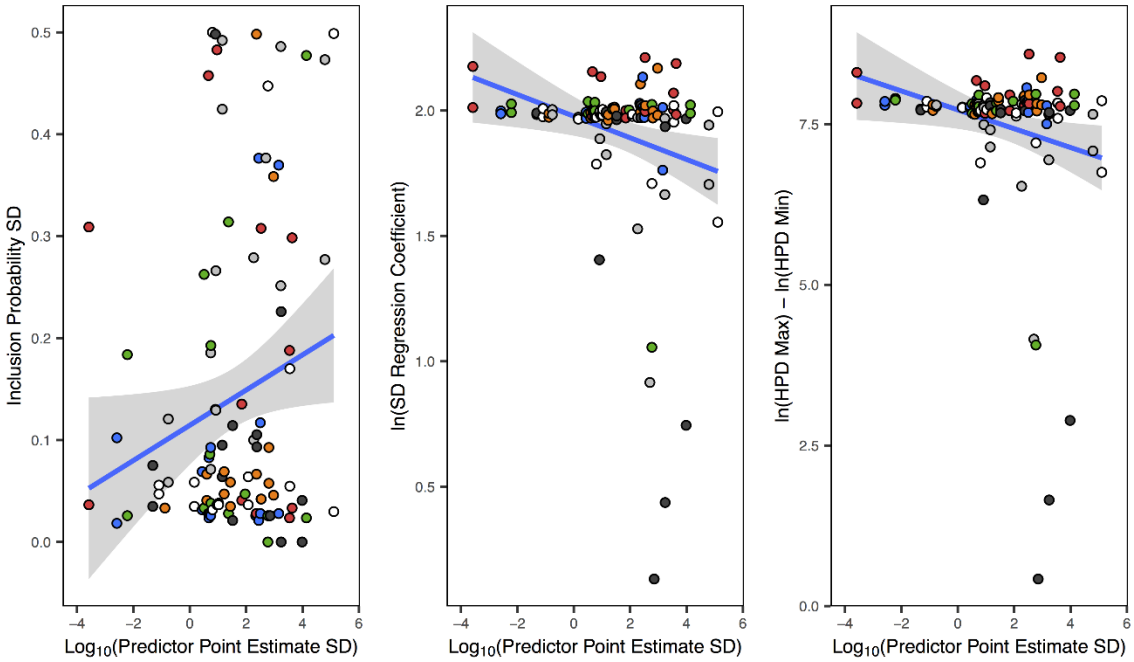

Figure 3.8. Linear correlations between the variance of predictor point estimates and the variance in posterior support metrics. The blue lines represent the lines of best fit and the shaded areas represent the 95% confidence intervals, and include data for all national and regional models.

I list the R^2 value for linear models between the predictor point estimate accuracy (independent variable) and posterior statistic variance (dependent variable) for each

individual analysis in Table 3. From Table 3, 20% of the posterior variance of the regression coefficient is explained by the predictor point estimate variance in the CBS-level aggregation, although just 8% of the posterior variances of the inclusion probability and HPD range of the regression coefficient are explained by the predictor point estimate variance. The state and national county aggregations do not yield $R^2 > 9\%$ in for any of the three statistics. For the regional county-level models, a modest amount of the variance in posterior estimates are explained by the predictor point estimate variances in the Midwest, Northeast, and South models. The Midwest analysis yields $R^2 > 17\%$ for each linear model, and 24% of the variance in the posterior regression coefficient is explained. Meanwhile, the West analysis shows $R^2 \leq 1\%$ for all three linear models.

Table 3.3

R^2 statistics for linear models between the variance of predictor point estimates and the variance in posterior support metrics

	Dependent Variable		
Model	SD P(δ^a=1)	SD (β^b)	HPD Range (β^b)
CBS	0.08	0.20	0.08
State	0.09	<0.01	0.09
County	0.07	0.07	0.07
Midwest	0.17	0.24	0.17
Northeast	0.14	0.01	0.14
South	0.10	0.18	0.10
West	0.01	<0.01	0.01
Overall	0.04	0.01	<0.01

[a] Inclusion probability
[b] Regression coefficient

Discussion

I found that the MCC topology and the age of its root were similar in the three national models that were successfully executed (CBS, state, and county-level

aggregations). The tMRCA for these viruses in each model is consistent with those presented in previous phylogeographic studies of WNV in the U.S. (Anez et al., 2013; Pybus et al., 2012). The observed molecular clock rates for each are slightly slower than the reported 5.06×10^{-4} in the open reading frame of human-origin isolates between 1999-2011 in the U.S. (Anez et al., 2013), but I note that our study includes one additional year and accounts for all sampled mosquito and avian species as well. In addition, the Bayesian Skyline plots are similar among the three national models (Figure 3.4). Given the fact that the best supported predictors vary across these three models, I conclude that the topology of the viral phylogenies is mainly determined by the sequence data rather than by the predictor data or discrete state partitioning.

I find the distance predictor to be of interest, especially pertaining to the three national models. Here, there is an increase in predictor support as the knowledge of the sampling location goes from most uncertain (CBS, BF = 8.1) to moderately uncertain (state, BF = 102.4) to least uncertain (county, BF = 30,185.0). Furthermore, the range of the HPD decreases as the sampling location certainty increases. The county-level aggregation suggests that distance is limiting the spread of WNV in the U.S., as its entire HPD is negative. The geographic diffusion of WNV in the U.S. is known to have occurred rapidly (Di Giallonardo et al., 2015). Thus, as the distribution of pairwise distances among discrete locations is largest at the county-level (Figure 3.10), it is plausible that the distance predictor would be protective. I note that the distance predictor is not supported in any of the four regional county-level models, which could indicate that geographic distance is less important at the local level but more important for widespread diffusion dynamics. Alternatively, this could simply be a result of the fewer

sequences and less genetic diversity available in the local analyses compared with the national analyses. The three national models share an alignment, with 79.2% identical sites, while the Midwest, Northeast, South, and West models contain 96.5%, 87.9%, 91.5%, and 95.8% identical sites, respectively. The strong negative correlation (Pearson's r = -0.99) between the number of viral sequences and the percent of identical sites per model demonstrates that an increase in the number of sequences results in a decrease in the number of fixed sites, and thus an increase in genetic diversity for the national models.

In addition to the distance predictor, human population density at the location of origin is not supported at the CBS level (BF < 0.1), well-supported at the state level (BF = 227.6), and was found to be included in every sample for the national county model (BF tends to infinity). This predictor is also supported in the Midwest, Northeast, and South regional county models (BF = 108.3, 39.1, and 3.8, respectively), but not for the West model (BF = 2.3). The point estimates of this predictor are the least uncertain at the county-level, so its unanimous support in the national model and frequent support in the regional county-level models provide evidence that population density is involved in WNV diffusion. As this predictor's contribution is strictly positive (regression coefficient = 3.9 and 95% HPD = [2.5, 5.4] in log-space), I can conclude that it is driving the diffusion of WNV from county-to-county, at least at the national level.

As the remaining predictors have variable support among the CBS, state, and county-level analyses, I reiterate several points about the point accurate estimations of two predictors which were outlined in Materials and Methods: human case counts and expected unvaccinated horses. For the human case counts, I collected the data at the state-

level (ArboNET, 2015) and not the specific county. Thus, our CBS and state aggregations have an accurate point estimate for this predictor, but for the county aggregation I assumed that the case counts per county were proportional to the county's population within the state. This assumption is not necessarily correct for the county-level aggregations. Case counts at the location of destination is supported in only the state model (BF = 13.1) so, the assumption for the county aggregations does not appear to have resulted in a potentially misleading supported predictor, although it is unknown how an alternative estimation assumption would change the posterior results. The expected unvaccinated horses predictor, however, does result in potentially suspect support metrics. The population of horses is known at the state-level (American Horse Council, 2005), and thus the population per CBS is simply the sum of the states in the region. For the county-level aggregation, I assumed that the number of horses was uniformly distributed across the state and thus that the number of horses per county was proportional to its land area. This predictor also required the vaccination rate per state, and several states in our dataset (Arizona, Connecticut, Ohio, Nebraska, South Dakota, and Texas) were absent from the survey from which we obtained this predictor (APHIS, 2006). I note that at least one state is absent from the Midwest, Northeast, South, and West regions of the USCB, which directly impacts the regional county-level analyses as well. I assumed that the absent states had the same vaccination rate as the most proximal geographic region from that survey, and that the CBS estimates were the average of the states in the region. I also assumed that the vaccination rate in each county was the same as the vaccination rate per state, which creates additional uncertainty. Overall, each of the seven completed models required a certain degree of assumption and potential error

introduction for this predictor, but the county-level aggregations required one additional assumption, thus increasing its uncertainty. I found that the expected unvaccinated horses at the region of origin is facilitating the diffusion of WNV in both the national and Northeast county-level aggregations (Figures 3.6 and 3.7) (BF tends to infinity and 95% HPD of the regression coefficient is strictly positive for each) and is supported at the CBS-level (BF = 18.4) as well (Figure 3.6), although its directionality is uncertain. Because the state-level point estimate is likely the most accurate for the expected unvaccinated horses predictor, and as I found that this predictor is supported from the location of destination in that model (BF = 10.8) I question the findings of the CBS and county-level aggregations. It is likely that multicollinearity is influencing the support of this predictor at the county level. For the CBS, state, and national county-level aggregations, the correlation between our point estimate for expected unvaccinated horses and the size of the discrete state is 0.23, 0.66, and 0.86, respectively. For the Midwest, Northeast, South, and West regional county-level models, the correlations are 0.62, 0.73, 0.67, 0.89, respectively. These data indicate that any support for the unvaccinated horses predictor in any of the county-level aggregations is rather indicative of the size of the discrete states, not the horse population, and should further caution researchers when aggregating predictors where assumptions must be made. In addition, horses infected with WNV are not known to be capable of passing the virus back to uninfected mosquitos, nor can they infect other horses or humans (Komar, 2000; Practitioners; Williams & Crans, 2004), so the suggestion that unvaccinated horses contributing to the spread of WNV seems suspect.

Although the correlations between the posterior predictor variance and variance in predictor point estimates (Figure 3.8, Table 3.3) fail to show linear trends, I do not consider this finding problematic. As the phylogenies are informed via both predictor data and sequence data (Lemey et al., 2014), identifying strong correlations would perhaps demonstrate a systematic bias within the GLM framework. Instead, these data may show the inherent stochasticity of this framework. Our inability to execute the CBR model due to its strong correlations between predictors (Table 3.1) indicates that discretizing locations and aggregating predictor data at highly uncertain levels for phylogeographic GLMs may require the elimination of predictors from the model and/or selection of alternate predictors such that the correlations are reduced. Either could result in a loss of pertinent information or misleading results regarding the dynamics of the virus in question.

Researchers that employ a GLM in Bayesian phylogeography may be tempted to create inferences based on posterior support of predictors and subsequently provide biological justification of these findings, but I believe that the results tell a cautionary story of the need to consider alternative discrete state construction. Here, I have shown how assigning identical nucleotide sequences into different discrete state sets with different degrees of spatial resolution can influence posterior support for predictors. As I am unable to pinpoint the source of the posterior variance as it pertains to the variance in point estimates across the discrete locations, I refrain from making any firm statements regarding which predictors are involved in the diffusion of WNV. Furthermore, as posterior predictor estimates obtained from regional county-level models are often inconsistent with those from the national level, it may also be important to perform

additional analyses at the local level prior to stating conclusions regarding more widespread epidemics (and vice versa). Finally, it is often the case that sampling locations are only known or annotated to low-level, uninformative, and ambiguous locations (Scotch et al., 2011; Tahsin et al., 2016). As I have shown here for the distance, population density, and expected unvaccinated horses predictors, aggregations that are averaged over a wider geographic area may not fully encapsulate or represent the true data at the precise sampling location, even though these same predictors at the county-level received strong support. Simply put, knowing the precise sampling location may enable local dynamics in viral diffusion to be revealed via GLM analyses, whereas this information may be lost when sequences are aggregated into coarser geographical units. Thus, I urge researchers that annotate and submit nucleotide sequences to public repositories to use the most precise sampling location possible so that these data can be used to accurately determine the factors that drive the diffusion of deadly viruses.

Materials and Methods

Model Parameters

Nucleotide Sequences. I obtained whole genome WNV sequences from the Virus Pathogen Database and Analysis Resource (Pickett et al., 2012) using the following search criteria: Family = *Flaviviridae*, Genus = *Flavivirus*, Species = *West Nile Virus*, Collection Year = *1999-2012*, Geography = *USA*, Host = *All*. This resulted in 781 sequences, 299 of which were annotated with a state of origin and county of origin. I aligned these 299 sequences using MAFFT v7.017 in Geneious Pro v.6.1.8 (Biomatters Ltd., Auckland, New Zealand). After exploratory Bayesian phylogeographic GLMs with

our sequence set failed to replicate molecular clock rates observed for WNV in the U.S. over a similar time period, I elected to focus on the envelope (E) protein of the WNV genome. I extracted the E protein for each record and aligned the sequences with the same parameters described above. I also performed four additional alignments, one for the sequences collected from each of the four CBRs: Midwest, Northeast, South, and West. I classified the hosts of these viruses using four categories: mosquito (n = 138), Corvidae (108), human (44), and other avian (9). The mosquito group contains members of the *Aedes* (11), *Culex* (101), *Culiseta* (15), *Ochlertotatus* (9), *and Psorophora* (2) genera. *Corvidae* is a family of birds, of which the American crow (*Corvus brachyrhynchos*, 81), blue jay (*Cyanocitta cristata*, 26), and black-billed magpie (*Pica hudsonia*, 1) were identified as hosts in these data. The remaining avian hosts include *Falco sparverius* (1), *Poecile atricapillus* (1), *Quiscalus quiscula* (1), *Accipiter cooperii* (1), *Buteo jamaicensis* (1), *Zenaida macroura* (1), *Mimus polyglottos* (2), and Loriidae (1). I provide the GenBank accession, discrete state assignment for each level of aggregation, and year of isolation of each sequence in Appendix C.

Bayesian Phylogeographic GLM. The phylogeographic model assumes that the location of each ancestral lineage is governed by a continuous-time Markov chain (CTMC) process that runs along the branches of an unknown phylogeny that is informed through sequence data. The infinitesimal rate matrix of the among-location CTMC process is parameterized as a log-linear GLM of predictors of interest (Lemey et al., 2014) to determine the probability of inclusion and the contribution of these predictor variables. Here, I selected predictors of interest to parameterize this rate matrix and estimate posterior probabilities of all 2^P linear models via a Bayesian stochastic search

variable selection (BSSVS) procedure (Lemey et al., 2014), where *P* is the number of predictors. I specified a 50% prior probability that no predictor is included in the model and evaluated the support of each predictor via Bayes factors (BFs), where I consider any predictor with BF > 3.0 to be supported for inclusion in the model (Kass & Raftery, 1995) following similar studies (Graf et al., 2015; Lemey et al., 2014; Magee et al., 2015; Magee, Suchard, & Scotch, 2017).

Levels of Predictor Aggregation. As I wish to investigate the differences in support for the GLM predictors when the sampling locations are specified with more or less resolution, I used four levels of aggregation for each predictor: USCB regions (CBR, $K = 4$), USCB subdivisions (CBS, $K = 8$), state ($K = 16$), and county ($K = 80$). At each level, I assume that the sampling location of each virus is known to one of the *K* discrete states. I define the aggregation levels as ranging from "most uncertain" (CBR) to "least uncertain" (county) as knowledge of the sampling location increases. I obtained internal latitude and longitude coordinates for each state in the contiguous U.S., including the District of Columbia, as well as every known sampled county from the USCB. For the CBR and CBS aggregations, the geospatial reference is the mean latitude and longitude of the states in the respective boundaries. To investigate whether regional dynamics of WNV match those at the national level for the four aforementioned aggregations, I also include a county-level aggregation for the sequences collected in each of the four CBR discrete states. That is, I selected the sequences from the four USCB regions, Midwest (n = 29), Northeast (n = 170), South (n = 64), and West (n = 36), and completed a regional analysis for each at the county-level aggregation. In Figure 3.9, I provide a map of the

discrete states in each level of aggregation and I detail the metadata for each sequence in Appendix C.

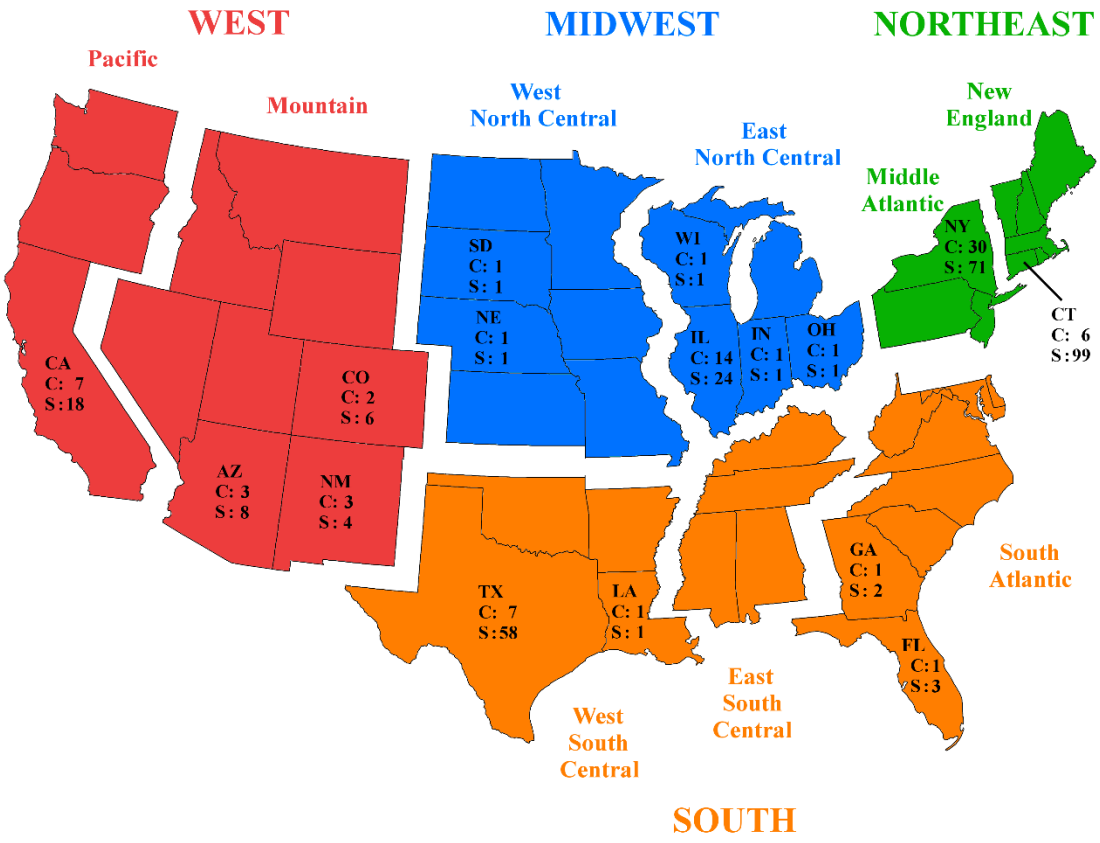

Figure 3.9. Map of the discrete state partitions used in this study. The four colors represent the discrete locations for the CBR model, which are further discretized into nine CBS locations, 16 states, and 80 counties. No samples were available for the East South Central subdivision at the CBS level. Each state is annotated with its number of unique sampled counties (C) and number of sequences (S). The Midwest, Northeast, South, and West regional models are county-level aggregations encapsulated by the *K* counties in their respective CBR.

GLM Predictors

I identified predictor data for each discrete state to represent the WNV epidemic in the U.S. from 1999-2012. Although the exact dates of the estimates for each predictor vary, each was accurate as of one specific point in time during the years of our study.

Distance (DS). I obtained a centroid latitude and longitude of each state and county from the USCB MAF/TIGER database. I calculated the pairwise distance from each location to the next using these coordinates for the state and county aggregations, respectively. For the CBR and CBS aggregations, I calculated the mean internal latitude and longitude for each of the states in the defined area. I used these means as the centroid coordinates for each area and calculated pairwise distances between them.

Population Density (PD). I obtained human census data from the USCB for the most recent full census in 2010. I obtained population data at both the state and county levels. For the CBR and CBS, I summed the total population in the respective states and divided by their total land area to obtain a density. For the state and county aggregations, I divided the population per sampled location by its land area to obtain a density.

Case Counts (CC). I obtained data on the number of cases per state from the Centers for Disease Control and Prevention's (CDC) ArboNET surveillance program (ArboNET, 2015). These data reflect the cumulative number of human cases per year from 1999-2012 at the state level. The predictor for the CBR and CBS aggregations reflect the total number of human cases in the defined states during this period. For the county aggregation, the point estimate assumes that the county observedseveral cases proportional to its population within the state. These data round to zero for Concho

County (Texas) and Wilkinson County (Georgia), so they were changed to a value of one to ensure positivity for the log transformation.

Unvaccinated Horses (UH). The American Horse Council Foundation completed the most comprehensive horse census in the U.S. during the time period of our study. This survey counted the number of horses per state, and includes horses found on farms, private homes, and those in the racing, showing, and recreation industries as of 2005 (Council, 2005). As data was only available for the state level, I assumed that these animals were uniformly distributed across the entire state. Thus, the state aggregation encapsulates all horses estimated to be in the state, and the county level reflects the expected number of horses given the county's land area. I used the sum of horses in each state for the CBR and CBS aggregations. Furthermore, the Animal and Plant Health Inspection Service (APHIS) of the U.S. Department of Agriculture (USDA) provided estimates of equid vaccination practices during the 2005 calendar year (APHIS, 2006). This study surveyed equid owners in 28 states, 12 of which were included in this study. The survey provided average vaccination rates of resident equids, discretized into four regions: South, Northeast, West, and Central. For the CBR and CBS aggregations, I matched each region to its most appropriate of the four USDA regions. At the state level, I used the average rates per region in the USDA study for the 12 states in this study and assumed the remaining four states to be part of the most proximal geographic region. At the county level, I use the same vaccination rate as its corresponding state. Given the horse census and the vaccination rates, I use the expected number of unvaccinated horses in each discrete state as the predictor at each level of aggregation.

***Corvidae* Counts (CA).** Of the 299 sequences used in this study, I identified 118 that were isolated from avian hosts, including 11 unique species, however 104 of these sequences (92%) were from birds of the *Corvidae* family. The Cornell Lab of Ornithology (Sullivan et al., 2009) provides a collection of census data obtained by birders throughout the world and can be selected by species, geographic region, and time. I obtained the total number of observations of the three *Corvidae* hosts (*Corvus brachyrhynchos*, *Cyanocitta cristata*, and *Pica hudsonia*) during 1999-2012 for each respective discrete state at the CBR, CBS, state, and county levels, as well as the total number of reports. To account for potential biases in the reporting of birders in various locations, I divided the cumulative counts of the three species by the cumulative number of reports to obtain an expected number of *Corvidae* sightings per observation in each discrete state and used this as the predictor at each level.

Wetlands Area (WL). I obtained GIS shapefiles of each of the states in this analysis from the U.S. Fish and Wildlife Service (USFWS, 2016). These files contain the wetlands polygon data for each state, and I extracted the total area of wetlands for the state level using ArcMap v10 (ESRI, Inc., Redlands, CA, USA). For the CBR and CBS aggregations, I used the sum of each state's wetlands to obtain the total wetlands per defined area. For the county level, I obtained the map of counties in each state from the USCB MAF/TIGER database and extracted all instances of wetlands contained in the respective counties. I divided each wetlands area by the total land area of each discrete state to obtain percentage wetlands cover, and used this as the predictor point estimate at each level of aggregation.

Temperature (TP) and Precipitation (PC). I obtained temperature and precipitation data from the 30-year normal datasets (1981-2010) provided by the National Climatic Data Center of the National Oceanic and Atmospheric Administration (NOAA) (Arguez et al., 2012). At the CBR, CBS, state, and county levels, I extracted the average annual temperature and precipitation data from each NOAA station in the respective areas. The temperature and precipitation predictors thus reflect the average 30-year normal observed by all stations in each discrete state. At the county level, there were several instances where either no NOAA station existed within the county boundaries or the station(s) in the county did not contain normal temperature or precipitation data. In these instances, I used the most proximal station within that state to the county's centroid coordinates that contained both temperature and precipitation normal.

I log-transformed and standardized all predictor data and created a separate predictor from both discrete state of origin and discrete state of destination, with the exception of the distance predictor, for a total of 15 predictors. I summarize the distributions of the predictors at level of aggregation in Figure 3.10.

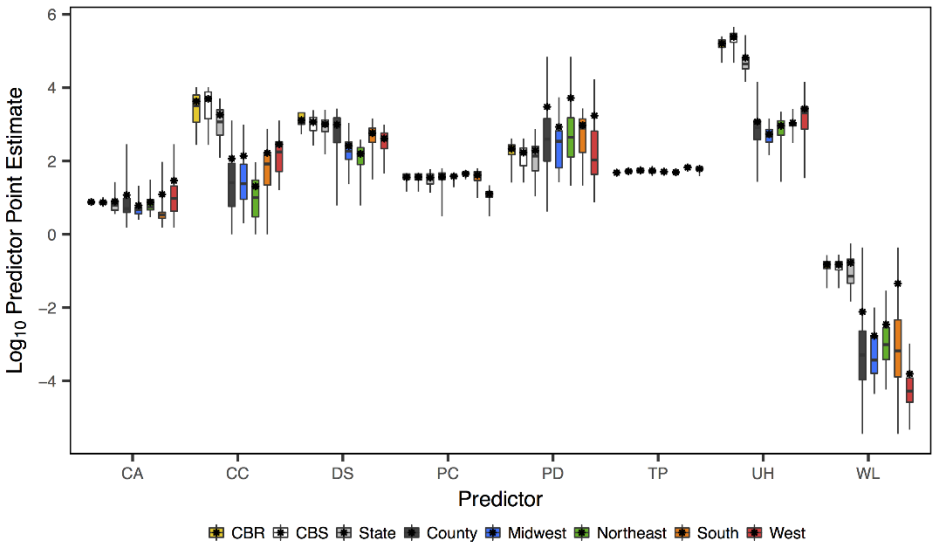

Figure 3.10. Boxplots of the predictors used in this study for each model. Predictor abbreviations are: *Corvidae* counts (CA), case counts (CC), distance (DS), precipitation (PC), population density (PD), temperature (TP), expected unvaccinated horses (UH), and wetlands area (WL).

BEAST Analyses

I specified a generalized time reversible substitution model following previous WNV studies (Anez et al., 2013; Di Giallonardo et al., 2015; Duggal et al., 2014; Lopez, Soto, & Gallego-Gomez, 2015; Mann et al., 2013; Pybus et al., 2012), also including invariant sites and a gamma heterogeneity (GTR+I+G) on our sequences. I set an uncorrelated lognormal relaxed molecular clock (Drummond, Ho, Phillips, & Rambaut, 2006) following previous studies (Ciccozzi et al., 2013; Mann et al., 2013; Pybus et al., 2012) with 0.001 substitutions per site per year and specified a Bayesian Skyline prior (Drummond et al., 2005). For each discrete space partitioning, I specified a phylogeographic GLM (Lemey et al., 2014) using the respective predictor data at each level of aggregation. I evaluated each using the BEAST v1.8.4 software package

(Drummond et al., 2012) with a chain length of 250 M and sampling every 25,000 steps for the CBR, CBS, state, and national county-level models. For the four regional models, we specified a chain length of 150 M with sampling every 15,000 steps. I used TreeAnnotator v1.8.4 to construct a maximum clade credibility (MCC) tree for each model after discarding the first 10% of trees as burnin and annotated the trees using FigTree v1.4.2. I obtained the mean posterior probability of inclusion, BF support, and the contribution of each GLM predictor for each model using Tracer v1.6.

Predictor Variance Correlations. From each model and for each predictor, I extracted the standard deviation of the inclusion probability, the standard deviation of the regression coefficient, and the upper and lower bounds of the regression coefficient's HPD. I used the "geom_smooth" function in the "ggplot" package in R v3.3.1 (R Core Development Team, 2008) to visualize correlations between the variance of predictor point estimates and variance in posterior support. I used the "lm" function to obtain these R^2 values for each individual model (Table 3.3) and with all models pooled together (Figure 3.8).

Data Availability. I have made all FASTA alignments, XML files, and MCC phylogenies freely available at https://figshare.com/projects/WNV_GLM_Aggregation_Study/19201.

CHAPTER 4

A PIPELINE FOR PRODUCTION OF BEAST XML FILES WITH GENERALIZED LINEAR MODEL SPECIFICATIONS

Introduction

Although Bayesian phylogeographic generalized linear models (GLMs) offer the benefit of simultaneously reconstructing the spatiotemporal history of the virus and assessing the contribution of each predictor to the process, few studies have utilized such an approach. As I addressed in Chapter 2, one possible hindrance to its widespread adoption could be a lack of research into the computational performance of GLMs compared to the traditional Bayesian stochastic search variable selection (BSSVS) framework (Lemey et al., 2009). Similarly, as I addressed in Chapter 3, researchers may struggle with discretizing locations and/or locating accurate predictor data for their selected geographic region. A different explanation could simply be that the GLM framework is not directly implementable via BEAUti, like other phylogeographic methods. Currently, the implementation of the GLM framework involves manual manipulation of XML files, as described by a tutorial (P. Lemey, Rambaut, & Suchard, 2014). In order to facilitate the process of performing a phylogeographic GLM, I introduce a Python script that was created to outfit BEAST-ready XML files (Drummond et al., 2012) with the GLM specification (P. Lemey et al., 2014) using a small number of command line options and preparation of applicable predictor data.

Program Requirements

The most recent version of this program can be found at https://github.com/djmagee5/BEAST_GLM. Here, one can access and download the code, a detailed README, and example files.

Python

The program is written in Python v3.4.3 (Python Software Foundation, 2015) and requires the built-in "xml", "os", and "math" packages. It also requires the external "numpy" package (van der Walt, Colbert, & Varoquaux, 2011), for which documentation and installation instructions are listed in the README.

BEAST XML File

As the purpose of the program is to outfit a BEAST-ready XML file with the GLM specification, a BEAST-ready XML file is needed. This file must specify a discrete trait (*e.g.* location or host) that will be modeled via a log-linear GLM of predictors of interest. It does not matter whether or not this discrete trait uses the BSSVS specification (Lemey et al., 2009).

Predictor Data File(s)

Point estimates for at least one predictor must be obtained for each state of the discrete trait that is to be modeled via the GLM. The predictor data must be in comma-delimited (.csv) or tab-delimited (.txt) format and can be presented either batch or single form, specifications of which are outlined below.

Batch Predictor File

A batch file of predictor data lists point estimates of multiple predictors for each discrete state. Users will be able to indicate whether a predictor should be taken from the

discrete trait of origin, discrete trait of destination, or both. An example of a batch predictor file is shown in Table 4.1. Batch predictor files must meet the following requirements:

1. The first value in the first line should be the name of the discrete trait that the user wishes to model as a GLM.
2. The remaining values in the first line must be the names of the predictors.
3. The first value in all remaining lines must be the names of the discrete states in the XML file. The order of the states does not matter as the program will sort them according to the order specified in the XML file. They should exactly match the names of the discrete states in the XML file to avoid any errors, although the program will strip whitespace and is case insensitive in order to avoid such issues.
4. The remaining values in each line must be the values of the predictor in the column for the line's discrete state.
5. A predictor will be created for each predictor name in the first row.

Table 4.1

Example format of a batch predictor file

Location	Population_Density	Temperature	Precipitation
Arizona	56.27	62.11	13.46
California	239.14	59.52	24.12
Colorado	48.07	46.00	16.63
Connecticut	738.08	49.50	50.38

Single Predictor File

A single file of predictor data lists the point estimates of one predictor in matrix form. Point estimates are directional, from the discrete state in row *i* to the discrete state in

column *j* for all $i \neq j$ (Lemey et al., 2014). Multiple input files can be placed in a single directory, <singlePredictorDir>, and a predictor will be created for each file in the specified directory. An example of a single predictor file is shown in Table 4.2. Single predictor files must meet the following requirements:

1. The name of the predictor should be the first value in the file (*i.e.* first row, first column).
2. The remaining values in the first line must be the names of the trait's discrete states. They should exactly match the names of the discrete states in the XML file to avoid any errors, although the program will strip whitespace and is case insensitive in order to avoid such issues.
3. The first value in each of the remaining lines must be the name of one of the discrete states. The same rules apply from Step 2 regarding discrete state names.
4. The remaining values in each line must represent the value in the matrix corresponding to the transition from the <discrete state in the row> to the <discrete state in the column>.
5. Values in the diagonal entries should be 0.
6. A predictor will be created for each single predictor file in <singlePredictorDir>.

Table 4.2

Example format of a single predictor file

Temperature_Origin	Arizona	California	Colorado	Connecticut
Arizona	0	62.11	62.11	62.11
California	59.52	0	59.52	59.52
Colorado	46.00	46.00	0	46.00
Connecticut	49.50	49.50	49.50	0

Predictor Data Point Estimates

For both the batch and single predictor files, all point estimates must be positive as these data will be log-transformed by the program, with two exceptions: diagonal entries in the single predictor files and coordinates in batch predictor files. For the former, diagonal entries in single predictor file matrices are ignored by the program, so the '0' entries (specified *Single Predictor File – Step 5*) are simply placeholders to ensure that the matrix is square. For the latter, if a batch predictor file has two columns labeled like coordinates (*e.g.* "Latitude" and "Longitude" or "LAT" and "LONG", case insensitive), the program will prompt the user to determine if a "distance" predictor is desired for the discrete trait (*i.e.* location), as it has been used as a predictor in multiple phylogeographic GLM studies (Lemey et al., 2014; Magee et al., 2015). If the user elects to use distance, the program will calculate the great circle distance between the coordinates for each pair of discrete states. The user will have the option to retain the raw coordinates as individual predictors if they so desire. Aside from these two exceptions, any predictor that contains non-positive point estimates will be flagged by the program and the user will be informed that said predictor(s) cannot be used in their log-linear GLM. This applies to coordinate predictors in their raw form. That is, if a user elects to include distance and also wishes to include raw latitude as a predictor, if some locations have a negative latitude (*i.e.* are located in the southern hemisphere) the program will indicate that raw latitude may not be used as a predictor. A user could, however, provide a workaround for this by including a separate "relative latitude" predictor that indicates a location's position relative to a certain point, which may ensure that all data are positive and thus can be used as a predictor. Finally, it is important to note that all predictor data

will be standardized by BEAST (Drummond et al., 2012). This essentially nullifies all units of predictor data point estimates, which enables flexibility of users that have inherently non-positive predictor data. For example, if one discrete state's point estimate for a "temperature" predictor is –0.5˚C, the program will not allow this predictor to be included. The user could, however, simply transform all point estimates from Celsius to Fahrenheit, which would yield 31.1˚F for this discrete state. The mean and variance of this predictor will not have changed after the transformation, but the program will now allow for its inclusion and the standardized predictor data will be identical, so it will not affect posterior estimates in any way.

Program Execution

The program is built for command line execution with a minimum of four and maximum of six arguments, which are outlined in Table 4.3. A general use case for the program is as follows:

```
$ python create_glm_xml.py <xmlFile> <discreteTrait> single
  <singlePredictorDir> batch <batchFile>
```

Table 4.3

Arguments for the Python script

Argument	Notes
<xmlFile>	BEAST-ready XML file that specifies some discrete trait.
<discreteTrait>	Name of the discrete trait to be modeled as a log-linear GLM. Case insensitive.
single	Indicates that single predictor file(s) will be used. Case insensitive.
<singlePredictorDir>	Directory containing all single predictor files to be written to the new XML.
batch	Indicates that a batch predictor file will be used. Case insensitive.
<batchFile>	Path to the batch predictor file to be written to the new XML.

For all use cases, the first two arguments must be <xmlFile> and <discreteTrait>, respectively, to indicate the XML file to process and the discrete trait to transform into a log-linear GLM with the desired predictor data. At least one of "single" or "batch" must also be specified, followed by the directory containing all single predictor files or the batch predictor file, respectively. The user may elect to use both single and batch files. In this case, the order for the arguments "single <singlePredictorDir>" and "batch <batchFile>" does not matter. An insufficient or excess number of arguments will prompt an error message and a new XML file will not be created.

Additional User Prompts

If a user only specifies single predictor file(s), the program will not require any additional user input. If a user specifies a batch predictor file that includes coordinate-like predictors, the user will be asked if a "Distance" predictor should be included as previously detailed. After the user indicates whether distance should be included and, subsequently, whether the raw latitude and longitude coordinates should be retained as

additional predictors, a list of predictors from the batch file will be echoed to the screen. The latter steps will be skipped and the list of predictors from the batch file will be immediately echoed to the screen if no coordinate-like predictors are contained in the batch predictor file. Table 4.4 shows how the example batch predictor file from Table 4.1 will be displayed by the program.

Table 4.4

Example output visible to a user that inputs a batch predictor file

Num	Predictor	Direction
(0)	Population_Density	Both
(1)	Temperature	Both
(2)	Precipitation	Both

From Table 4.4, the "Direction" column may hold one of four values: "both" (default), "origin", "destination", or "** REMOVE **". This column represents the direction(s) from which the predictors are to be represented (*i.e.* from discrete state of origin, discrete state of destination, or both). A prompt will ask a user if the list is correct. If the list is not correct, the user may remove any predictor or modify its directionality. A modified list will be echoed to the screen with each change made by the user, and this process will continue in a loop until the user indicates that they are satisfied with the final list. Once the list is finalized, no more input is required from the user.

Algorithm

Once the user calls the program, the following steps occur:

1. Ensure that a correct number of command line arguments are entered.

2. Check the specified XML file and ensure that it contains the specified discrete trait.
3. Extract all discrete states for that discrete trait.
4. Read in predictor data from batch and/or single predictor file(s).
5. Extract the list of discrete states from each predictor file and ensure that the list matches the discrete states from the XML file.
6. Read in all predictor data and check for non-positive values. Exceptions are detailed in *Program Requirements – Predictor Data Point Estimates*. Log-transform all data.
    a. Echo to the screen any data points that are non-positive, including the row and column numbers in the specified file(s).
7. If a batch predictor file is uploaded, complete the *Additional User Prompts* until the user is satisfied with the final predictor list.
8. Process the original XML file line-by-line and write its fields to a new XML file. The name of the new XML file will indicate that a GLM is specified (*e.g.* "originalXMLFileName.xml" to "originalXMLFileName_GLMedits.xml").
    a. Comment out all sections of the XML file that must be removed in order to model the discrete trait with the log-linear GLM specification and replace them with the required GLM sections as outlined by the BEAST tutorial (P. Lemey et al., 2014).
    b. Write the log-transformed predictor data in the correct order for the predictor design matrix and calculate its rank. Output a file containing the

predictors in the order that they were written to the new XML file, for the user's reference, titled "originalXMLFileName_predictorNames.txt".

c. Change the names of all logfiles to indicate that the data stem from a GLM (*e.g.* "originalLogFileName.log" to "originalLogFileName_GLMedits_discreteTrait.log".

9. Echo to the screen the number of predictors and the rank of the design matrix. As the new XML file will not execute in BEAST if the design matrix is not of full rank, echo to the screen a statement regarding whether or not the new XML file is likely to run based on the design matrix's rank.
10. Echo to the screen a message that the program has terminated, the name of the new XML file with the GLM specification, and the name of the discrete trait for which the GLM will be modeled.

The program's physical output is a new, renamed XML file and a plain text (.txt) file that lists the predictors in the order that they were written to the XML file. The former is renamed, as are all logfiles contained in the original XML file, such that both files can be executed in BEAST without fear of inadvertent overwriting of some or all crucial data. The latter is done in order to provide the user with a reference to the predictor logfile that will be outputted by BEAST. This logfile will contain column titles like as "coefIndicator1" and "glmCoefficient1" to represent the indicator variable and regression coefficient for the first predictor written in the XML file. The list of predictors, titled "originalXMLFileName_predictorNames.txt", informs the user which predictors correspond to which variables.

Error Messages

The program anticipates several possible errors that will either render the program unable to create a new XML file or will result in a new XML file that will cause a known BEAST error. Identification of these errors will result in a termination of the program and a new XML file will not be created, but will result in an informative error message that will be echoed to the screen in the event that one is encountered by a user. Some of the possible errors include:

1. Incorrect number of command line arguments.
2. Failure to specify "single" or "batch" as the third and/or fifth command line arguments.
3. The XML file does not contain the specified discrete trait.
4. Different number of discrete states in the XML file and predictor data file.
5. Discrete state name(s) listed in a predictor data file cannot be matched to a discrete state name listed in the XML file.
6. Invalid values (*e.g.* non-floating point or negative) provided in a predictor data file.
7. Single predictor file is not a square matrix.

Conclusion

I developed this program to facilitate the currently-tedious nature of creating of GLM-outfitted BEAST XML files. This function is not currently supported in BEAUti (Drummond et al., 2012). It will ideally enable researchers to seamlessly produce these files for investigating the contribution of predictors to the overall diffusion process of the

virus 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 my knowledge, this program will create GLM-outfitted XML files that will properly execute in BEAST v1.8.3 and v1.8.4. I will work to promptly update the code to incorporate any changes that future versions of BEAST may require, including the rapidly-developing BEAST2 framework. I have posted a brief description of this code, as well as the GitHub address, to the "beast-users" Google Group in order to promote this time-saving program to its target audience.

CHAPTER 5

DISCUSSION

Summary of Chapters

The purpose of my dissertation was to study the use of generalized linear models (GLMs) in Bayesian phylogeography. In Chapter 1 (Magee et al., 2015), I provided a case study that showed how such a method could be used to explain the diffusion of an RNA virus. In this example, I studied influenza A/H5N1 in Egypt, which is an on-going public health concern. The results indicate that, in addition to strong support for sample size predictors, the overall density of bird species, as well as the density of specific avian hosts, may have been involved in the diffusion process of this virus in Egypt. As H5N1 is an avian virus and Egyptian citizens often obtain their poultry via live bird markets (Abdelwhab & Hafez, 2011), these results are biologically justifiable. Also supported for inclusion in the model were longitude of the location, the lack of a genetic motif corresponding to increased transmissibility of the virus, human population density, climate factors, and elevation. Each of these variables were suspected to have been involved with the circulation of an avian influenza virus. Although none of the predictors, aside from sample size, were found to suggest a driving force or protective effect of the diffusion of H5N1, the results do show how the GLM framework can be implemented to provide a direct biological interpretation of the spread of a virus. At the time, this was just the third publication that utilized a GLM in Bayesian phylogeography (Faria et al., 2013; Lemey et al., 2014), so I focused the remainder of my dissertation on properties of the GLM framework that had yet to be analyzed. My goal was to provide researchers

with an established foundation of this framework, including its limitations and other factors that researchers should consider when using it.

In Chapter 2 (Magee et al., 2017), I assessed the GLM framework's performance against the popular Bayesian stochastic search variable selection (BSSVS) framework and a primitive model that does not use BSSVS for the influenza A/H3N2 virus during the 2014-15 flu season in the United States. Across six scenarios for population growth, six random sequence samples, and five total methods entailed by the three frameworks of ancestral state reconstruction, the GLMs provided the most statistically favorable phylogeographic reconstructions. Furthermore, the GLMs showed strong support for temperature and precipitation at the location of origin as drivers of the virus, which provided a biological interpretation that was consistent with global source-sink dynamics of influenza viruses. These results appeared to show the GLM, arguably, as a better method for this particular virus and time period, but the GLM was also found to be the most influenced by sampling bias among the three frameworks. Meanwhile, the BSSVS framework showed significantly lower correlations between the posterior probability of each region at the root of the maximum clade credibility (MCC) phylogeny and the number of samples from that region than the GLMs under three of the six population growth scenarios. Chapter 2 showed that caution should be taken when using a GLM and interpreting its results, as they could be strongly impacted by sampling bias compared to alternative methods.

Still unknown, however, was how the partitioning of the geographic area into discrete states influences the identification of predictors when using a GLM. Namely, as the sampling location of virus sequences are typically annotated at a high level of spatial

order (*e.g.* a country or a state in the U.S.) it was unclear if aggregating predictor data at these levels would result in a loss of posterior information gain. Conversely, it was unknown if aggregating predictor data at a low level of spatial order (*e.g.* county) would reveal the predictors that are involved in the diffusion process to a better extent. Therefore, In Chapter 3, I addressed whether the way in which discrete states are selected, and, thus, how sequences are pooled, makes a difference in posterior support metrics for predictors. For this analysis, I selected West Nile virus in the U.S., as 299 sequences were annotated with the county of isolation. I pooled the sequences into four discrete U.S. Census Bureau regions, eight U.S. Census Bureau subdivisions, 16 states, and 80 counties. I then collected and aggregated predictor data at each of these four levels, then performed a GLM analysis for each. The results indicate that the level of aggregation clearly makes an impact in the support metrics for predictors. In fact, when the sequences were discretized by the four regions of the U.S. Census Bureau, the predictor point estimates became so correlated that the predictor design matrix could not achieve full rank and thus could not be executed in BEAST. For the U.S. Census Bureau subdivision, state, and county-level aggregations, the predictors that achieved BF > 3.0 varied between the analyses, although the MCC trees showed consistent times to the most recent common ancestor, molecular clock rates, root state posterior probabilities, and their Bayesian Skyline plots showed similar population sizes over time. Four additional analyses performed with a county-level aggregation that encapsulated the counties from each individual U.S. Census Bureau region showed that the support for predictors region-by-region did not necessarily reflect the national trends. These results demonstrate that caution should be taken by researchers when selecting a spatial partition, and that the

most specific discrete states possible should be used in order to truly identify the local variables that may impact the diffusion of a virus. Furthermore, a predictor that represented the expected number of unvaccinated horses at the county level showed strong support, although it is likely that this was a product of collinearity with the size of the county. Predictor data was not directly measurable at the county level and its point estimate was obtained by assuming that the horse population was proportional to the size of the county. This shows that researchers should also use caution with their assumptions and ensure that predictor support is not an artifact ofcollinearity with another, perhaps unrelated, measurement.

The GLM framework may be used to address complex epidemiological questions, and my studies in Chapters 1-3 demonstrated strengths and weaknesses of using this approach. Despite its potential, it has yet to gain the popularity that might be expected for such an innovative method. One possible reason for its lack of popularity could be the difficulty of implementing the framework in the BEAST software package (Drummond et al., 2012), as the software used to create BEAST XML files, BEAUti, currently does not support the GLM. Thus, BEAST XML files must be manually manipulated in order to use the GLM specification. Although there is a tutorial (P. Lemey et al., 2014), this process is, in my experience, extremely tedious and may be hindering the widespread adoption of GLMs by phylogeographic researchers. Therefore, in Chapter 4, I introduced a pipeline that may facilitate expanded use of the GLM framework by the general public. The pipeline that I created and have made public (https://github.com/djmagee5/BEAST_GLM), allows individuals to simply pass a BEAST XML file, the name of a discrete trait, and formatted predictor data to a Python

script which will then produce a new XML file outfitted with all necessary components to use the GLM specification in BEAST. I will update the program to patch any bugs discovered by users, and will also provide support for future versions of BEAST.

Future Research

There are several opportunities to capitalize on the research contained in this dissertation. First and foremost, although Chapter 2 provides a side-by-side comparison of how ancestral state reconstructions, including the GLM, compare to one another for a given sequence set, time frame, and region, it does not empirically test whether one is "better" at obtaining a correct phylogeny. A future study could provide BEAST with sequence data simulated using a known phylogeny and allow the three ancestral state reconstruction frameworks to attempt replicate this target phylogeny so that the accuracy of each method could be directly assessed. Predictors for the GLMs could be simulated based on factors that vary across space, such as transmission rates, in order to directly assess the phylogeny-trait-predictor relationship. A different study could also analyze the effects of using a GLM on multiple discrete partitions (*e.g.* host and location). There has yet to be a study that utilizes such an approach, so it would be scientifically interesting to observe whether predictors for one discrete trait dominate the resulting phylogeny, predictors from each discrete trait are involved, or the sequence data dominates the phylogeny. A simple approach would be to select a sequence set, time frame, and location where the discrete states of multiple traits are known. Under this example, a researcher could perform a phylogeographic assessment with: (i) no GLM, (ii) a GLM on the host trait, (iii) a GLM on the location trait, and (iv) a GLM on both the host and

location traits in a single analysis. Two such studies may build upon the work presented in my dissertation and further reveal the true capabilities and limitations of the GLM framework.

REFERENCES

Abdelwhab, E. M., & Hafez, H. M. (2011). An overview of the epidemic of highly pathogenic H5N1 avian influenza virus in Egypt: epidemiology and control challenges. *Epidemiology & Infection, 139*(05), 647-657. doi:doi:10.1017/S0950268810003122

Abdelwhab, E. M., Selim, A. A., Arafa, A., Galal, S., Kilany, W. H., Hassan, M. K., . . . Hafez, M. H. (2010). Circulation of Avian Influenza H5N1 in Live Bird Markets in Egypt. *Avian Diseases, 54*(2), 911-914. doi:10.1637/9099-100809-RESNOTE.1

Ahmed, S. S. U., Ersbøll, A. K., Biswas, P. K., Christensen, J. P., Hannan, A. S. M. A., & Toft, N. (2012). Ecological Determinants of Highly Pathogenic Avian Influenza (H5N1) Outbreaks in Bangladesh. *PLoS One, 7*(3), e33938. doi:10.1371/journal.pone.0033938

American Horse Council. (2005). *Most Comprehensive Horse Study Ever Reveals A Nearly $40 Billion Impact on the U.S. Economy*. Retrieved from

Anez, G., Grinev, A., Chancey, C., Ball, C., Akolkar, N., Land, K. J., . . . Rios, M. (2013). Evolutionary dynamics of West Nile virus in the United States, 1999-2011: phylogeny, selection pressure and evolutionary time-scale analysis. *PLoS Negl Trop Dis, 7*(5), e2245. doi:10.1371/journal.pntd.0002245

APHIS. (2006, Dec 2006). Vaccination Practices on U.S. Equine Operations. Retrieved from https://www.aphis.usda.gov/animal_health/nahms/equine/downloads/equine05/Equine05_is_Vaccination.pdf

Arafa, A., El-Masry, I., Kholosy, S., Hassan, M. K., Dauphin, G., Lubroth, J., & Makonnen, Y. J. (2016). Phylodynamics of avian influenza clade 2.2.1 H5N1 viruses in Egypt. *Virol J, 13*, 49. doi:10.1186/s12985-016-0477-7

ArboNET. (2015). West Nile virus disease cases reported to CDC by state of residence, 1999-2014. Retrieved from https://www.cdc.gov/westnile/resources/pdfs/data/2-west-nile-virus-disease-cases-reported-to-cdc-by-state_1999-2014_06042015.pdf

Arguez, A., Durre, I., Applequist, S., Russell, V. S., Squires, M. F., Yin, X., . . . Owen, T. W. (2012). *NOAA's 1981-2010 U.S. Climate Normals*. Retrieved from ftp://ftp.ncdc.noaa.gov/pub/data/normals/1981-2010/documentation/1981-2010-normals-overview.pdf

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., . . . Suchard, M. A. (2012). BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Syst Biol, 61*(1), 170-173. doi:10.1093/sysbio/syr100

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol, 29*(9), 2157-2167. doi:10.1093/molbev/mss084

Baele, G., Li, W. L., Drummond, A. J., Suchard, M. A., & Lemey, P. (2013). Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol Biol Evol, 30*(2), 239-243. doi:10.1093/molbev/mss243

Baele, G., Suchard, M. A., Rambaut, A., & Lemey, P. (2016). Emerging Concepts of Data Integration in Pathogen Phylodynamics. *Syst Biol*. doi:10.1093/sysbio/syw054

Bahl, J., Nelson, M. I., Chan, K. H., Chen, R., Vijaykrishna, D., Halpin, R. A., . . . Smith, G. J. (2011). Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc Natl Acad Sci U S A, 108*(48), 19359-19364. doi:10.1073/pnas.1109314108

Beard, R., Magee, D., Suchard, M. A., Lemey, P., & Scotch, M. (2014). Generalized linear models for identifying predictors of the evolutionary diffusion of viruses. *AMIA Jt Summits Transl Sci Proc, 2014*, 23-28.

Bedford, T., Riley, S., Barr, I. G., Broor, S., Chadha, M., Cox, N. J., . . . Russell, C. A. (2015). Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature, 523*(7559), 217-220. doi:10.1038/nature14460

Beerli, P., & Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A, 98*(8), 4563-4568. doi:10.1073/pnas.081068098

Bouma, A., Claassen, I., Natih, K., Klinkenberg, D., Donnelly, C. A., Koch, G., & van Boven, M. (2009). Estimation of Transmission Parameters of H5N1 Avian Influenza Virus in Chickens. *PLoS Pathog, 5*(1), e1000281. doi:10.1371/journal.ppat.1000281

CAPMAS. (2012a, 2012 Jul 05). Arab Republic of Egypt. Retrieved from http://www.citypopulation.de/Egypt.html

CAPMAS. (2012b). Statistical Tables for Population at Governorate Level. Retrieved from http://www.capmas.gov.eg/reports_eng/cens/form_cns_e.aspx?parentid=2940&id=3455&free=1

CDC. (2016a, 17 Sep 2015). 2014-15 Flu Season. Retrieved from http://www.cdc.gov/flu/fluvaxview/1415season.htm

CDC. (2016b, 18 Feb 2016). What You Should Know for the 2014-2015 Influenza Season. Retrieved from http://www.cdc.gov/flu/pastseasons/1415season.htm

Chen, Y., Liu, T., Cai, L., Du, H., & Li, M. (2013). A One-Step RT-PCR Array for Detection and Differentiation of Zoonotic Influenza Viruses H5N1, H9N2, and H1N1. *Journal of Clinical Laboratory Analysis, 27*(6), 450-460. doi:10.1002/jcla.21627

Chipman, H., George, E., & McCulloch, R. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics, 4*(1), 266-298.

Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, 65-134.

Ciccozzi, M., Peletto, S., Cella, E., Giovanetti, M., Lai, A., Gabanelli, E., . . . Zehender, G. (2013). Epidemiological history and phylogeography of West Nile virus lineage 2. *Infect Genet Evol, 17*, 46-50. doi:10.1016/j.meegid.2013.03.034

Cobbin, J. C., Verity, E. E., Gilbertson, B. P., Rockman, S. P., & Brown, L. E. (2013). The source of the PB1 gene in influenza vaccine reassortants selectively alters the hemagglutinin content of the resulting seed virus. *J Virol, 87*(10), 5577-5585. doi:10.1128/jvi.02856-12

Couch, R. B., & Kasel, J. A. (1983). Immunity to influenza in man. *Annu Rev Microbiol, 37*, 529-549. doi:10.1146/annurev.mi.37.100183.002525

Council, T. A. H. (2005). *Most Comprehensive Horse Study Ever Reveals a Nearly $40 Billion Impact on the U.S. Economy*. Retrieved from

Di Giallonardo, F., Geoghegan, J. L., Docherty, D. E., McLean, R. G., Zody, M. C., Qu, J., . . . Holmes, E. C. (2015). Fluid Spatial Dynamics of West Nile Virus in the United States: Rapid Spread in a Permissive Host Environment. *J Virol, 90*(2), 862-872. doi:10.1128/jvi.02305-15

Dolberg, F. (2009). Poultry sector country review: Bangladesh. *Food and Agricultural Organization of the United Nations*.

Drummond, A. J., Ho, S. Y., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol, 4*(5), e88. doi:10.1371/journal.pbio.0040088

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol, 22*(5), 1185-1192. doi:10.1093/molbev/msi103

Drummond, A. J., & Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology, 8*(1), 114.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol, 29*(8), 1969-1973. doi:10.1093/molbev/mss075

Duggal, N. K., Bosco-Lauth, A., Bowen, R. A., Wheeler, S. S., Reisen, W. K., Felix, T. A., . . . Brault, A. C. (2014). Evidence for co-evolution of West Nile Virus and house sparrows in North America. *PLoS Negl Trop Dis, 8*(10), e3262. doi:10.1371/journal.pntd.0003262

FAO. (2014a). FAOSTAT. Retrieved 2014 21 July http://faostat3.fao.org/faostat-gateway/go/to/home/E

FAO. (2014b). Global Livestock Production and Health Atlas. Retrieved 2014 Jul 21 http://kids.fao.org/glipha/index.html

Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., & Lemey, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philos Trans R Soc Lond B Biol Sci, 368*(1614), 20120196. doi:10.1098/rstb.2012.0196

Frost, S. D., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S., & Bedford, T. (2015). Eight challenges in phylodynamic inference. *Epidemics, 10*, 88-92. doi:10.1016/j.epidem.2014.09.001

Gilbert, M., Xiao, X., Pfeiffer, D. U., Epprecht, M., Boles, S., Czarnecki, C., . . . Slingenbergh, J. (2008). Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences, 105*(12), 4769-4774. doi:10.1073/pnas.0710581105

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., & Suchard, M. A. (2013). Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol, 30*(3), 713-724. doi:10.1093/molbev/mss265

Graf, T., Vrancken, B., Maletich Junqueira, D., de Medeiros, R. M., Suchard, M. A., Lemey, P., . . . Pinto, A. R. (2015). Contribution of Epidemiological Predictors in Unraveling the Phylogeographic History of HIV-1 Subtype C in Brazil. *J Virol, 89*(24), 12341-12348. doi:10.1128/jvi.01681-15

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A., & Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science, 303*(5656), 327-332. doi:10.1126/science.1090727

Griffiths, R. C., & Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci, 344*(1310), 403-410. doi:10.1098/rstb.1994.0079

Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol, 22*(2), 160-174.

Herrick, K., Huettmann, F., & Lindgren, M. (2013). A global model of avian influenza prediction in wild birds: the importance of northern regions. *Veterinary Research, 44*(1), 42.

HHS. (2014, 15 Apr 2014). Regional Offices. Retrieved from http://www.hhs.gov/about/agencies/iea/regional-offices/index.html

Horm, S. V., Mardy, S., Rith, S., Ly, S., Heng, S., Vong, S., . . . Buchy, P. (2014). Epidemiological and virological characteristics of influenza viruses circulating in Cambodia from 2009 to 2011. *PLoS One, 9*(10), e110713. doi:10.1371/journal.pone.0110713

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association, 90*(430), 773-795.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications, 13*(3), 235-248. doi:http://dx.doi.org/10.1016/0304-4149(82)90011-4

Koelle, K., & Rasmussen, D. A. (2015). The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *Elife, 4*, e07361. doi:10.7554/eLife.07361

Komar, N. (2000). West Nile viral encephalitis. *Rev Sci Tech, 19*(1), 166-176.

Krauss, H. (2003). *Zoonoses: Infectious Diseases Transmissible from Animals to Humans*: ASM Press.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65-81.

Lam, T. T. Y., Hon, C. C., Lemey, P., Pybus, O. G., Shi, M., Tun, H. M., . . . Leung, F. C. C. (2012). Phylodynamics of H5N1 avian influenza virus in Indonesia. *Molecular Ecology, 21*(12), 3062-3077. doi:10.1111/j.1365-294X.2012.05577.x

Lemey, P., Rambaut, A., Bedford, T., Faria, N., Bielejec, F., Baele, G., . . . Suchard, M. A. (2014). Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog, 10*(2), e1003932. doi:10.1371/journal.ppat.1003932

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol, 5*(9), e1000520. doi:10.1371/journal.pcbi.1000520

Lemey, P., Rambaut, A., & Suchard, M. A. (2014, Sep 2014). Phylogeographic inference in discrete space: a hands on practical. Retrieved from https://perswww.kuleuven.be/~u0036765/crashcourse/BEAST_files/discretePhylogeography_RABV_1.8.1_1.zip

Lopez, R. H., Soto, S. U., & Gallego-Gomez, J. C. (2015). Evolutionary relationships of West Nile virus detected in mosquitoes from a migratory bird zone of Colombian Caribbean. *Virol J, 12*, 80. doi:10.1186/s12985-015-0310-8

Loth, L., Gilbert, M., Wu, J., Czarnecki, C., Hidayat, M., & Xiao, X. (2011). Identifying risk factors of highly pathogenic avian influenza (H5N1 subtype) in Indonesia. *Preventive Veterinary Medicine, 102*(1), 50-58. doi:http://dx.doi.org/10.1016/j.prevetmed.2011.06.006

Magee, D., Beard, R., Suchard, M. A., Lemey, P., & Scotch, M. (2015). Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion. *Arch Virol, 160*(1), 215-224. doi:10.1007/s00705-014-2262-5

Magee, D., Suchard, M. A., & Scotch, M. (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. doi:10.1371/journal.pcbi.1005389

Mann, B. R., McMullen, A. R., Swetnam, D. M., & Barrett, A. D. (2013). Molecular epidemiology and evolution of West Nile virus in North America. *Int J Environ Res Public Health, 10*(10), 5111-5129. doi:10.3390/ijerph10105111

Nelson, M. I., Viboud, C., Vincent, A. L., Culhane, M. R., Detmer, S. E., Wentworth, D. E., . . . Lemey, P. (2015). Global migration of influenza A viruses in swine. *Nat Commun, 6*, 6696. doi:10.1038/ncomms7696

NOAA. (2014). National Climatic Data Center. Retrieved from http://www.ncdc.noaa.gov/

Parker, J., Rambaut, A., & Pybus, O. G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect Genet Evol, 8*(3), 239-246. doi:10.1016/j.meegid.2007.08.001

Pfeiffer, D. U., Minh, P. Q., Martin, V., Epprecht, M., & Otte, M. J. (2007). An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal, 174*(2), 302-309. doi:http://dx.doi.org/10.1016/j.tvjl.2007.05.010

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., . . . Scheuermann, R. H. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res, 40*(Database issue), D593-598. doi:10.1093/nar/gkr859

Pollett, S., Nelson, M. I., Kasper, M., Tinoco, Y., Simons, M., Romero, C., . . . Bausch, D. G. (2015). Phylogeography of Influenza A(H3N2) Virus in Peru, 2010-2012. *Emerg Infect Dis, 21*(8), 1330-1338. doi:10.3201/eid2108.150084

Practitioners, A. A. o. E. West Nile Virus. Retrieved from http://www.aaep.org/info/west-nile-virus-

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., . . . Delwart, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc Natl Acad Sci U S A, 109*(37), 15066-15071. doi:10.1073/pnas.1206598109

Python Software Foundation. (2015). Python Language Reference, version 3.4.3. Retrieved from http://www.python.org

R Core Development Team. (2008). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.R-project.org

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., & Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature, 453*(7195), 615-619. doi:10.1038/nature06945

Sardelis, M. R., Turell, M. J., Dohm, D. J., & O'Guinn, M. L. (2001). Vector competence of selected North American Culex and Coquillettidia mosquitoes for West Nile virus. *Emerg Infect Dis, 7*(6), 1018-1022. doi:10.3201/eid0706.010617

Scotch, M., Mei, C., Makonnen, Y. J., Pinto, J., Ali, A., Vegso, S., . . . Rabinowitz, P. (2013). Phylogeography of influenza A H5N1 clade 2.2.1.1 in Egypt. *BMC Genomics, 14*, 871. doi:10.1186/1471-2164-14-871

Scotch, M., Sarkar, I. N., Mei, C., Leaman, R., Cheung, K. H., Ortiz, P., . . . Gonzalez, G. (2011). Enhancing phylogeography by improving geographical information from GenBank. *J Biomed Inform, 44 Suppl 1*, S44-47. doi:10.1016/j.jbi.2011.06.005

Si, Y., de Boer, W. F., & Gong, P. (2013). Different Environmental Drivers of Highly Pathogenic Avian Influenza H5N1 Outbreaks in Poultry and Wild Birds. *PLoS One, 8*(1), e53362. doi:10.1371/journal.pone.0053362

Slatkin, M., & Maddison, W. P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics, 123*(3), 603-613.

Statistics Canada. (2013, October 2013). Population Estimate. Retrieved from http://www.statcan.gc.ca/start-debut-eng.html

Streicker, D. G., Turmelle, A. S., Vonhof, M. J., Kuzmin, I. V., McCracken, G. F., & Rupprecht, C. E. (2010). Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science, 329*(5992), 676-679. doi:10.1126/science.1188836

Su, Y. C., Bahl, J., Joseph, U., Butt, K. M., Peck, H. A., Koay, E. S., . . . Smith, G. J. (2015). Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun, 6*, 7952. doi:10.1038/ncomms8952

Suchard, M. A., Weiss, R. E., & Sinsheimer, J. S. (2005). Models for Estimating Bayes Factors with Applications to Phylogeny and Tests of Monophyly. *Biometrics, 61*(3), 665-673. doi:10.1111/j.1541-0420.2005.00352.x

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation, 142*(10), 2282-2292.

Tahsin, T., Weissenbacher, D., Rivera, R., Beard, R., Firago, M., Wallstrom, G., . . . Gonzalez, G. (2016). A high-precision rule-based extraction system for expanding geospatial metadata in GenBank records. *J Am Med Inform Assoc, 23*(5), 934-941. doi:10.1093/jamia/ocv172

Tamerius, D., Shaman, J., Alonso, W. J., Bloom-Feshbach, K., Uejio, C. K., Comrie, A., & Viboud, C. (2013). Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. *PLoS Pathog, 9*(3), e1003194. doi:10.1371/journal.ppat.1003194

USFWS. (2016). Geospatial Services. Retrieved from https://www.fws.gov/GIS/

Van Boeckel, T. P., Thanapongtharm, W., Robinson, T., Biradar, C. M., Xiao, X., & Gilbert, M. (2012). Improving Risk Models for Avian Influenza: The Role of Intensive Poultry Farming and Flooded Land during the 2004 Thailand Epidemic. *PLoS One, 7*(11), e49528. doi:10.1371/journal.pone.0049528

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering, 13*(2), 22-30.

Viboud, C., Alonso, W., & Simonsen, L. (2006). Influenza in Tropical Regions. *PLoS Med, 3*(4), e89. doi:10.1371/journal.pmed.0030089

Viboud, C., Bjornstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science, 312*(5772), 447-451. doi:10.1126/science.1125237

WHO. (2011, July 2011). West Nile Virus. Retrieved from http://www.who.int/mediacentre/factsheets/fs354/en/

WHO. (2012). Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. *Influenza and other respiratory viruses, 6*(1), 1-5. doi:10.1111/j.1750-2659.2011.00298.x

WHO. (2013). Cumulative number of confirmed human cases for avian influenza A (H5N1) reported to WHO, 2003-2013.

Williams, C. A., & Crans, W. (2004). West Nile Virus in Horses: Frequently Asked Questions. Retrieved from http://esc.rutgers.edu/fact_sheet/west-nile-virus-in-horses-frequently-asked-questions/

WMO. (2013). Climate Data and Data Related Products. Retrieved from http://www.wmo.int/pages/themes/climate/climate_data_and_products.php

Yang, W., Lipsitch, M., & Shaman, J. (2015). Inference of seasonal and pandemic influenza transmission dynamics. *Proc Natl Acad Sci U S A, 112*(9), 2723-2728. doi:10.1073/pnas.1415012112

Yoon, S. W., Kayali, G., Ali, M. A., Webster, R. G., Webby, R. J., & Ducatez, M. F. (2013). A Single Amino Acid at the Hemagglutinin Cleavage Site Contributes to the Pathogenicity but Not the Transmission of Egyptian Highly Pathogenic H5N1 Influenza Virus in Chickens. *J Virol, 87*(8), 4786-4788. doi:10.1128/jvi.03551-12

Ypma, R. J. F., Bataille, A. M. A., Stegeman, A., Koch, G., Wallinga, J., & van Ballegooijen, W. M. (2012). Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences, 279*(1728), 444-450. doi:10.1098/rspb.2011.0913

Ypma, R. J. F., van Ballegooijen, W. M., & Wallinga, J. (2013). Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics*. doi:10.1534/genetics.113.154856

APPENDIX A

SEQUENCE METADATA FOR CHAPTER 1

Accession[a]	Governorate[b]	Host	Year
CY041290	Al Gharbiyah	Chicken	2008
CY044032	Al Gharbiyah	Chicken	2008
CY061552	Al Qalyubiyah	Chicken	2008
CY062464	Bani Suwayf	Human	2010
CY062466	Ad Daqahliyah	Human	2010
CY062468	Al Qalyubiyah	Human	2010
CY062470	Cairo	Human	2010
CY062472	Al Minufiyah	Human	2010
CY062474	Ad Daqahliyah	Human	2010
CY062476	Kafr ash Shaykh	Human	2010
CY062478	Al Qalyubiyah	Human	2010
CY062480	Al Qalyubiyah	Human	2010
CY062482	Kafr ash Shaykh	Human	2010
CY062484	Cairo	Human	2010
CY062486	Al Fayyum	Human	2010
CY125961	Al Fayyum	Duck	2010
CY125969	Al Qalyubiyah	Duck	2010
CY126034	Al Minufiyah	Chicken	2010
CY126049	Al Minufiyah	Chicken	2010
CY126096	Al Fayyum	Duck	2010
CY126144	Al Minufiyah	Chicken	2010
CY126240	Al Minufiyah	Chicken	2010
CY126248	Al Minufiyah	Chicken	2010
CY126264	Al Minufiyah	Chicken	2010
EU496388	Al Qalyubiyah	Chicken	2007
EU496389	Qina	Chicken	2007
EU496390	Ash Sharqiyah	Turkey	2007
EU496395	Ash Sharqiyah	Chicken	2007
EU496397	Al Buhayrah	Chicken	2007
EU496398	Ad Daqahliyah	Chicken	2008
EU496399	Ad Daqahliyah	Chicken	2008
EU623467	Ash Sharqiyah	Chicken	2007
EU623468	Ash Sharqiyah	Chicken	2007
FJ686831	Al Jizah	Chicken	2008
FJ686832	Cairo	Chicken	2008
FJ686833	Ash Sharqiyah	Chicken	2008
FJ686834	Ad Daqahliyah	Chicken	2008
FJ686835	Al Qalyubiyah	Chicken	2008
FJ686836	Ad Daqahliyah	Chicken	2008
FJ686837	Al Qalyubiyah	Chicken	2008
FJ686838	Al Qalyubiyah	Chicken	2008
FJ686839	Ad Daqahliyah	Chicken	2008
FJ686840	Al Qalyubiyah	Chicken	2008
FJ686841	Ad Daqahliyah	Chicken	2008
FJ686843	Ad Daqahliyah	Chicken	2008

FJ686844	Ash Sharqiyah	Chicken	2008
FJ686845	Al Qalyubiyah	Chicken	2008
FJ686846	Ash Sharqiyah	Chicken	2008
FJ686848	Cairo	Chicken	2008
FJ686849	Ash Sharqiyah	Chicken	2008
FR687256	Al Qalyubiyah	Chicken	2010
FR687257	Al Qalyubiyah	Chicken	2010
FR687258	Al Qalyubiyah	Chicken	2010
GQ184221	Al Qalyubiyah	Chicken	2008
GQ184223	Al Qalyubiyah	Chicken	2008
GQ184227	Al Qalyubiyah	Chicken	2008
GQ184230	Al Gharbiyah	Chicken	2008
GQ184231	Al Qalyubiyah	Chicken	2008
GQ184232	Al Jizah	Chicken	2008
GQ184233	Ash Sharqiyah	Chicken	2008
GQ184236	Al Qalyubiyah	Chicken	2008
GQ184238	Al Jizah	Chicken	2008
GQ184239	Al Qalyubiyah	Chicken	2008
GQ184247	Al Qalyubiyah	Chicken	2008
GQ184248	Al Minufiyah	Chicken	2008
GU002678	Ash Sharqiyah	Duck	2009
GU002683	Kafr ash Shaykh	Chicken	2009
GU002684	Al Qalyubiyah	Chicken	2009
GU002689	Ash Sharqiyah	Chicken	2009
GU002692	Ash Sharqiyah	Chicken	2009
GU002693	Al Qalyubiyah	Chicken	2009
GU002698	Al Qalyubiyah	Chicken	2009
GU002702	Ash Sharqiyah	Turkey	2009
GU002703	Al Uqsur	Chicken	2009
GU002705	Al Qalyubiyah	Chicken	2009
GU064350	Ash Sharqiyah	Chicken	2008
GU064351	Ash Sharqiyah	Chicken	2008
GU064352	Al Qalyubiyah	Chicken	2008
GU064354	Cairo	Chicken	2008
GU064355	Al Qalyubiyah	Chicken	2008
GU811722	Al Gharbiyah	Chicken	2009
GU811726	Al Uqsur	Chicken	2009
GU811745	Qina	Goose	2009
HQ198251	Al Qalyubiyah	Chicken	2009
HQ198252	Al Qalyubiyah	Chicken	2009
HQ198255	Ash Sharqiyah	Chicken	2010
HQ198256	Al Fayyum	Duck	2010
HQ198257	Al Fayyum	Environment	2010
HQ198258	Bani Suwayf	Environment	2010
HQ198261	Al Minufiyah	Chicken	2010
HQ198262	Al Gharbiyah	Chicken	2010

HQ198263	Al Minufiyah	Chicken	2010
HQ198265	Al Qalyubiyah	Chicken	2010
HQ198266	Cairo	Chicken	2010
HQ198268	Al Ismailiyah	Chicken	2010
HQ198269	Al Qalyubiyah	Chicken	2010
HQ198270	Al Buhayrah	Chicken	2010
HQ198271	Al Iskandariyah	Duck	2010
HQ198272	Al Qalyubiyah	Duck	2010
HQ198273	Al Qalyubiyah	Chicken	2010
HQ198274	Cairo	Chicken	2010
HQ198275	Al Fayyum	Chicken	2010
HQ198276	Al Jizah	Turkey	2010
HQ198277	Al Wadi al Jadid	Chicken	2010
HQ198278	Ad Daqahliyah	Chicken	2010
HQ198279	Kafr ash Shaykh	Duck	2010
HQ198280	Al Iskandariyah	Chicken	2010
HQ198281	Al Qalyubiyah	Chicken	2010
HQ198282	Dameitta	Duck	2010
HQ198283	Al Minufiyah	Goose	2010
HQ198284	Ad Daqahliyah	Chicken	2010
HQ198285	Al Gharbiyah	Duck	2010
HQ198287	Al Buhayrah	Duck	2010
HQ198288	Ash Sharqiyah	Chicken	2010
HQ198290	Al Iskandariyah	Chicken	2010
HQ198292	Al Uqsur	Chicken	2010
HQ198293	Al Gharbiyah	Duck	2010
HQ198295	Al Qalyubiyah	Chicken	2010
HQ198296	Al Minya	Chicken	2010
JN807772	Ad Daqahliyah	Chicken	2010
JN807774	Al Minya	Duck	2010
JN807775	Al Iskandariyah	Chicken	2010
JN807776	Al Jizah	Chicken	2010
JN807777	Bani Suwayf	Chicken	2010
JN807778	Al Jizah	Chicken	2010
JN807779	Al Fayyum	Goose	2010
JN807780	Ash Sharqiyah	Duck	2010
JN807782	Ad Daqahliyah	Chicken	2010
JN807783	Bani Suwayf	Duck	2010
JN807784	Al Qalyubiyah	Chicken	2010
JN807785	Al Qalyubiyah	Chicken	2010
JN807786	Ad Daqahliyah	Duck	2010
JN807788	Bani Suwayf	Chicken	2010
JN807789	Ad Daqahliyah	Chicken	2010
JN807790	Al Uqsur	Chicken	2010
JN807791	Kafr ash Shaykh	Turkey	2010
JN807792	Bani Suwayf	Chicken	2010

JN807793	Al Minufiyah	Chicken	2010
JN807794	Al Minufiyah	Duck	2010
JN807795	Al Fayyum	Duck	2010
JN807796	Al Fayyum	Duck	2010
JN807797	Al Minufiyah	Chicken	2010
JN807798	Al Fayyum	Duck	2010
JN807799	Bani Suwayf	Duck	2010
JN807800	Al Minufiyah	Duck	2010
JN807801	Al Jizah	Chicken	2010
JN807802	Al Fayyum	Chicken	2010
JN807803	Ad Daqahliyah	Chicken	2010
JN807804	Al Minya	Chicken	2010
JN807806	Al Qalyubiyah	Chicken	2011
JN807807	Ad Daqahliyah	Chicken	2011
JN807808	Al Qalyubiyah	Chicken	2011
JN807809	Al Jizah	Chicken	2011
JN807810	Al Minufiyah	Duck	2011
JN807811	Ad Daqahliyah	Duck	2011
JN807812	Al Minufiyah	Chicken	2011
JN807813	Al Qalyubiyah	Chicken	2011
JN807814	Al Gharbiyah	Duck	2011
JN807815	Al Fayyum	Goose	2011
JN807816	Al Minya	Duck	2011
JN807817	Al Qalyubiyah	Chicken	2011
JN807818	Al Minufiyah	Chicken	2011
JN807819	As Suways	Chicken	2011
JN807820	Ad Daqahliyah	Chicken	2011
JN807821	Al Fayyum	Chicken	2011
JN807822	Al Fayyum	Chicken	2011
JN807824	Al Fayyum	Chicken	2011
JN807825	Al Fayyum	Chicken	2011
JN807826	Al Qalyubiyah	Chicken	2011
JN807827	Al Fayyum	Chicken	2011
JN807829	Al Fayyum	Duck	2011
JN807830	Al Fayyum	Chicken	2011
JN807832	Bani Suwayf	Chicken	2011
JN807833	Al Fayyum	Duck	2011
JN807834	Al Fayyum	Chicken	2011
JN807835	Al Fayyum	Chicken	2011
JN807836	Ad Daqahliyah	Chicken	2011
JN807837	Al Fayyum	Chicken	2011
JN807838	Al Gharbiyah	Goose	2011
JN807839	Ash Sharqiyah	Chicken	2011
JN807840	Al Buhayrah	Chicken	2011
JN807841	Al Minufiyah	Chicken	2011
JN807842	Al Fayyum	Duck	2011

JN807843	Al Buhayrah	Chicken	2011
JN807844	Al Buhayrah	Chicken	2011
JN807845	Ash Sharqiyah	Chicken	2011
JN807846	Al Gharbiyah	Chicken	2011
JN807847	Al Fayyum	Chicken	2011
JN807848	Al Jizah	Duck	2011
JN807849	Al Qalyubiyah	Duck	2011
JN807850	Al Fayyum	Duck	2011
JN807851	Ash Sharqiyah	Duck	2011
JN807852	Al Minufiyah	Chicken	2011
JN807854	Al Fayyum	Duck	2011
JN807855	Bani Suwayf	Chicken	2011
JN807856	Cairo	Chicken	2011
JN807857	Ash Sharqiyah	Duck	2011
JN807858	Al Fayyum	Chicken	2011
JN807859	Al Minya	Duck	2011
JN807860	Al Fayyum	Duck	2011
JN807861	Al Fayyum	Duck	2011
JN807862	Al Jizah	Goose	2011
JN807863	Al Fayyum	Chicken	2011
JN807865	Al Uqsur	Chicken	2011
JN807866	Bour Said	Quail	2011
JN807867	Al Uqsur	Chicken	2011
JQ858469	Al Minya	Chicken	2011
JQ858470	Al Fayyum	Duck	2011
JQ858471	Al Jizah	Chicken	2011
JQ858472	Al Qalyubiyah	Chicken	2011
JQ858473	Al Minufiyah	Chicken	2011
JQ858475	Al Minufiyah	Chicken	2011
JQ858476	Al Qalyubiyah	Duck	2011
JQ858477	Al Fayyum	Chicken	2011
JQ858478	Al Jizah	Duck	2011
JQ858479	Al Jizah	Chicken	2011
JQ858480	Al Minya	Duck	2011
JQ858481	Al Jizah	Chicken	2011
JQ858482	Al Minufiyah	Chicken	2011
JQ858483	Al Minya	Chicken	2012
JQ858484	Al Minufiyah	Chicken	2012
JQ858485	Al Minufiyah	Chicken	2012
JQ858486	Al Minufiyah	Chicken	2012
JX456101	Al Buhayrah	Human	2012
JX456104	Al Jizah	Human	2012
JX576786	Ad Daqahliyah	Duck	2011

[a] GenBank

[b] Governorate of Egypt

APPENDIX B

SEQUENCE METADATA FOR CHAPTER 2

Region[a]	Sample 1[b]	Sample 2[b]	Sample 3[b]	Sample 4[b]	Sample 5[b]	Sample 6[b]
1	168130	168702	167940	167411	168127	167933
1	168703	168703	168130	168702	168715	168702
1	169334	168715	168715	168703	169908	168716
1	169484	169902	169285	169487	170037	169285
1	169490	169926	169490	169906	170111	169902
1	169906	170111	170037	169926	170150	172493
1	169908	170685	170107	170111	170693	172505
1	170685	170693	172565	170692	172508	172508
1	170693	172493	174122	170696	172730	172730
1	172495	172495	174187	172500	174159	174159
1	172565	172502	175183	172565	174175	174187
1	172730	172505	175185	174122	174187	174188
1	174122	172508	176510	174169	176537	175183
1	174132	174122	176535	174188	176540	175198
1	174149	174159	176625	175183	176628	175207
1	174159	175185	176651	175185	176651	176535
1	174169	176532	176659	176535	177537	176537
1	174187	176540	178990	176700	178979	176642
1	174188	178980	178992	178981	178990	177537
1	176733	178991	178993	178993	178993	178996
1	178980	178997	178996	178996	191047	181072
1	178996	191048	178997	191044	191048	188889
1	188889	191065	191044	191046	193343	191047
1	191048	191691	191434	191058	193349	191058
1	191065	193343	192159	194128	194128	191069
1	193343	193349	195891	194168	194134	193349
2	169296	169476	169307	169506	169296	169501
2	169307	169501	169309	172543	170035	169505
2	169476	169507	169476	172579	173852	170035
2	169501	172543	170035	172831	174128	172543
2	169507	172579	172831	174152	174146	172578
2	169508	174128	174152	174153	174152	174185
2	174152	174158	174161	174161	174162	175229
2	174158	175206	174185	174185	174192	176640
2	174192	175219	175167	175219	175167	176736
2	176746	176641	175220	176743	175219	178988
2	181071	176743	176736	176750	175220	178999
2	191043	191427	176746	178987	181071	181071
2	194179	191669	193319	194179	191669	191669
3	168129	167952	169942	169331	167952	169119
3	169331	169119	170039	169338	169311	169338
3	169942	169504	170049	169498	169331	169504
3	170136	170014	170136	170015	169337	169934
3	170712	170125	170697	170039	169504	170730
3	171385	170730	170702	170125	169912	171389

3	172564	171389	170715	170697	170039	171840
3	172574	171841	171843	170702	170125	172501
3	172575	172513	172564	170736	170702	172513
3	172590	172564	172575	172501	170737	172544
3	172610	172575	172594	172512	171843	172564
3	172751	172610	172610	172513	172501	172574
3	172754	172782	172751	172544	172588	172590
3	172782	175227	172754	172571	172590	172738
3	172783	175228	172783	175216	172782	174140
3	174140	176546	175196	176494	175227	175181
3	175181	176562	175227	176544	175228	175228
3	176494	176624	176503	176545	176503	176550
3	176511	177526	176546	176546	176509	176556
3	176543	177527	176559	176632	176511	176645
3	176545	177545	176562	176643	176550	176646
3	176632	178983	176644	176645	176663	176666
3	176644	188879	176666	188879	176666	176740
3	176666	188892	176740	188892	176745	177524
3	177526	191063	177535	192161	177535	177526
3	177535	192165	178982	192165	178985	177535
3	177545	192191	182625	192191	188893	178982
3	192174	193357	192186	193332	192174	191063
3	194132	195543	194180	194132	192191	192161
4	167396	167405	167404	167417	167406	167405
4	167405	167406	167405	167927	167413	167406
4	167407	167407	167407	167946	167414	167407
4	167414	167414	167413	168117	167926	167415
4	167418	167925	167416	168699	168116	167417
4	168108	167926	168116	169093	168699	167924
4	168699	168116	168124	169094	169095	167926
4	168705	169092	168698	169096	169126	168124
4	169095	169096	168699	169286	169317	169115
4	169304	169458	169093	169343	169340	169126
4	169315	169470	169094	169456	169470	169321
4	169470	169477	169115	169470	169935	169323
4	169477	170106	169126	169477	169947	169343
4	169929	170133	169315	169929	170023	169456
4	169935	170711	169323	170043	170718	170021
4	169947	170720	169343	170132	170720	170023
4	170016	170726	169456	170721	170726	170043
4	170133	172550	170021	171361	171354	170132
4	170714	172585	170028	171366	171362	170716
4	170723	172595	170133	172538	171367	170718
4	171354	172740	170134	172550	172538	171361
4	171362	172741	170704	172551	172585	171366
4	172559	172760	170723	172560	172596	171367

4	172745	172793	171387	172580	172801	171379
4	173217	172801	172536	172595	172806	172536
4	173225	172807	172596	172602	172809	172551
4	173246	173217	172740	172742	173220	172559
4	173247	173222	172743	172743	173225	172585
4	173256	173223	172748	172748	173246	172743
4	174157	173225	172793	172794	173247	172748
4	174173	174129	173220	172806	174151	172805
4	174177	174157	173225	173246	174157	172809
4	175162	174171	174189	175157	174191	173238
4	175171	175162	175176	175158	175162	173256
4	175176	175184	175251	175184	175176	174129
4	176488	176571	176489	175251	176572	175157
4	176490	176706	176571	176490	176715	175169
4	176571	177525	176751	176576	177531	175171
4	176751	177534	177531	176706	177538	176490
4	178972	177538	181075	176715	178971	176717
4	179003	178969	188868	176722	179002	178969
4	179010	179009	188890	178967	179003	179003
4	181077	179010	191050	179004	179004	181077
4	188883	188867	191052	179010	188880	188880
4	191052	191031	191055	181074	188883	188883
4	191067	191050	191067	188887	191052	188887
4	191703	191062	191678	191031	191067	191050
4	193351	191678	193310	191052	191678	191067
4	193353	191703	193359	191067	193328	193309
4	193354	193354	197491	193351	197486	197491
5	169453	169104	169085	169087	168706	168706
5	169909	169479	169120	169106	169086	169101
5	169910	169480	169130	169130	169289	169120
5	169933	169483	169344	169453	169344	169344
5	170042	169915	169479	169488	169488	169482
5	170122	169916	169482	169911	169491	169909
5	170684	170691	169483	169948	169494	169910
5	172496	172504	169494	170118	169924	169915
5	172504	172516	169917	170123	169930	169933
5	172506	172541	169930	170126	169933	170040
5	172547	172557	170029	170724	169948	170122
5	172553	172744	170047	172515	170029	170724
5	172758	173229	170123	172542	170047	172504
5	173236	173236	170724	172563	170684	172506
5	173240	173242	170729	172758	172496	172515
5	173242	173244	172548	172804	172516	172533
5	173258	173245	172563	173218	172533	172542
5	174172	173255	172758	173221	172553	172553
5	175159	174165	173219	173239	172557	172804

5	175163	174172	173224	173240	173234	173219
5	175192	175192	173255	173255	173239	173258
5	176555	176491	175156	173258	173255	174165
5	176567	176566	175160	173860	173258	175156
5	176575	176656	175163	174165	174181	175159
5	176652	176657	175170	175156	175163	175160
5	176656	176658	175192	175170	175170	176575
5	176657	176721	176538	176491	176491	176656
5	176701	178994	176569	176538	176539	176657
5	176721	179000	176701	176555	176560	176708
5	176732	179001	176723	176656	176647	176709
5	178994	188865	178977	176708	176656	176732
5	179008	188877	178989	178977	176710	178976
5	188865	191425	178994	188873	176721	178994
5	188873	191429	181078	188877	176742	179007
5	188874	191709	188874	191424	178976	188866
5	191701	191713	188882	191689	188877	191429
5	191713	191715	191701	191710	191670	193307
5	192183	192183	191710	191715	191701	193311
5	193311	193352	193331	192183	194130	193331
5	194175	194130	194183	194131	194131	194131
6	168709	168098	167950	168709	168709	166984
6	168711	169102	168098	168725	169102	168708
6	168712	169293	168118	169293	169292	169089
6	169102	169294	168707	169329	169467	169099
6	169108	169903	168708	169467	169481	169290
6	169291	169938	168725	169485	169485	169291
6	169329	169949	168726	169493	169903	169329
6	169335	170041	168727	169913	169907	169335
6	169486	170674	169102	169932	169932	169485
6	169944	170708	169329	170022	169940	169486
6	170022	170731	169481	170038	169944	169502
6	170038	170738	169502	170128	170676	169907
6	170041	171839	169932	170146	170701	169940
6	170675	172761	170022	170674	170708	169949
6	170679	172832	170674	170677	170709	170022
6	170701	172833	170679	170695	172832	170680
6	170709	172834	170680	172761	172834	170701
6	170742	173231	170687	172834	173231	170740
6	172739	173232	170695	173849	173232	172231
6	172834	173250	170701	174135	173250	172582
6	173216	173252	170742	174156	173845	173231
6	174135	173845	172608	175189	173857	173232
6	175175	173849	172833	176501	175175	173250
6	175224	174156	174136	176561	175188	173857
6	176501	174170	174144	176565	175189	174138

6	176506	176497	175175	176650	176495	175175
6	176507	176633	175189	176675	176506	175189
6	176553	176635	176501	176677	176508	176495
6	176622	176675	176636	176681	176565	176497
6	176634	176676	176637	176685	176650	176637
6	176678	176679	176638	176693	176676	176685
6	176680	176687	176639	176731	176680	176739
6	176693	176739	176693	178964	176739	178959
6	176739	178962	176739	178975	178962	178964
6	178959	178975	178964	181079	188895	178975
6	182624	178986	182624	188870	191037	188870
6	191702	188894	188895	191688	191688	191688
6	193308	191037	191688	194154	194154	193308
7	169103	169103	169107	169107	169103	169318
7	169336	169118	169283	169118	170108	169346
7	169914	169283	169925	169336	172498	169925
7	170713	169925	169937	170017	174134	169937
7	172499	170113	170108	170113	174168	172499
7	172507	172498	171384	172507	174179	172539
7	172539	172499	174145	172545	175193	172545
7	174134	172545	174186	176504	175214	172598
7	174168	175161	175172	176738	176524	174168
7	176524	175172	176522	177529	176564	176504
7	176629	175193	176534	191035	176688	176697
7	177529	176536	176536	191038	176697	176738
7	191035	176629	176564	194171	176738	181073
7	193356	176697	176660	194176	177529	191038
7	194176	191035	176688	194182	193350	194182
7	194177	193356	194139	195872	194177	195872
7	195882	195882	194176	195895	194182	195882
8	167402	168099	168103	168101	167402	168113
8	168102	168700	168109	169098	167408	168115
8	168701	169110	168700	169928	168101	169098
8	169110	169112	169098	170024	168102	169125
8	169113	169113	169112	170703	168109	169300
8	169306	169300	169128	170719	169111	170024
8	169904	169904	169306	170722	169112	170705
8	170020	170127	169322	172520	169113	170707
8	170690	170706	169452	172561	169308	170725
8	172593	170707	170020	172562	169475	172558
8	172736	170717	170027	172573	169928	172562
8	173850	170725	170036	172736	169939	172746
8	173859	170734	170698	173235	170024	174123
8	174154	172561	172554	173853	170706	174154
8	175154	172562	172749	173859	170734	174155
8	175180	172593	173850	174143	172747	174178

8	176705	172749	173859	174155	172749	174190
8	176711	172753	174143	175154	173850	175165
8	176713	174143	174154	175179	173853	175180
8	178968	174155	175164	176714	174190	176705
8	178974	175154	175180	176718	176671	176711
8	191423	175164	176671	176719	176711	178973
8	191671	175165	176719	178968	176719	191423
8	191676	175180	191676	191671	191423	191671
8	191682	178974	191683	191673	191677	191673
8	191712	179005	191711	191674	191714	191681
8	193333	191675	191714	191675	191716	191712
8	194140	191682	193360	191676	194146	192170
8	194147	191683	194150	191714	194150	193333
8	194150	191711	194158	194140	194158	193344
9	169123	169124	170044	170045	169314	169123
9	169124	169314	170137	170688	170110	169314
9	170045	169923	170688	170710	170137	170137
9	170110	170045	170699	172497	172497	171351
9	172497	170137	170739	174176	174131	172494
9	174147	170681	172607	175190	174133	172607
9	175191	170682	172611	176531	174174	172609
9	176523	170699	176528	176689	175187	174130
9	176673	170710	176689	176696	176493	174131
9	176741	172497	176734	191039	176516	174176
9	176748	172607	176749	191696	176529	175187
9	188864	175187	178957	192182	176531	176493
9	191039	175190	191032	193312	176689	176523
9	191696	176689	191033	193346	176749	176741
9	194133	176735	191698	193347	191033	191039
9	194137	176741	192182	194133	193346	191695
9	194144	178957	193347	194149	194148	191698
9	194161	194144	194148	194161	194153	194144
9	195890	194149	195890	195890	195890	194161
10	168721	167953	167953	168718	169298	167953
10	169129	168720	168720	168720	170033	168718
10	169298	169129	169945	169918	170689	168719
10	169495	169298	170129	169945	170728	168721
10	169918	169918	171837	170120	172583	170025
10	169945	170025	172830	170728	173237	170033
10	170025	170728	173227	172510	173251	170129
10	170120	176502	173241	172576	176502	170689
10	170129	176526	174141	172756	176521	171838
10	170728	176551	176502	172830	176526	172763
10	173227	176648	176654	175213	176648	173241
10	173251	176653	176726	176684	176672	173251
10	175213	176683	176737	176724	176684	173254

10	176725	176724	178963	176737	176726	176683
10	176726	178958	178978	177539	178965	176684
10	178978	178995	188878	178958	178995	176692
10	178995	188878	188888	191029	188878	178995
10	188878	188881	191059	191040	191431	188876
10	191040	191040	191432	191431	191432	191029
10	191041	191685	191692	191432	192176	191034
10	191059	192177	192187	192176	192177	192177
10	192177	192178	192188	192187	194138	193345
10	192178	193361	195893	195893	195893	193355

[a] Region of the U.S. Department of Health and Human Services

[b] GISAID accessions for whole genomes; hemagglutinin genes were used in the study

APPENDIX C

SEQUENCE METADATA FOR CHAPTER 3

Accession[a]	CBR[b]	CBS	State	County	Year
DQ164186	NE	Middle Atlantic	South Dakota	San Bernardino	2002
DQ164187	NE	Middle Atlantic	New York	Broome	2002
DQ164188	NE	Middle Atlantic	New York	Westchester	2003
DQ164189	NE	Middle Atlantic	New York	Albany	2003
DQ164190	NE	Middle Atlantic	New York	Suffolk	2003
DQ164191	NE	Middle Atlantic	New York	Chautauqua	2003
DQ164192	NE	Middle Atlantic	New York	Rockland	2003
DQ164193	NE	Middle Atlantic	New York	Clinton	2002
DQ164194	NE	Middle Atlantic	New York	Suffolk	2001
DQ164195	NE	Middle Atlantic	New York	Nassau	2002
DQ164196	South	South Atlantic	Georgia	Wilkinson	2002
DQ164197	South	South Atlantic	Georgia	Wilkinson	2002
DQ164198	South	West South Central	Texas	Concho	2002
DQ164199	South	West South Central	Texas	Concho	2003
DQ164200	MW	East North Central	Indiana	Hendricks	2002
DQ164201	West	Mountain	Arizona	Yavapai	2004
DQ164202	MW	East North Central	Ohio	Licking	2002
DQ164203	West	Mountain	Colorado	Park	2003
DQ164204	West	Mountain	Colorado	Park	2003
DQ164205	South	West South Central	Texas	Concho	2002
DQ164206	South	West South Central	Texas	Harris	2004
DQ431693	South	West South Central	Texas	Randall	2003
DQ431695	MW	East North Central	Illinois	Cook	2003
DQ431696	MW	East North Central	Wisconsin	Milwaukee	2003
DQ431697	South	South Atlantic	Florida	Hillsborough	2003
DQ431698	South	South Atlantic	Florida	Hillsborough	2003
DQ431699	South	South Atlantic	Florida	Hillsborough	2003
DQ431700	West	Pacific	California	San Francisco	2004
DQ431701	West	Mountain	Colorado	Mesa	2004
DQ431702	West	Mountain	Colorado	Mesa	2004
DQ431703	West	Mountain	Colorado	Mesa	2004
DQ431704	West	Mountain	Colorado	Mesa	2004
DQ431705	MW	West North Central	South Dakota	Pennington	2004
DQ431706	West	Mountain	New Mexico	Sandoval	2004
DQ431707	West	Mountain	New Mexico	Sandoval	2004
DQ431708	West	Pacific	California	San Diego	2004
DQ431709	West	Pacific	California	San Bernardino	2004
DQ431710	West	Pacific	California	Orange	2004
DQ431711	West	Mountain	Arizona	Maricopa	2004
DQ431712	West	Mountain	Arizona	Maricopa	2004
EF530047	NE	Middle Atlantic	New York	Richmond	2000
EF657887	NE	Middle Atlantic	New York	Richmond	2000
FJ151394	NE	Middle Atlantic	New York	New York	1999
FJ527738	South	West South Central	Louisiana	Jefferson	2001
GQ507468	South	West South Central	Texas	El Paso	2005

GQ507469	West	Mountain	New Mexico	Dona Ana	2005
GQ507470	South	West South Central	Texas	El Paso	2006
GQ507471	South	West South Central	Texas	El Paso	2007
GQ507472	West	Pacific	California	Orange	2003
GQ507473	West	Pacific	California	Los Angeles	2004
GQ507474	West	Pacific	California	San Bernardino	2004
GQ507475	West	Pacific	California	San Bernardino	2005
GQ507476	West	Pacific	California	San Bernardino	2005
GQ507477	West	Pacific	California	Los Angeles	2005
GQ507478	West	Pacific	California	Los Angeles	2005
GQ507479	West	Mountain	Arizona	Pima	2005
GQ507480	West	Pacific	California	Los Angeles	2005
GQ507481	MW	West North Central	Nebraska	Douglas	2006
GQ507482	West	Mountain	Arizona	Pima	2006
GQ507483	West	Pacific	California	Los Angeles	2007
GQ507484	West	Pacific	California	Los Angeles	2007
GU827998	South	West South Central	Texas	Harris	2002
GU827999	South	West South Central	Texas	Montgomery	2003
GU828000	South	West South Central	Texas	Harris	2003
GU828001	South	West South Central	Texas	Harris	2003
GU828002	South	West South Central	Texas	Harris	2003
GU828003	South	West South Central	Texas	Jefferson	2003
GU828004	South	West South Central	Texas	Montgomery	2003
HM488114	NE	New England	Connecticut	Fairfield	2002
HM488115	NE	New England	Connecticut	Fairfield	2005
HM488116	NE	New England	Connecticut	Fairfield	2005
HM488117	NE	New England	Connecticut	Fairfield	2005
HM488118	NE	New England	Connecticut	Fairfield	2005
HM488119	NE	New England	Connecticut	Fairfield	2005
HM488120	NE	New England	Connecticut	Fairfield	2005
HM488121	NE	New England	Connecticut	Fairfield	2005
HM488125	NE	New England	Connecticut	Fairfield	1999
HM488126	NE	New England	Connecticut	Fairfield	1999
HM488127	NE	New England	Connecticut	Fairfield	1999
HM488128	NE	New England	Connecticut	Fairfield	1999
HM488129	NE	New England	Connecticut	New Haven	2000
HM488130	NE	New England	Connecticut	New Haven	2000
HM488131	NE	New England	Connecticut	New Haven	2000
HM488132	NE	New England	Connecticut	Fairfield	2000
HM488133	NE	New England	Connecticut	Fairfield	2001
HM488134	NE	New England	Connecticut	Fairfield	2001
HM488135	NE	New England	Connecticut	Fairfield	2001
HM488136	NE	New England	Connecticut	Fairfield	2001
HM488137	NE	New England	Connecticut	Fairfield	2002
HM488138	NE	New England	Connecticut	Fairfield	2003
HM488139	NE	New England	Connecticut	Fairfield	2003

HM488140	NE	New England	Connecticut	Fairfield	2003
HM488141	NE	New England	Connecticut	Fairfield	2003
HM488142	NE	New England	Connecticut	Fairfield	2004
HM488143	NE	New England	Connecticut	Fairfield	2004
HM488144	NE	New England	Connecticut	Fairfield	2004
HM488145	NE	New England	Connecticut	Fairfield	2004
HM488146	NE	New England	Connecticut	Fairfield	2004
HM488147	NE	New England	Connecticut	Fairfield	2004
HM488148	NE	New England	Connecticut	Fairfield	2004
HM488149	NE	New England	Connecticut	Fairfield	2005
HM488150	NE	New England	Connecticut	Fairfield	2005
HM488151	NE	New England	Connecticut	Fairfield	2005
HM488152	NE	New England	Connecticut	Fairfield	2005
HM488153	NE	New England	Connecticut	Fairfield	2005
HM488154	NE	New England	Connecticut	Fairfield	2005
HM488155	NE	New England	Connecticut	Fairfield	2006
HM488156	NE	New England	Connecticut	Fairfield	2006
HM488157	NE	New England	Connecticut	Fairfield	2006
HM488158	NE	New England	Connecticut	Fairfield	2006
HM488159	NE	New England	Connecticut	Fairfield	2006
HM488160	NE	New England	Connecticut	Fairfield	2006
HM488161	NE	New England	Connecticut	Fairfield	2007
HM488162	NE	New England	Connecticut	Fairfield	2007
HM488163	NE	New England	Connecticut	Fairfield	2007
HM488164	NE	New England	Connecticut	Fairfield	2007
HM488165	NE	New England	Connecticut	Fairfield	2007
HM488166	NE	New England	Connecticut	Fairfield	2008
HM488167	NE	New England	Connecticut	Fairfield	2008
HM488168	NE	New England	Connecticut	Fairfield	2008
HM488169	NE	New England	Connecticut	Fairfield	2008
HM488170	NE	New England	Connecticut	Fairfield	2008
HM488171	NE	New England	Connecticut	Fairfield	2003
HM488172	NE	New England	Connecticut	Fairfield	2003
HM488173	NE	New England	Connecticut	New Haven	2003
HM488174	NE	New England	Connecticut	New Haven	2003
HM488175	NE	New England	Connecticut	Hartford	2003
HM488176	NE	New England	Connecticut	New Haven	2003
HM488177	MW	East North Central	Illinois	Cook	2002
HM488178	MW	East North Central	Illinois	Cook	2002
HM488180	MW	East North Central	Illinois	Cook	2002
HM488181	MW	East North Central	Illinois	Iroquois	2002
HM488182	MW	East North Central	Illinois	Clinton	2002
HM488183	MW	East North Central	Illinois	Douglas	2002
HM488184	MW	East North Central	Illinois	Moultrie	2002
HM488185	MW	East North Central	Illinois	Cook	2003
HM488186	MW	East North Central	Illinois	Champaign	2003

HM488188	MW	East North Central	Illinois	Vermilion	2004
HM488189	MW	East North Central	Illinois	Will	2004
HM488190	MW	East North Central	Illinois	Cook	2004
HM488191	MW	East North Central	Illinois	Cook	2004
HM488192	MW	East North Central	Illinois	Rock Island	2005
HM488193	MW	East North Central	Illinois	St. Clair	2005
HM488194	MW	East North Central	Illinois	Lake	2005
HM488195	MW	East North Central	Illinois	Kendall	2005
HM488196	MW	East North Central	Illinois	Cook	2005
HM488197	MW	East North Central	Illinois	McHenry	2005
HM488203	NE	Middle Atlantic	New York	Putnam	2008
HM488204	NE	Middle Atlantic	New York	Suffolk	2008
HM488205	NE	Middle Atlantic	New York	Albany	2008
HM488206	NE	Middle Atlantic	New York	Erie	2008
HM488207	NE	Middle Atlantic	New York	Nassau	2008
HM488208	NE	New England	Connecticut	Fairfield	2002
HM488209	NE	New England	Connecticut	Fairfield	2003
HM488210	NE	New England	Connecticut	New Haven	2003
HM488212	NE	New England	Connecticut	New Haven	2003
HM488213	NE	New England	Connecticut	Fairfield	2003
HM488214	NE	New England	Connecticut	Fairfield	2003
HM488215	NE	New England	Connecticut	Fairfield	2003
HM488216	NE	New England	Connecticut	New London	2003
HM488217	NE	New England	Connecticut	New Haven	2003
HM488218	NE	New England	Connecticut	Fairfield	2003
HM488219	NE	New England	Connecticut	Hartford	2003
HM488220	NE	New England	Connecticut	New Haven	2003
HM488221	NE	New England	Connecticut	New London	2003
HM488222	NE	New England	Connecticut	New London	2003
HM488223	NE	New England	Connecticut	Fairfield	2003
HM488224	NE	New England	Connecticut	Fairfield	2003
HM488225	NE	New England	Connecticut	New Haven	2003
HM488226	NE	New England	Connecticut	New Haven	2003
HM488227	NE	New England	Connecticut	New Haven	2003
HM488228	NE	New England	Connecticut	New Haven	2003
HM488229	NE	New England	Connecticut	New Haven	2003
HM488230	NE	New England	Connecticut	Windham	2003
HM488231	NE	New England	Connecticut	Middlesex	2003
HM488232	NE	New England	Connecticut	Middlesex	2003
HM488233	NE	New England	Connecticut	New Haven	2003
HM488234	NE	New England	Connecticut	New Haven	2003
HM488235	NE	New England	Connecticut	Fairfield	2003
HM488236	NE	New England	Connecticut	Middlesex	2003
HM488237	NE	Middle Atlantic	New York	Onondaga	2008
HM488238	NE	Middle Atlantic	New York	Onondaga	2008
HM488239	NE	Middle Atlantic	New York	Putnam	2008

HM488240	NE	Middle Atlantic	New York	Suffolk	2008
HM488241	NE	Middle Atlantic	New York	Niagara	2008
HM488242	NE	Middle Atlantic	New York	Dutchess	2008
HM488243	NE	Middle Atlantic	New York	Suffolk	2008
HM488244	NE	Middle Atlantic	New York	Erie	2008
HM488245	NE	Middle Atlantic	New York	Putnam	2008
HM488246	NE	Middle Atlantic	New York	Kings	2001
HM488247	NE	Middle Atlantic	New York	New York	2001
HM488248	NE	Middle Atlantic	New York	Herkimer	2001
HM488249	NE	Middle Atlantic	New York	Onondaga	2001
HM488250	NE	Middle Atlantic	New York	Broome	2003
HM488251	NE	Middle Atlantic	New York	Cortland	2003
HM488252	NE	Middle Atlantic	New York	Onondaga	2005
HM756648	NE	New England	Connecticut	Fairfield	2002
HM756649	NE	New England	Connecticut	Fairfield	2006
HM756650	NE	New England	Connecticut	New Haven	2003
HM756651	NE	New England	Connecticut	Fairfield	2003
HM756652	NE	New England	Connecticut	Middlesex	2003
HM756653	NE	New England	Connecticut	Middlesex	2003
HM756654	NE	New England	Connecticut	Fairfield	2003
HM756656	NE	New England	Connecticut	New London	2003
HM756657	NE	New England	Connecticut	Fairfield	2003
HM756658	NE	New England	Connecticut	New London	2003
HM756659	NE	New England	Connecticut	Middlesex	2003
HM756660	NE	Middle Atlantic	New York	Livingston	2008
HM756661	NE	Middle Atlantic	New York	Bronx	2001
HM756662	NE	Middle Atlantic	New York	Albany	2001
HM756663	NE	Middle Atlantic	New York	Albany	2001
HM756664	NE	Middle Atlantic	New York	Albany	2002
HM756665	NE	Middle Atlantic	New York	Dutchess	2002
HM756666	NE	Middle Atlantic	New York	Saratoga	2003
HM756667	NE	Middle Atlantic	New York	Onondaga	2003
HM756668	NE	Middle Atlantic	New York	Columbia	2003
HM756669	NE	Middle Atlantic	New York	Saratoga	2003
HM756670	NE	Middle Atlantic	New York	Queens	2003
HM756671	NE	Middle Atlantic	New York	Cortland	2004
HM756672	NE	Middle Atlantic	New York	Nassau	2004
HM756673	NE	Middle Atlantic	New York	Oswego	2004
HM756675	NE	Middle Atlantic	New York	Monroe	2005
HM756676	MW	East North Central	Illinois	Perry	2003
HM756677	West	Mountain	New Mexico	Bernalillo	2005
HM756678	NE	Middle Atlantic	New York	Jefferson	2007
HQ671721	NE	Middle Atlantic	New York	Tompkins	2008
HQ671722	NE	Middle Atlantic	New York	Jefferson	2002
HQ671723	NE	Middle Atlantic	New York	Putnam	2003
HQ671724	NE	Middle Atlantic	New York	Broome	2005

HQ671725	NE	Middle Atlantic	New York	Lewis	2005
HQ671726	NE	Middle Atlantic	New York	Putnam	2005
HQ671727	NE	Middle Atlantic	New York	Orleans	2006
HQ671728	NE	Middle Atlantic	New York	Richmond	2006
HQ671729	NE	Middle Atlantic	New York	Suffolk	2006
HQ671730	NE	Middle Atlantic	New York	Onondaga	2007
HQ671742	MW	East North Central	Illinois	Perry	2002
HQ705660	NE	Middle Atlantic	New York	Orange	2003
HQ705669	MW	East North Central	Illinois	Clinton	2002
JF415914	South	West South Central	Texas	Harris	2005
JF415915	South	West South Central	Texas	Harris	2006
JF415916	South	West South Central	Texas	Harris	2006
JF415917	South	West South Central	Texas	Harris	2007
JF415918	South	West South Central	Texas	Harris	2007
JF415919	South	West South Central	Texas	Harris	2007
JF415920	South	West South Central	Texas	Harris	2007
JF415921	South	West South Central	Texas	Harris	2008
JF415922	South	West South Central	Texas	Harris	2009
JF415923	South	West South Central	Texas	Harris	2009
JF415924	South	West South Central	Texas	Harris	2009
JF415925	South	West South Central	Texas	Harris	2009
JF415926	South	West South Central	Texas	Harris	2009
JF415927	South	West South Central	Texas	Harris	2009
JF415928	South	West South Central	Texas	Harris	2009
JF415929	South	West South Central	Texas	Harris	2005
JF415930	South	West South Central	Texas	Harris	2006
JF488094	NE	Middle Atlantic	New York	Dutchess	2004
JF488095	NE	Middle Atlantic	New York	Albany	2009
JF488096	NE	Middle Atlantic	New York	Suffolk	2009
JF488097	NE	Middle Atlantic	New York	Suffolk	2007
JF703161	West	Pacific	California	Imperial	2004
JF703162	West	Pacific	California	Riverside	2003
JF703163	West	Pacific	California	Imperial	2005
JF703164	West	Pacific	California	Riverside	2003
JF730042	NE	Middle Atlantic	New York	Niagara	2007
JF899528	NE	Middle Atlantic	New York	Suffolk	2004
JN183885	NE	Middle Atlantic	New York	Orleans	2008
JN183886	NE	Middle Atlantic	New York	Niagara	2008
JN183887	NE	Middle Atlantic	New York	Oswego	2002
JN183891	MW	East North Central	Illinois	Perry	2002
JN367277	NE	Middle Atlantic	New York	Niagara	2004
JX015515	South	West South Central	Texas	El Paso	2005
JX015516	South	West South Central	Texas	El Paso	2007
JX015517	South	West South Central	Texas	El Paso	2008
JX015519	South	West South Central	Texas	El Paso	2009
JX015521	South	West South Central	Texas	El Paso	2009

JX015522	South	West South Central	Texas	El Paso	2010
JX015523	South	West South Central	Texas	El Paso	2010
KC736486	South	West South Central	Texas	Montgomery	2012
KC736487	South	West South Central	Texas	Montgomery	2012
KC736488	South	West South Central	Texas	Montgomery	2012
KC736489	South	West South Central	Texas	Montgomery	2012
KC736490	South	West South Central	Texas	Montgomery	2012
KC736491	South	West South Central	Texas	Dallas	2012
KC736492	South	West South Central	Texas	Dallas	2012
KC736493	South	West South Central	Texas	Dallas	2012
KC736494	South	West South Central	Texas	Montgomery	2012
KC736495	South	West South Central	Texas	Dallas	2012
KC736496	South	West South Central	Texas	Montgomery	2012
KC736497	South	West South Central	Texas	Montgomery	2012
KC736498	South	West South Central	Texas	Montgomery	2012
KC736499	South	West South Central	Texas	Montgomery	2012
KC736500	South	West South Central	Texas	Dallas	2012
KC736501	South	West South Central	Texas	Dallas	2012
KC736502	South	West South Central	Texas	Dallas	2012
KF704147	West	Mountain	Arizona	Maricopa	2010
KF704153	West	Mountain	Arizona	Maricopa	2010
KF704158	West	Mountain	Arizona	Maricopa	2010
KJ786935	South	West South Central	Texas	Harris	2012
KJ786936	South	West South Central	Texas	Harris	2012

[a] GenBank

[b] Midwest (MW); Northeast (NE)

APPENDIX D

STATEMENTS FROM CO-AUTHORS IN PUBLISHED WORK

Chapters 1 and 2 of this document have been published in peer-reviewed journals. Citations for these chapters are listed below and are included in the References section of this document. I have received permission to use those publications in this document from all co-authors: Rachel Beard, Dr. Philippe Lemey, Dr. Marc A. Suchard, and Dr. Matthew Scotch.

Chapter 1

Magee, D., Beard, R., Suchard, M. A., Lemey, P., & Scotch, M. (2015). Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion. Arch Virol, 160(1), 215-224. doi:10.1007/s00705-014-2262-5

Chapter 2

Magee, D., Suchard, M. A., & Scotch, M. (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. doi:10.1371/journal.pcbi.1005389