MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS

by

LAMBODHAR DAMODARAN

(Under the Direction of JUSTIN BAHL)

ABSTRACT

Human seasonal influenza viruses cause significant morbidity and mortality on a global scale. The constant evolution and seasonal epidemic transmission of these viruses has allowed the virus to continually evade human population immunity and produce novel strains. To develop effective preventative measures, it is critical to leverage available data and algorithmic frameworks to accurately characterize and track virus evolution. This body of work describes methods that aim to robustly characterize influenza virus evolution on the national and international scale. To perform this characterization each chapter interrogates and introduces different statistical methodologies for the study of influenza. In the first aim of this thesis, phylogeographic methods were employed to study the nature of viral diffusion in the United States. A data informed geographic partitioning schema was used to developed and leverages in jointly estimated discrete trait diffusion models over a decade of H3N2 influenza transmission. This work identified major geographic sources and sinks for influenza across seasonal epidemics and identified important predictors for the transmission process. The second aim of this work combines Bayesian phylogenetic methods with antigenic cartographies inferred using inhibition assay data in a generalized additive model to study the influence of different predictors, such as climate and demographic information, on the evolutionary landscape for multiple seasonal influenza viruses (H3N2, H1N1, B-Yamagata, and B-Victoria). This work introduces a novel methodology for using phylogenetic metrics as a response variable to study the partial effects of antigenic space on virus evolution. In the third and final aim of this work antigenic cartographies for H3N2 viruses are compared to assess the assays they are derived from. Antigenic cartography methods use log transformed titer data, traditionally hemagglutinin inhibition assay titer data, to study the differences between isolates of influenza. Newer neutralization-based assays have been introduced to address recent changes in HI assay sensitivity. This chapter compares the ordinations made by different assays as well as with phylogenetic distances and shows a stronger correlation between neutralization-based assays as well as low correlation with phylogenetic history across assays. Overall, this thesis provides a framework to make actionable inferences about influenza transmission and evolution.

INDEX WORDS: Molecular Epidemiology, Phylogenetics, Phylodynamics, Genomics, Influenza

MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS

by

LAMBODHAR DAMODARAN

B.S. Biology, University of North Carolina at Charlotte, 2018

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2023

MOLECULAR EPIDEMIOLOGY OF CONTEMPORARY SEASONAL INFLUENZA EPIDEMICS

by

LAMBODHAR DAMODARAN

Major Professor:	Justin Bahl
Committee:	Andrew Park
	Liliana Salvador
	Stephan Mark Tompkins

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2023

DEDICATION

This work is dedicated to my wife Erica Cherian. The one who knows me better than anyone, and my rock in the storm. You are my best friend, and I am continually made a better person because you. I love you.

ACKNOWLEDGEMENTS

It takes a village to make every person and I am so fortunate to have so many amazing people in my life that I have been supported by and learned from. I would like to thank my advisor Justin Bahl for giving me so many opportunities to learn and grow as a scientist and giving me the freedom to develop my own ideas. I would like to thank my committee members Liliana Salvador, Andrew Park, and Mark Tompkins for all the guidance and support you have provided me during my Ph.D.

I would like to thank my wife Erica Cherian for her constant support and love. I would like to thank my family. Thank you to my mother and father, Subangi Ragupathy and T.V. Damodaran, for supporting me in my decisions and for all your sacrifices that you have made for my success and wellbeing. Thank you to my brother Pranav Damodaran for being a constant companion through everything growing up. Thank you to my mother-in-law and father-in-law, Mercy Paulose and Cherian Varghese, for your constant support and encouragement. Thank you to my brothers-in-law, Andrew Cherian and Christopher Cherian, for your warmth and the love that you give.

I would like to thank my closest friends in the world in the order of appearance in my life: William T. Fowler, Darvlyn R. McLean, and Matthew J. Gibson. You all have given me support throughout my life that is hard to put into words. You all know how much I love you and how much you mean to me.

I would like to thank the members of the Bahl lab for your great support over these years. I especially would like to thank Jiani Chen, starting in the lab at same time allowed us to help each other in our goals and learning and I am forever grateful to have such a great friend and colleague during this process, it has been a privilege being able to work with such a talented person. I

additionally want to thank Swan Tan and Cody Aaron Dailey, two amazing friends that I had the honor of working with as my go to experts on immunology and epidemiology, I could not have gotten this far without your help. I would also like to thank the members of the Salvador lab, Noah Legall and Rujie Xu for your friendship and camaraderie during all of this. Thank you to Venkata Duvvuri for providing me with great advice and guidance throughout my research. I would like to thank Rebecca Garten Kondor at the CDC for her time, guidance, and mentorship.

I would like to thank Daniel Janies for giving me the opportunity to start doing research as an undergrad and giving me your time to teach me and allow me to become an active participant in the scientific process. Additionally, I would like to thank Adriano Schneider and Tamar Carter for their time and guidance during my time as an undergraduate researcher and afterwards.

I would like to thank April Mosley and Sandra Getz for your tireless work to keep things running in the IOB office, I am so appreciative for all your help over the years.

Thanks to David Preissman for being a friendly face and for all the great conversation and friendship over the years.

Thank you to every teacher, mentor, and friend that I have had. I am who I am because of the time that you spent and your efforts to make me a better and intelligent person. I will always strive to take the lessons you have given me and make the world a better place.

Finally, I would like to thank my dogs Rey and Riot. The ones who have literally been at my side as I worked every day and provided me a level of support that is hard to put into words. If you are reading this and considering a furry companion, please consider adopting, it was one of the best decisions I have ever made.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Seasonal influenza viruses are respiratory viruses that are implicated in annual global epidemics in humans which cause significant morbidity and mortality. Globally between 5-15% of the human population is infected with seasonal infauna viruses resulting in ~500 thousand annual deaths (1). Seasonal influenza epidemics are responsible for 9.2 - 35.6 million infections annually in the United States (2). Influenza viruses belong to the family *Orthomyxoviridae* and have a segmented single stranded negative sense RNA genome, and have four types, influenza A-D (3). Human seasonal influenzas are comprised of two influenza A viruses (H3N2 and H1N1) and two influenza B viruses (B-Yamagata and B-Victoria). These viruses cause several symptoms including but not limited to fever, cough, sore throat, and fatigue (4). The severity of infections is more pronounced in younger and older individuals as well as individuals with compromised immune function (5)(6)(7).

At the cellular level, infection with influenza viruses is primarily facilitated by the Hemagglutinin (HA) and Neuraminidase (NA) surface proteins (8). The HA surface protein is responsible for cellular attachment where the receptor binding domain of the HA protein attaches to the sialic acid terminating surface receptors of the host cell to facilitate cellular attachment and invasion via membrane fusion endosomes. The NA surface protein allows for the release of newly formed virions via the cleavage of sialic acid/HA protein bonds (9). The process of antigenic drift allows influenza viruses to evolve and evade population immunity and vaccine coverage through the gradual accumulation of genetic changes over time in gene segments of the surface proteins (10).

Evidence shows differential rates of evolution in the head region of HA protein as opposed to the stalk domain (11). In addition to antigenic drift, the influenza viruses segmented genome allow a process called reassortment to occur resulting in "antigenic shift". Reassortment occurs when two or more different virions with differing gene segments co-infect a given cell where their gene segments can mix, this results in the creation of hybrid virions containing different combinations of gene segments. This diversity in reassortant viruses allows them to evade previously developed population immunity causes zoonoses from avian/swine hosts to humans (12). Human seasonal influenzas historically have origins in avian reservoirs of waterfowl where the process of reassortment is common (13). The zoonoses (infection jumping from a nonhuman animal to human) of these avian influenzas viruses has been observed to be facilitated by human reared swine populations which act as major reassortment vessels, the 2009 swine flu pandemic was caused by a reassorted H1N1 strain (called H1N1 pdm) which later became the strain that replaced the previously predominate H1N1 seasonal influenza (14).

The emergence of novel strains of influenza virus by antigenic drift can result in increased transmission in a naive population (15). A measure of the transmission potential of a pathogen is the basic reproductive number R_0 which is defined as the average number of secondary cases caused by a given infection in a susceptible population (16). The reproductive number for seasonal influenza has varied over time with a median value of 1.28, and major pandemics (1918, 1957,1968, and 2009 pandemics) having median values between 1.46 and 1.8 (17). Viral sequence data is a key tool in understanding the evolution of influenza viruses where the rate of evolution over time can be estimated sequenced data that has associated temporal metadata. The intensity of sampling and genomic sequencing for seasonal influenzas viruses has intensified after the 2009 swine flu pandemic, spurred by the adoption of major data sharing platforms such as NCBI and

GISAID for real-time genomic data sharing during epidemics (18)(19). The availability of sequence data for influenza viruses has grown as sequencing technologies have become cheaper and surveillance apparatuses were established. An example of the breadth of the sequence data that was generated is seen in the number of H3N2 genomes submitted to the sequence data sharing platform GISAID (Figure 1.1).

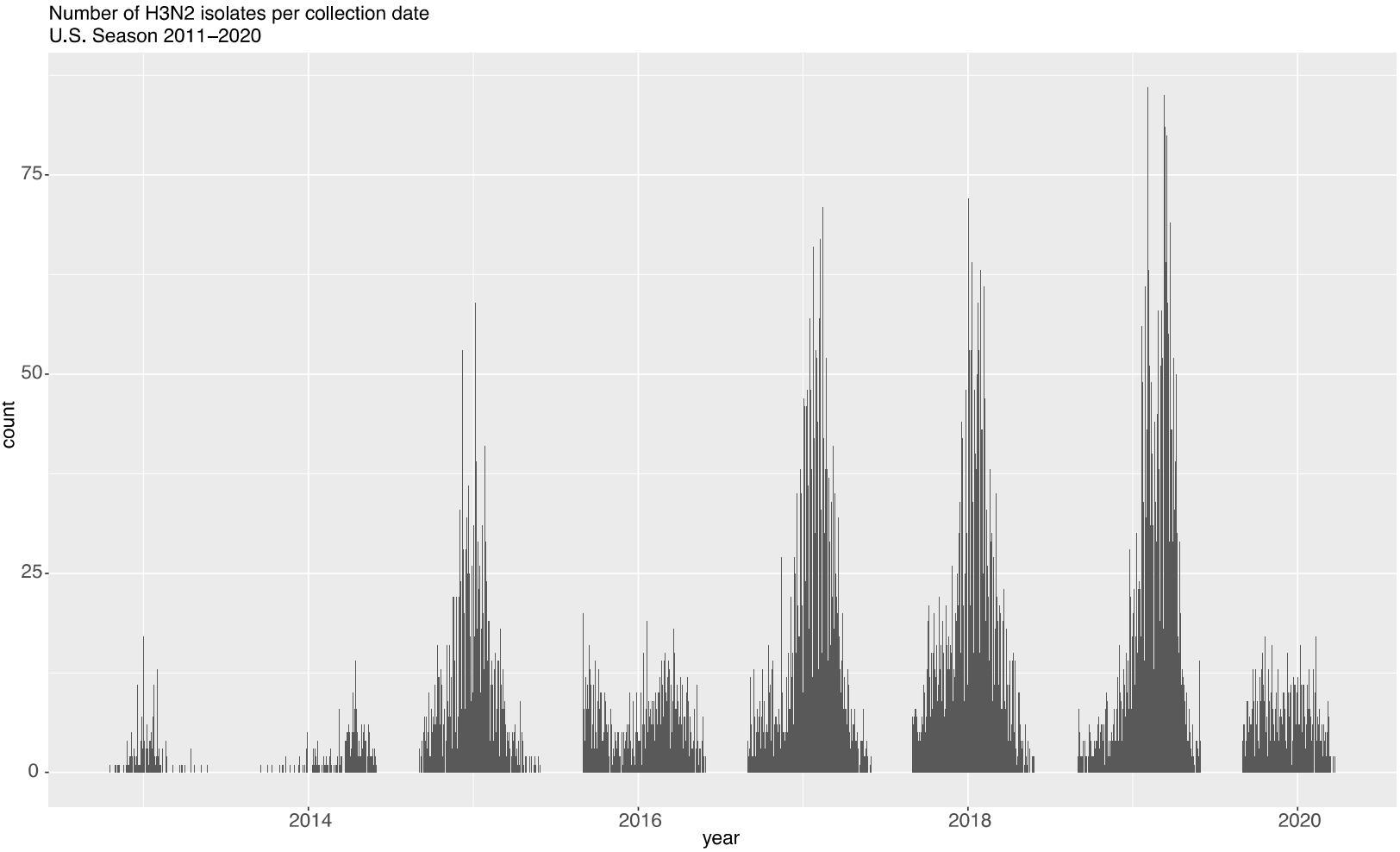


Figure 1.1. Number of H3N2 genomes submitted to GISAID by collection date for isolates collected in the United States during epidemic seasons.

The amount of sequence data generated daily during the epidemic season (typically between September and April of the following year) has increased considerably after the 2009 pandemic.

The importance and power of molecular surveillance became extremely apparent as genomic data began to be used to study both on finer and wider scales how viruses evolve over time and geographic space.

When the collection of sequence data is further broken down by the constituent lineage (or clade) of the H3N2 subtype, the importance of the breadth of sampling becomes more apparent were circulating diversity is not restricted to a single lineage in each season (Figure 1.2).

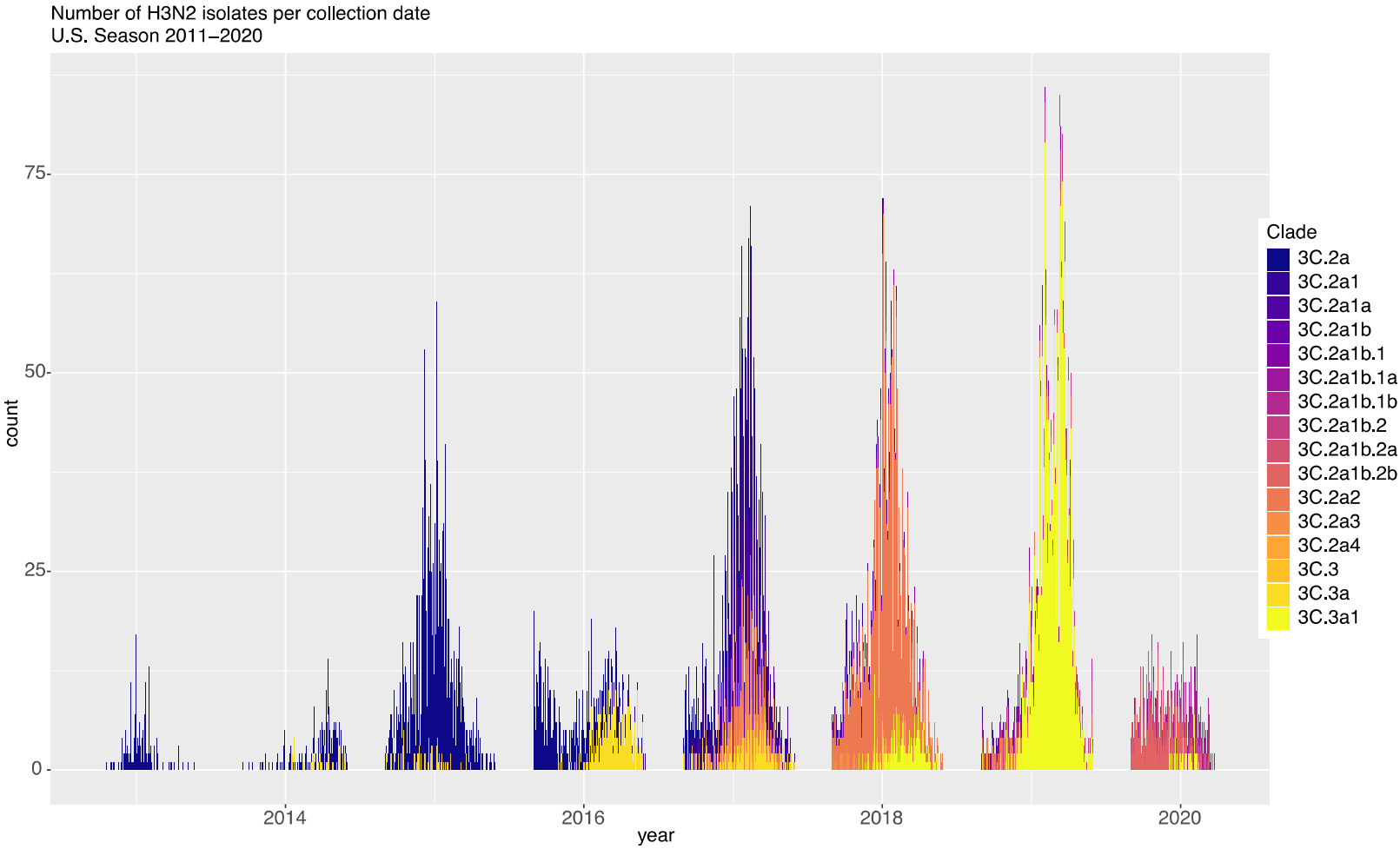


Figure 1.2. Number of H3N2 genomes submitted to GISAID by collection date for isolates collected in the United States during epidemic seasons, colored by the constituent clade of the isolate.

When attempting to perform phylodynamic analyses, responsible sampling strategies are critical in making meaningful and unbiased inferences. The amount of sequence data has grown over time, but the sampling is not uniform across the U.S. (Figure 1.3).

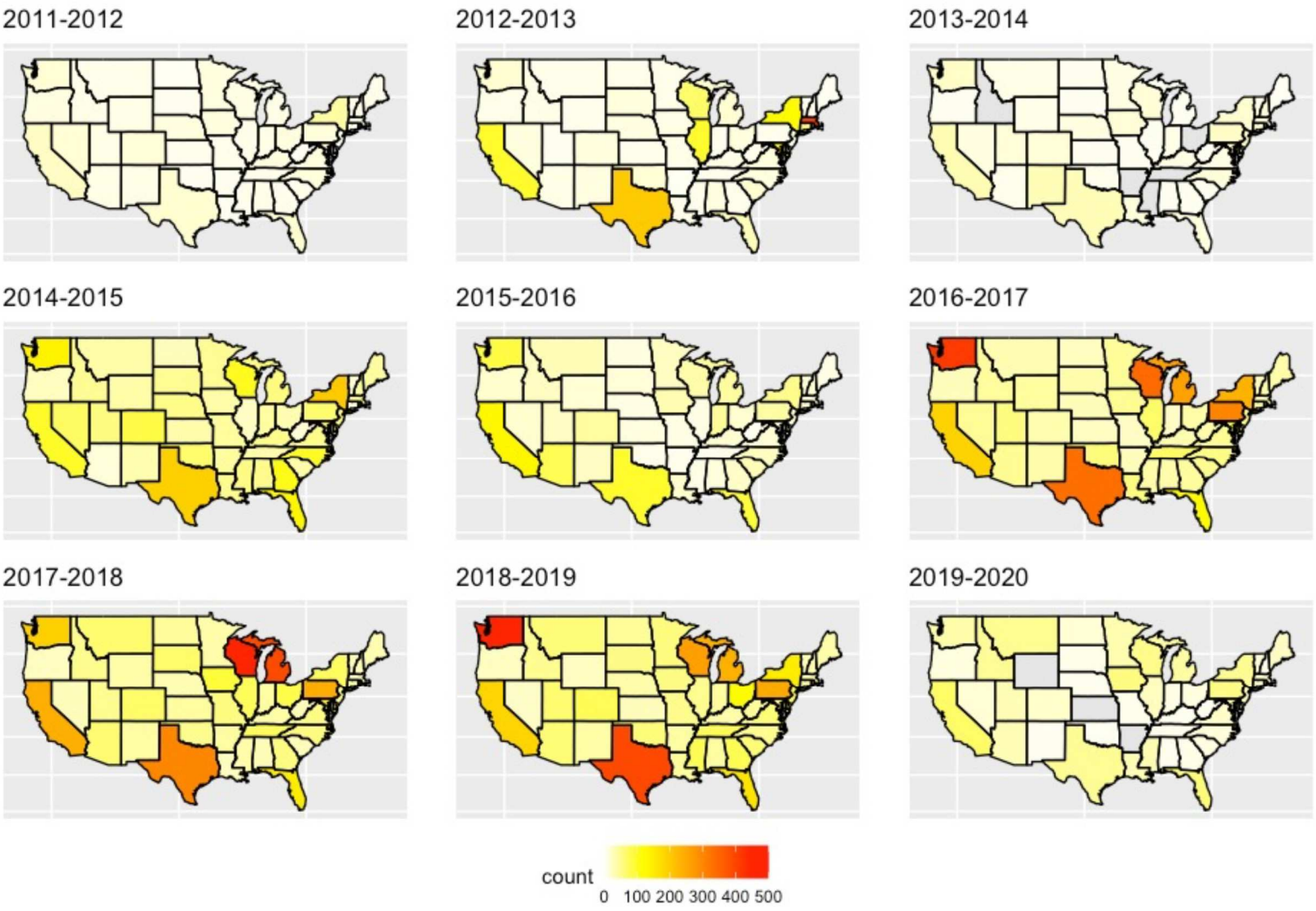


Figure 1.3. Number of available sequences by U.S. State by epidemic season between 2011 and 2020.

Despite the greater magnitude of sampling over time, it is critical that appropriate subsampling methods are utilized to create uniform datasets with less geographic bias. In this time the collection of sequence-paired antigenic data, as well as environmental data is more widely spread and greater

in intensity than in the previous decades which has allowed for greater breadth of analysis to be performed. The Right Size sampling methodology proposed by the Centers for Disease Control and Prevention (CDC) and Association of Public Health Laboratories (APHL) attempts to address sampling biases by creating guidelines for the genomic and antigenic sampling of strains and their reporting to both the CDC and sharing in public repositories (20).

The heterochronous sampling of viruses allows for the estimation of an evolutionary rate and allows for selection of an appropriate molecular clock model which describes the rate of evolution across the diversity of a phylogeny (21). The calculation of evolutionary rate, in terms of nucleotide substitutions per site and time, can be used as a proxy of antigenic drift within the population. Additionally antigenic data in the form of assay data from Hemagglutinin Inhibition (HI) assays and Focus Reduction assays (FRA) can be used to create antigenic cartographies where the change in antigenic units over time in the cartography is used as a measure of antigenic drift (22)(23). These laboratory assays allow for a measure of antibody reactivity to given isolates and transforms the value associated with the level of reactivity, the titer, into a measure of distance that can be used to estimate the Euclidean distances between isolates.

The term “Phylodynamics” was coined by Grenfell et al. and refers to the use of phylogenetic history to infer the effects of population immunity and study the transmission dynamics of pathogens (24). Phylodynamic methods utilize a Bayesian approach to studying the evolutionary history of pathogens by allowing for the integration of different models of population evolution (Coalescent models), evolutionary clock models, and nucleotide substation models (21)(25)(26). The primary programmatic framework to carry out Bayesian phylodynamic analyses globally is the BEAST suite of programs. The BEAST program, which stands for Bayesian Evolutionary

Analysis Sampling Trees, has two distinct versions BEAST 1 and BEAST 2 which are unrelated in development which offer different tools and packages to augment phylodynamic analyses (27)(28).

There has been a wide range of analysis that have aimed to use phylodynamic frameworks to study the evolution and transmission of influenza viruses. Previous work has attempted to correlate antigenic drift with the transmission using several different methods. Antigenic drift has been compared to weekly incidence purely though the use of Influenza-Like Illness data (ILI) where drift was estimated through the use of time-series data to estimate a measure of infective pressure, it was found that antigenic drift was non-uniform between years and that certain years showed substantially increased infection pressure due to new strains (29). The epochal nature of influenza virus evolution, where major phenotypic changers occur in a step-wise manner was explored where the genetic diversity within a given period was compared to the overall infectivity and it was found that the typical boom-and-bust nature of influenza epidemics occurs during a period when antigenic diversity grows and episodic reduction in diversity during antigenic cluster transitions (30). Other phylodynamic studies have also focused on the circulation of the seasonal influenza globally and found that that when global circulation estimated as the discrete trait transition rates between major geographic regions, was low, there was a coincident reduction in antigenic evolution (estimated calculating several measures of strain persistence) for A/H1N1, B-Yamagata, and B-Victoria viruses (31). The nature of major source populations for novel strains of influenza has been another major focus of phylodynamic analysis which have shown that shifting meta-populations between major geographic locales might act as seeding regions for global epidemics (32). The evolutionary rate which was estimated as the nucleotide substitution rate was compared to the R_0 which was calculated from the initial exponential growth rate estimated using parametric

growth models (33). Phylodynamic analyses have allowed for inferences about the size of pathogen populations through the estimation of the effective population size (Ne) and the growth rate from coalescent-based phylodynamic analysis, these estimates can be treated as proxies number of infections and colonization of a given pathogen in a population (34).

The objective of the work contained in this thesis is to understand the diffusion and evolution of seasonal influenza viruses across demographically, climatologically, and socially heterogeneous landscapes through Bayesian phylodynamic methods. Additionally, the evaluation of existing methods for the characterization of seasonal influenza molecular epidemiology and the proposal of novel methods is a major focus of this body of work. Seasonal influenza evolution is primarily marked by gradual antigenic drift in which a population gains immunity to introduced viruses over time. The United States is a country with many large metropolitan regions which host major hubs for international and regional travel which facilitate the introduction and rapid spread of n novel virus strains into vulnerable populations (35)(36). These regions have large populations and are typically densely populated. Previous characterizations of transmission history in the United States used phylogenetic analyses inferred from sequence data and antigenic characterization from Hemagglutination Inhibition (HI) assay data, on a global scale across many seasons (31). The pairing of genetic and antigenic methods has allowed for a more accurate estimation of transmission patterns between major demographic regions and can aid in determining regions that experience significant evolutionary pressures. These analyses can be further augmented through the inclusion of other data such as case numbers, transportation statistics, and climate data. The first experimental chapter of this body of work studies the effects population demographic factors and region of spread within the United States and how this impacts the evolution and transmission of influenza viruses. The second experimental chapter of this body of work characterizes the

evolution of seasonal influenzas in the context of the interruption of typical seasonal patterns due to the SARS-Cov-2 pandemic and aims to integrate phylogenetic data with antigenic data to characterize recent influenza evolution. The third experimental chapter compares different laboratory assays utilized to antigenically characterize influenzas viruses and compare them both to each other and phylogenetic history.

Chapter 2 of this dissertation takes a data informed approach to partitioning phylogeographic space to allow for the better characterization of the evolution of H3N2 influenza and the effects of major co-variates on the diffusion process. Quantifying the importance of different external factors on virus evolution is critical when trying to understand the transmission of important human pathogens like seasonal influenza viruses. Characterizing the evolutionary features of virus transmission can give measures of strain diversification and evolutionary pressures within a given geographic range and timeframe. Additionally, understanding the transmission of viruses across a large geographic range and determining key transmission regions is important in control efforts. In this chapter the geographic diffusion of influenza viruses within the United States is elucidated using discrete trait diffusion models. These discrete trait models are used to describe the geographic dispersal of viruses as a study of the phylogeography; the diffusion between geographic regions will be analyzed and the most significant transition rates can be described using the Bayesian Stochastic Search Variable Selection (BSSVS) (37). An important aspect of this study is the development of meaningful geographic regions to describe the diffusion process. Regions such as those created by the U.S. Census Bureau provide a geographic breakdown that is not necessarily data-driven and can be arbitrarily drawn (38). A regionalization schema for geographic discrete traits was developed, informed by state level influenza like illness data which was used in the creation of adjacency matrices for community detection algorithms. This community detection

approach can create regions that more accurately describe a geographic space with variable populations densities. The diffusion process for influenza viruses between different geographic regions can be affected by different environmental factors, the different metadata collected during the period of virus circulation was analyzed as evolutionary covariates through the Generalized Linear Model (GLM) (39) . The GLM allows us to determine the impact of a specific type of data i.e., mean temperature, population size, road coverage, and calculate a level of support for that data's inclusion in describing the diffusion process. In addition to these data types, antigenic data in the form of HI assay data provided from the Centers for Disease Control (CDC) for assays performed before 2020 was used to create profiles of reactivity to vaccines across the study space through antigenic cartography. The working hypothesis of this work states that covariates relating to climate, population demography, and the antigenic profile of circulating viruses are the most consistent drivers of transmission and evolution across seasons. This study identified the major drivers of evolution in influenza as climate and transportation factors, this has implications for future surveillance and resources towards important evolutionary factors and significant geographic locales of infection.

The second experimental chapter, Chapter 3, explored the molecular epidemiology of seasonal influenza viruses for a contemporary dataset of influenza isolates for each seasonal influenza subtype between 2017-2022. Recent transmission of seasonal influenzas has been interrupted by the ongoing COVID-19 pandemic and understanding how this might have affected the evolution and transmission of seasonal influenzas viruses is important in making future determinations about public health interventions. Characterizing the evolutionary space for each subtype of seasonal influenza is important to determine whether influenza evolution was fundamentally affected by the introduction of another major respiratory pathogen. The availability of paired antigenic and

sequence data provides valuable information for the characterization of this evolutionary space and allow for the implementation of models to study the impact of different predictors on the evolutionary process. The generalized additive model (GAM) was used in this chapter to study antigenic evolutionary space (in the form of antigenic cartographies) as a function of phylogenetic distance for each isolate. The GAM offers a method to integrate metadata and estimate the effects of that metadata on the overall evolutionary rate. The working hypothesis of this chapter posits that the interruption of typical seasonal influenza epidemic patterns due to the COVID-19 pandemic has led to marked reduction in evolutionary rate and diversity in seasonal influenzas. The findings of this study illustrate that influenzas rates of evolution remained constant despite the disruption in the intensity of seasonal influenzas epidemics. Furthermore, this study introduced a new data integration method, via GAMs, that allows for the integration of phylogenetic history and antigenic cartography with secondary metadata to highlight major drivers of evolution and provides a novel framework for studying the evolution of influenza viruses using paired genomic and antigenic data.

The third and final experimental chapter of this thesis, Chapter 4, examines the different antigenic assays and their resulting antigenic cartographies for H3N2 seasonal influenza. Different neutralization based antigenic assay, FRA and HINT have been introduced characterize the antigenic characteristics of circulating H3N2 influenza viruses and the antigenic distance between isolates can be estimated using antigenic cartography methods, as is done for HI data. These neutralization assays were introduced due to the reduction in sensitivity of HI assay (40). The relative sensitivity and ability to discern major changes in antigenic space for each assay is important to determine to make better decisions regarding vaccine strain selection. The comparison of the antigenic cartographies and how ordinations differ between assays via the Procrustes

analysis allows for a quantification of differences and examine the importance of methodology when studying antigenic space. Additionally, the phylogenetic history of isolates that is paired with these antigenic cartographies is used to study the correlation of phylogenetic history with the different estimated cartographies. The working hypothesis of Chapter 4 posits that antigenic cartographies of HI data will be less correlated with ordinations built for FRA and HINT assay data. The results of the analysis show that the HI assay is less correlated with the neutralization-based assays, with the neutralization assays are better correlated to each other. Additionally, each of the antigenic assays were poorly correlated with phylogenetic distance, with neutralization assays showing the greatest overall correlation with phylogenetic history.

This body of work presents important contributions to the fields of phylogeography and molecular epidemiology and ultimately will help to augment public health response to these major human pathogens.

CHAPTER 2: ANALYSIS OF SEASONAL H3N2 INFLUENZA DIFFUSION IN THE UNITED STATES, 2011-2020

Damodaran, Lambodhar et al. To be submitted to PLOS computational Biology

Abstract

Seasonal influenza virus epidemics in the United States cause significant morbidity and mortality every year. One of the pathogens which make up the typical influenza season is the H3N2 influenza A virus which typically co-circulates with the H1N1, B-Yamagata, and B-Victoria influenza virus sub-types. An important aspect of understanding influenza evolution is the characterization of diffusion patterns across large geographic space to understand which locales need significant attention and resources for mitigation strategies. In our study, we employ a data-driven geographic partitioning schema to study the phylogeographic diffusion patterns of H3N2 seasonal influenza viruses across the United States in the past decade. We integrate genomic sequence data and associated metadata and antigenic assay data to characterize the transmission patterns across the United States between epidemic seasons. We used influenza-like-illness data to create a transmission network across the United States and then employed community detection algorithms to create the optimal regional schema for use in a jointly estimated discrete trait diffusion model across multiple seasons. The regions defined through community detection algorithms differed from administratively defined regions such as the U.S. Health and Human Services and U.S. Census regions. Our model shows a significant diffusion activity from regions in the Western and Middle Atlantic United States to other sink regions. Furthermore, we aimed to study the influence of different co-variates on the diffusion process. We organized antigenic, climate, demographic, and transportation metadata based on the regional schema developed and employed a Generalized Linear Model to study the co-variate influence on influenza diffusion. We found that antigenic distance to vaccine candidates for circulating strains, climate incidences of temperature, population density at the origin region, and airline passenger numbers were positively correlated with the diffusion process.

Introduction

Seasonal influenza virus epidemics in the United States cause 9.2-35.6 million infections and 140,000 - 710,000 hospitalizations each year (2). Influenza viruses are single-stranded negative sense RNA viruses belonging to the family *Orthomyxoviridae* with three major types that cause respiratory disease in humans; types A and B are responsible for seasonal epidemics while type C is less prevalent in human populations (41). Influenza seasons are characterized by the co-circulation of two influenza A virus sub-types (H3N2 and H1N1pdm) and two influenza B virus sub-types (B-Yamagata and B-Victoria). A major factor in the evolution of influenza viruses is the gradual accumulation of mutations over time leading to the emergence of new lineages in a process called antigenic drift (42)(31)(43). These gradual changes not only allow the pathogens to continually burden the public health but also provide an opportunity for the spatiotemporal characterization of transmission among populations using molecular epidemiology approaches (44).

Phylogenetic models can be used to study the transmission and evolutionary dynamics of a pathogen across space and time by assessing the gradual changes in pathogen genomes. These changes between pathogens are used to estimate the relatedness between collected virus isolates and can divulge larger-scale immunological and epidemiological characteristics, the use of these phylogenetic methods and the addition of different mathematical models has broadly been described as “Phylodynamics” (24)(45). Data for virus isolates, such as the time and place of isolation, can be used to augment models based solely on molecular sequence data to characterize the diffusion of the virus more accurately between different geographic locations by modeling them as discrete traits across a phylogenetic history (46)(47)(48). Additionally, when implementing geographic discrete trait diffusion models, data on the place of isolation are used to

create phylogeographic discrete trait models and the impact of co-variates on the diffusion process can be investigated (37)(49). Modeling the diffusion process using discrete traits allows the estimation of rates of transition from any given state to another. In this context, transitions between locations as discrete traits correspond to the geographic diffusion of virus strains, a cross-scaling inference of virus transmission (48). Past characterizations of population dynamics in H3N2 influenza viruses indicate fluid population dynamics where shifting meta-populations are responsible for the emergence of seasonal epidemics (32). Other characterizations of H3N2 population dynamics indicate a major source region that acts to seed epidemics in the other areas seasonally (50)(51). The majority of phylogeographic studies typically aggregate study space into administrative levels. These levels can vary from smaller scales such as the city/town and can be larger at the state/country level (52)(53). Population structure can have profound effects on the transmission of pathogens and their epidemic behavior. Previous work has shown that urban centers and areas of high population density are positively correlated with incidence of influenza viruses (54). A major limitation of using administrative boundaries is that population structure and travel within a region or between regions might be obscured. An example of this lost signal would be highly mobile populations which travel for work across administrative borders (55). Previous studies in public health and livestock health have sought to create "regionalizations" using spatial models that consider different data to capture regional variation and identify major regions of increased incidence for a given disease, but these have not been applied to phylogenetic discrete trait models (55)(56)(57)(58). Taking a data-informed approach to the regionalization of geographic transmission space can potentially create more meaningful regions that account for differences in regional characteristics such as population density and connectedness.

In this study, incidence data in the form of Influenza-Like Illness (ILI) count was used to create a data-informed regionalization schema for phylogeographic discrete trait diffusion modelling and the characterization of co-variate influence on the diffusion process of H3N2 influenza virus in the United States. We estimated a case data-informed regional structure across a decade of influenza transmission, which partitioned the United States into regions. The results of this study suggest that a data-informed region schema provides granularity to the source-sink dynamics described via the discrete trait diffusion process. The regions determined were then used to estimate the influence of different data types on the diffusion process, indicating the strongly supported importance of population demographic structure and climate.

Materials and Methods

Sequence data collection and primary phylogenetic analysis

Sequence data sets for the hemagglutinin segment and associated metadata was downloaded from the public repository GISAID for H3N2 viruses for each season in the United States between May 2011 and April 2020 (67). The timing of the season was determined using case data observing the initial rise and fall of the epidemic (generally between September of the start year and May of the following year). Multiple sequence alignment was performed for each season data set using MAFFT v7.453, and visual inspection of alignments was conducted using Geneious v 9.0.5 (68)(69).

Initial phylogenetic trees were created using the maximum likelihood program IQTREE v 1.6.12 for each season data set using all available sequence data (70). The maximum likelihood trees were then used to sub-sample the sequence data using the Phylogenetic Diversity Analyzer (PDA) program (v.1.0.3) to create data sets for use in Bayesian phylodynamic analysis by creating subsets

which preserve clade diversity across the phylogenetic tree (71). The tree construction and temporal outlier detection process was performed for each season subset to identify and remove outliers. The temporal outliers and the "clockliness" of each of the season datasets were analyzed through root-to-tip regression using TempEst v 1.5.3 (46).

Antigenic Cartography

HI assay titer data for H3N2 influenza isolates collected in the United States between 2011 and 2020 was provided from the Centers for Disease Control. The lineage of the isolates analyzed with paired antigenic data was determined using the Nextclade program (72). The "Racmacs" R package was used for the calculation of antigenic cartographies. The Euclidean distance between vaccine candidates and antigens within the cartography were calculated with an in-house python script. The Euclidean distances to vaccine candidates was further used in and implementation of the generalized linear model described below.

Multiple season phylogenetic tree estimation

To reproduce and confirm the seasonal epidemic behavior of H3N2 in the United States an initial Bayesian phylogenetic analysis was carried out to estimate the effective population size over time via the Skyride coalescent. All following Bayesian phylogenetic reconstructions and analyses were performed using BEAST v 1.10.4 (27). To create a multi-season phylogeny, a sub-sample of 150 sequences were taken for each season using the PDA as well as two independent random samples of 150 sequences each (Table S4). Initial phylogenetic reconstruction was performed for this multi-season dataset using IQTREE v 1.6.12, and temporal signal and outliers were diagnosed using TempEst v 1.5.3. Bayesian phylogenetic reconstruction was performed using the SRD06 codon partitioning model, the GMRF skyride coalescent, and a lognormal relaxed clock (73)(74)(75). Six

independent Markov Chain Monte Carlo (MCMC) runs were performed with a chain length of 100 million states, sampling every 10,000 states. The results of these runs were visualized in Tracer v1.7.2 and runs with inadequate effective sampling score (ESS) were discarded; an ESS above 200 is generally accepted as an adequate score and indicates proper mixing in the Bayesian algorithm (76). The three best runs were selected, which maximized the combined ESS score. The runs were combined using LogCombiner v1.10.4 for both log and tree files, 10% burn-in was removed, and the runs were re-sampled at a frequency of 30,000 states to produce a posterior set of 9,000 trees and a log file with the corresponding 9000 states logged. The maximum clade credibility (MCC) tree was estimated using the posterior set of trees though the program TreeAnnotator v1.10.4, using annotation cutoff for nodes with at least 95% posterior support. These files were used in Tracer v1.10.4 to estimate the effective population size over time using the GMRF Skyride reconstruction tool.

Empirical tree set estimation

Empirical tree sets of 500 posterior trees were created for further analysis of discrete trait diffusion. These empirical tree sets were created for each season data set, and identical sequences were removed, retaining one representative sequence. Bayesian phylogenetic reconstruction was performed using BEAST v1.10.4, a chain length of 100 million states sampling every 10000 states. The SRD06 codon partition model was chosen as the nucleotide substitution and codon partitioning model. The Gaussian Markov Random Field (GMRF) skyride coalescent was used as the population model. Following temporal analysis of the ML trees, an uncorrelated relaxed clock model was used for the temporal model. The rate of nucleotide substitution observed in primary Bayesian phylogenetic analysis of the multi-seasons dynamics was used to set prior rate of 10^-3 substitutions/per site/per year for seasonal influenza, therefore a ucld.mean prior of 0.0033 was

used. Six runs were performed, and the four best runs with appropriate parameter estimations and adequate ESS were chosen to create an empirical set of 9000 trees by re-sampling at 40000 states and removing the burning of 10 million states from each run. The empirical tree set of 9000 trees was down sampled to 500 posterior trees for use in further phylogenetic analysis.

Regionalization using incidence data

Data for the ILI activity at the state level in the United States were downloaded from ILInet. Due to reporting policies at the state level, ILI data for the state of Florida was requested directly from the Florida Department of Public Health.

We used the ILI incidence data to explore clustering using spatial scan statistics with SaTScan™ v 9.6 (77). The data were fit to each week's estimates separately as to only assess the spatial autocorrelation. The scan statistics were estimated assuming a Poisson distribution probability model with rates consisting of ILI cases per total patients seen in each week. We used the clusters detected by this algorithm to generate an adjacency matrix representing the pairwise frequencies of co-occurrence within clusters. Using the adjacency matrix to represent the regional network of the states, we then assessed the modularity of the network and implemented the Louvain community detection algorithm (78). Communities, i.e., possible divisions within the network, were iteratively created by combining adjacent nodes and assessing the resulting effects on the network modularity. The community detection algorithm was used twice in a hierarchical fashion; that is, we first implemented the Louvain algorithm on the entire network, and then, we re-implemented the algorithm on the detected communities separately to identify sub-communities. The communities that were detected using these methods were used in the phylogeographic analysis for discrete trait diffusion.

Newman and Girvan (2004) define the modularity of a weighted network as:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j)$$

where A_{ij} represents the weight of the edge between *i* and *j*, = $\sum_{ij} A_{ij}$ is the sum of the weights of the edges attached to the vertex *i*,c_i is the community to which vertex *i* is assigned, the δ-function $\delta(\mu,\nu)$ is 1 if $\mu = \nu$ and 0 otherwise, and m = $\frac{1}{2}\sum_{ij} A_{ij}$

The regions that were developed using the Louvain method were used in downstream phylodynamic analyses of discrete trait diffusion and will be referred to as the "Louvain" regional schema.

Discrete trait diffusion BSSVS

To study the viral exchange between geographic locations an asymmetric substitution model for discrete traits was employed using a non-reversible continuous time Markov chain for the regionalization schema defined below for every season in the study space.

The taxa were labeled according to a place of isolation metadata available with molecular sequence data from GISAID, this metadata was used to categorize sequence data into regions based on the data-informed regionalization schema described previously (Louvain schema) and two administrative regions schema. The administrative regions used in this study are the Health and Human Services (HHS) regions and the U.S. Census divisions (79)(38). Broader regions (< 6 distinct regions) were excluded from further analysis due to the lack of characteristic value.

To determine the most parsimonious diffusion network and reduce the number of rates to only those with significant non-zero transition rates, the BSSVS was utilized (37). The level of support

for each transition was calculated using the BF support, the program SpreaD3 v0.9.7 used to calculate this BF support (80). The relative significance and support of a given rate is determined by the value of the BF where the following ranges apply BF > 3 substantial support, BF > 10 strong support, BF > 30 strong support, and BF > 100 decisive support. The median transition rates and the Bayesian credible interval (BCI) were calculated from the non-zero actual rates of the BSSVS log files using the python module PyMC (81).

Generalized Linear Models

The GLM was used to determine the importance of a given co-variate on the diffusion dynamics between regions. The GLM was implemented using a jointly estimated diffusion matrix calculated via the joint estimation method described previously for the BSSVS. The following schema was used for each regional schema.

The joint matrix of across multiple seasons, λ_{ij} is defined:

$$\lambda_{ij} = \sum(A_1 + A_2 + \ldots A_n) \text{ / n}$$

Where A_n represents the number of distinct discrete trait diffusion matrices, one for each season in this analysis.

And the GLM is implemented as a function of this joint matrix:

$$\log(\lambda_{ij}) = \beta_1\delta_1\log(p1_{ij}) + \beta_2\delta_2\log(p2_{ij}) + \ldots \beta_n\delta_{np}\log(pn_{ij})$$

Where β represents the coefficient, δ represents the the binary inclusion/exclusion indicator value, and pn_{ij} represents the predictor. Here n represents the number of independent predictors analyzed in the GLM.

The following data was downloaded at the state level and was further organized into respective regions based on the regional schema. Transportation data was collected from the U.S. Department of Transportation's Bureau of Transportation Statistics. The Monthly Transportation Statistics data set was utilized (82). Population estimates were collected from the U.S. Census Bureau for each state using the "State Population Totals and Components of Change: 2010-2019" data set (83). The "nclimdiv" data set was utilized to download and organize climate metrics at the state level (using averages across the area of a given state) for the different regionalization schema used in this study. NARR data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site: https://psl.noaa.gov/cgi-bin/data/timeseries/timeseries1.pl. The ILI incidence data collected previously was used in this analysis as well and was organized respectively. The mean antigenic distance to vaccine candidates previously described were calculated for each regionalization schema, additionally the inter-region distance was calculated.

All data were filtered for the relevant time period of a given epidemic season for each season between 2011 and 2020. Each predictor was organized for a given influenza season and the average of the values across all seasons in the study was used in the joint estimation GLM for each predictor. The full list of predictors and their descriptions can be found in Supplemental Table S5.

Batch predictor data sets were created for each regionalization schema. The GLM was implemented in BEAST v 1.10.4 using an MCMC chain of 10 million states, logging every 1000 states. Three independent runs were performed and combined. The results were summarized and visualized using in-house scripts (provided in GitHub).

Investigation of model fit

To investigate model fit the marginal likelihood estimation was performed using the generalized stepping stone sampling method which combines the methodology of path sampling and stepping-stone sampling methodologies to yield reliable estimates of the log marginal likelihood (84)(85). The MLE was used to calculate the BF support for a given model compared to another model. The log marginal likelihood of two models can be compared using the framework described by Kass and Raftery (1995) where:

log(BF) = log(M1)-log(M2))

Additionally, to investigate the fit of the regionalization models in a phylogenetic context a Bayesian association of tip-states (BaTS) analysis was performed using the program Befi-BaTS v 0.1.1 (86).

All analytic scripts and BEAST XML files used for described analyses can be found at the following github repo: https://github.com/ldamodaran/US-H3N2-Diffusion-2011-2020

Results

Regionalization of the United States

The initial pass of the community detection algorithm used for ILI data collected between 2011 and 2020 resulted in 3 large clusters which, in a secondary iteration, were further subdivided to detect 7 sub-clusters (Figure 2.1, Figure S2.1).

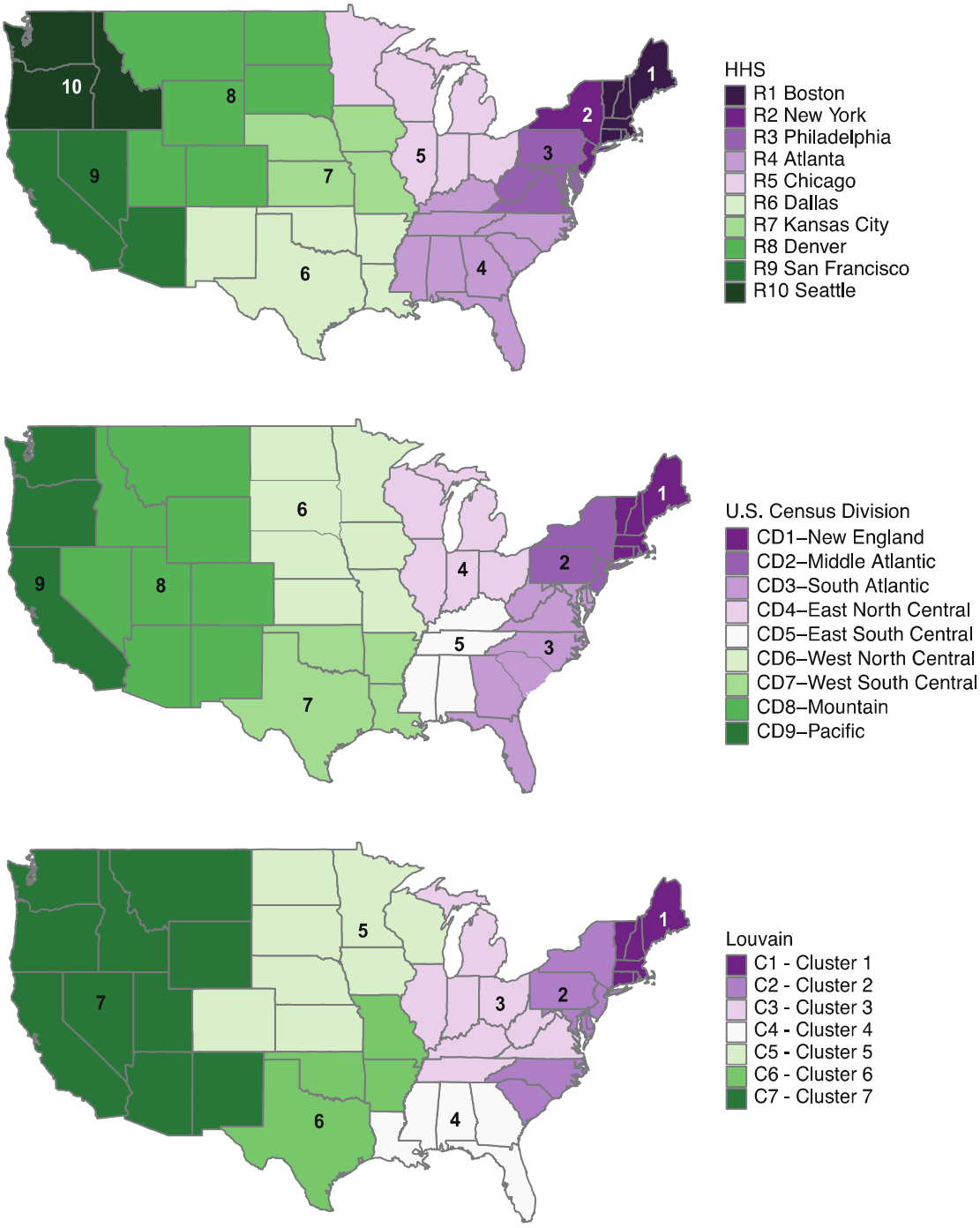


Figure 2.1. Map of the United States colored by geographic grouping schema. The maps for the HHS and U.S. Census Division show administrative regions created by the United States government. The Louvain map shows regions created using the Louvain community detection algorithm applied to incidence data for influenza.

These sub-clusters (C1-C7) are the regions of the Louvain community detection algorithm which will be referred to as the Louvain schema. All regions were continuous except region C2, which separates the mid-Atlantic states from North and South Carolina between the state of Virginia. This non-contiguous region might be the result of population movement differences across the eastern seaboard, where the DC metro area, which brings in commuters from several surrounding states, might act as an abnormal centroid of travel. Additionally, the existence of a major airport hub on the North Carolina/South Carolina border can act as a major point of connectivity to the middle Atlantic region, potentially causing the disjointed region. These regions show some similarities to the structure of the HHS and U.S. Census regions, but some regions show different subdivisions of states. For example, the western United States is divided into two and three regions for the U.S. Census and HHS regions, respectively, whereas in the Louvain regional schema the western United States is a contiguous large region comprised of 10 states.

When comparing the regional structure of data-informed regions to administrative regions, there are some similarities and differences depending on the administrative region being compared (Figure 2.1). In the broader data-informed Louvain region the large region C3, encompasses the western and upper-Midwestern regions which is divided in the U.S. Census regions. The U.S. Census regions divide the continental United States into four administrative regions, the region designated “South” in the Census regions is divided between states further south, which belong to the Louvain C1 region, and states which are part of Louvain C2, which are designated as the Northeast region. When comparing the different regionalization schema, the administrative regions have a smaller number of constituent states than the data-informed regions. This could be due to the relative sizes of populations between states where states with greater densities might

create better administrative regions while a regional schema formed based on infectious activity might have more regional fluidity due to travel.

Phylogenetics of H3N2 across a 10-year period

Phylogenetic reconstruction was performed for sub-samples of sequences from each epidemic season between 2011 and 2020. A Bayesian phylogenetic reconstruction was performed across the 10-year period. The "ladder-like" structure of influenza evolution where selection for novel strains is evident by extinction of clades over time and the emergence of new clades from the most immediately persisting clade (Figure 2.2).

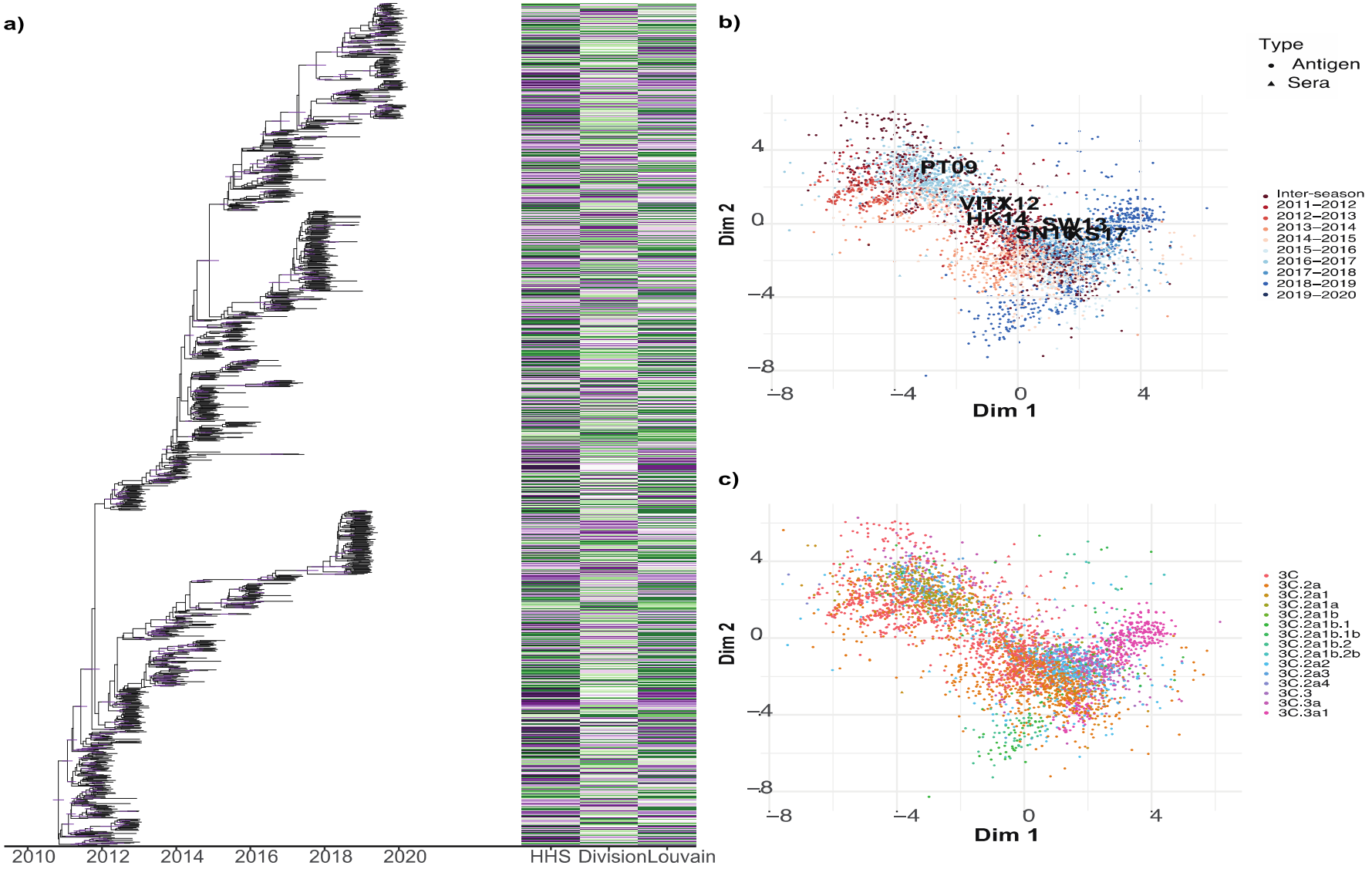


Figure 2.2. Genetic and antigenic analysis of H3N2 in the United States between 2011 and 2020 correlated with the geographic schema as defined in Figure 2.1. A) Maximum Clade Credibility tree of H3N2 isolates collected between 2011 and 2020 in the United States. The corresponding heatmap shows the geographic discrete trait associated for a given tip in the tree for three different geographic schema. Nodes with a posterior probability greater than 90% are annotated with a purple bar for the 90% BCI of the given node. B) Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States. Isolates are colored by season of collection and the major vaccine candidates for the study period were labeled. C) Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States with isolates colored by respective clade determined using paired nucleotide data with the Nextclade program.

The temporal signal was analyzed, a nucleotide substitution rate of 4.587E-3 sub/site/per was observed with a tMRCA of 2010.392. The GMRF skyride estimates of the mean effective population size over time were visualized and captured the seasonal epidemic pattern with oscillating waves of infection which occur in the northern hemisphere, these correspond to oscillations seen in when visualizing ILI data at the state level across the study period (Figure S2.2-3). The tips of the phylogenetic tree were colored to reflect the discrete character associated with a given taxa for each geographic schema. All clades were heterogeneous for a given discrete trait which illustrates the diffuse nature of the virus across the United States (Figure 2.2a).

In order to study the diffusion patterns across the 10-year period, the Markov jumps between states across the tree were analyzed for each geographic schema (Figure S2.3). The western U.S. showed the greatest source activity with the Louvain region C7, HHS region 9, and Census region 9 showing the greatest number of Markov jumps origination from these regions to other regions. Particularly, the major sink region for each regional schema seemed to be the Southern Atlantic regions of the United States. The Louvain regional schema partitions the Southern Atlantic region and middle Atlantic regions by including North and South Carolina as a part of the major region. This might contribute to the major sink indicated as region C2.

The marginal likelihood estimates were calculated for each geographical schema to determine statistical support for each discrete trait model. The Marginal Likelihood estimates (MLE) indicated that the Louvain geographic schema has slightly greater marginal likelihood support than the HHS and Census schema. The Louvain model had greater marginal likelihood support (loglik = -71246.42 BCI: [-94484.94, -56708.76]) than the HHS (loglik = -71712.17 BCI: [-94515.89,-57774.84]) and Census (loglik = -71342.15, BCI: [-94258.52, -57247.86]) models. The Bayes Factor (BF) support between the Louvain model and the HHS and U.S. Census models was 465.75

and 95.73 respectively. A log BF support > 100 is considered overwhelming support while a BF support > 20 is considered strong support (Kass and Raftery (1995))(59).

To further investigate the fit of the discrete traits to the phylogeny, the program BaTS was employed and showed a lower mean association index score (AI) and mean Fitch parsimony score (PS), 108.2 and 773.81 respectively compared to the HHS (AI: 119.97, PS: 888.55) and Census regions (AI: 117.82, PS: 862) (Table S1-S2.3). The maximum exclusive single state clade size, which describes the maximum frequency of a given discrete trait on clades consisting only of taxa with one discrete trait, showed that the Louvain method had the largest observed mean value (4.6) compared to the HHS (3.97) and Census regions (3.62). The larger mean value indicates that transmission events that were linked closely in time were also captured in the regional structure of the schema through the community detection algorithm.

The data-informed regional schema developed shows similar statistical support to the administrative regions used. Administrative regions are developed centering around major population centers and regions and encompass large areas of commerce which overlap with regions of transmissions determined using ILI data. This similarity in geographic space can potentially capture transmission routes across the community network in each area which allows for comparable phylogenetic reconstructions of regional schema.

Antigenic Cartography of H3N2 across 10-year period

Antigenic data in the form of Hemagglutinin inhibition (HI) assay titer values were used to create an antigenic cartography showing the antigenic distance between isolates (22). The resulting antigenic cartography shows the wide range of antigenic diversity occurring over the 10-year period (Figure 2.2b-c). When the isolates were visualized according to their geographic schema

there was no apparent geographic structuring for any of the regional schema (Figure S2.4-S2.6). Visual inspection of antigenic cartographies indicates temporal structuring when coloring isolates by epidemic season where isolates are grouped by season. The temporal structuring is most apparent when plotting each antigenic dimension by time and coloring isolates by season, inter-epidemic isolates demarcate the space between epidemic seasons (Figure S2.9-S2.10). When annotating the northern hemisphere vaccine candidate (WHO) positions on the antigenic cartography there is an antigenic space ~2 antigenic units wide where most of the vaccines are situated. A given antigenic unit corresponds to a 2x dilution of antisera in the HI assay, this indicates that the antigenic space that the vaccines are found is tightly constrained. The overall antigenic space is large with about 8 antigenic units in space in both dimensions indicating a wide antigenic diversity. This antigenic diversity can also be compared to the diversity of lineages as determined by paired sequence data. The Nextclade program was used to determine the clade that a given influenza strain belongs to, when visualized in the antigenic cartography there is a wide diversity of lineages that co-circulate within epidemic periods. The persistence of clades such as 3C.3a is evident when visualizing the antigenic dimensions against time (Figure S2.7-S2.8). The wide distance of circulating strains to the vaccine candidates indicates that there is a wide range of circulating viruses whose antigenic diversity is not captured with chosen candidates, potentially allowing for the persistence and escape of variants that can cause future epidemics.

Discrete trait diffusion model of data informed regions of the United States.

Discrete trait model across different seasons

The individual season phylogeny was estimated to characterize each epidemic season individually and to jointly estimate a phylogenetic discrete trait diffusion model across the 10-year study space. The diffusion patterns across multiple seasons, each independently estimated, were examined, and

summarized to identify major sources and sinks of transmission across the United States. Broader regional schema (U.S. Census Regions and initial broad partitioning of communities) does not provide enough granularity to make strong inferences about the overall diffusion across the U.S. To observe the finer scale diffusion patterns, observation of the U.S. Census divisions, HHS regions, and Louvain 2 regions is more informative. When examining the Louvain 2 the clusters 5 and 6, which encompass the Western and South Atlantic regions, are shown to be major sources for all other regions across the 10-year period (Figure S2.11). These results of the data-informed regions show similar source-sink patterns across the study space, when compared to the administrative regions where the Pacific and South Atlantic regions for the U.S. Census divisions are major sources to all other regions. The predominance of these major regions, which host large population centers and major transportation hubs for air travel which, would allow for greater diffusion to different regions across the United States and sustained transmission.

Observing each season's transition rates individually and comparing the results across seasons, shows the major source of the Western and South Atlantic regions, which was summarized across the 10-year period (Figure S2.11). By taking a season-by-season approach, the signal of the South Atlantic regions becomes more defined as opposed to the initial phylogenetic analysis of a large sub-sample across multiple seasons. In addition to the major sources, some seasons show source activity from clusters 1, 2, and 3 (Louvain 2), which encompass the Southern and Mideast regions. The seasons where clusters 1, 2, and 3 show source activity indicate that clusters 5 and 6 are major sinks as well. Clusters 5 and 6 of the Louvain 2 regions were previously identified as major source regions, and they may also act as important sinks due to their importance as major population centers and air traffic hubs. This can allow for a wider diversity of circulating viruses and a greater frequency of importations from other regions.

The Markov rewards were estimated for each regional schema jointly and the trunk reward proportion was visualized for each epidemic season (Figure S2.14-S2.16). The results indicate that transmission is seemingly sporadic between different regions during the epidemic season. The proportions of regions across the tree varies depending on the season where certain seasons are observed with periodic waves of high proportions of a given region are seen as opposed to other seasons where the proportion is variable throughout the season.

Jointly estimated discrete trait model

The jointly estimated discrete trait model allows for the characterization of multiple phylogenetic histories of transmission to be combined to create a more comprehensive characterization of the diffusion process. When jointly estimating the diffusion across all seasons in the study space, the major sources of transmission become evident. The jointly estimated BSSVS diffusion model across the 10-year period shows clusters 5 and 6 (Louvain 2) as the major sources, and this reaffirms the results seen which estimating individual seasons that showed the same clusters as major sources (Figure 2.3).

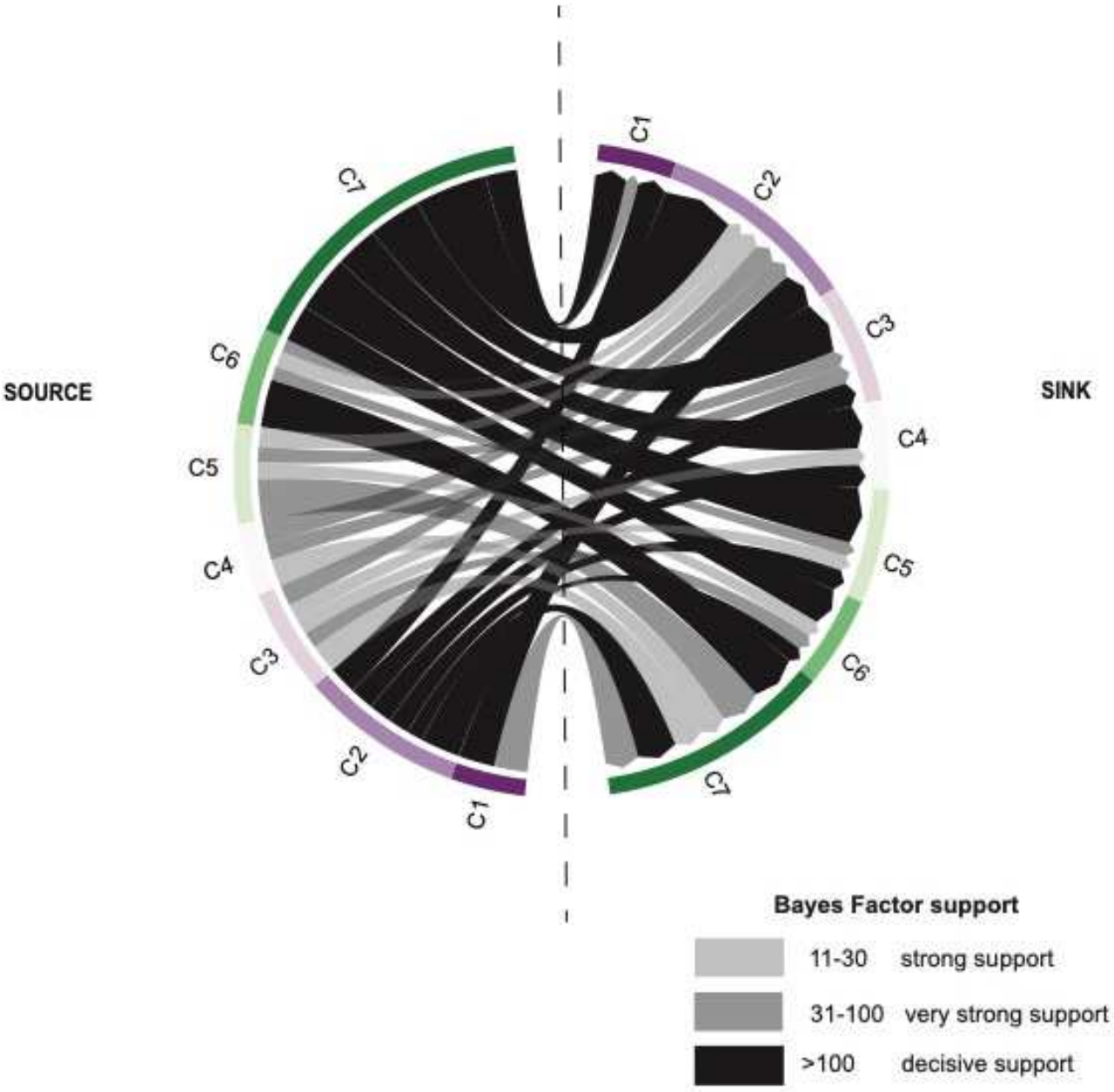


Figure 2.3. jointly estimated BSSVS of epidemic seasons between 2011 and 2020 for the Louvain regional schema. The thickness of the chord corresponds to the discrete trait transition rate and the color corresponds to the Bayes Factor support. Only chords with a posterior probability of 50% or greater are visualized. Color of source and sink bins correspond to the colors seen in Figure 2.1.

These results show decisive support for the major sources, and the relative intensity of the mean transition rate is much greater for clusters 5 and 6. The Markov jump history was jointly estimated and showed cluster 7 was a major source population with the most jumps occurring from cluster 7 to all other sink regions (Figure S2.12). The result of the jointly estimated Markov jumps for the HHS and U.S. Census division schema show the western region of the United States acting as major sources, Region 9 and Census Division 9 respectively. The jointly estimated BSSVS results for the HHS and Census Divisions show similar results to the Louvain model where western U.S. regions act as major sources, however the Southern United States (HHS: Region 4, Division: South-Atlantic, Louvain: Cluster 2) is seen as another major source population across all three regional models (Figure S2.13). The Southern United States and Western United States act as major sink populations across models.

Jointly estimated Generalized Linear Model

A GLM of antigenic, climate, population demographic, transportation, and epidemiological data allowed for the identification of important co-variates for the diffusion rates across multiple seasons for each regional schema. The Louvain GLM model indicated a positive correlation with climatological incidences of temperature, population density at the origin region, airline passenger numbers with the diffusion process and had BF > 3, indicating support for the conditional effect size calculated (Figure 2.4).

Figure 2.4. Results from Generalized Linear Model implemented in BEAST v 1.10.4 for co-variates of the diffusion process between regions of the Louvain regional schema. The conditional effect size panel on the right indicates the level of inclusion for a given variable as a covariate for the diffusion process of the jointly estimated diffusion matrix. The posterior probability panel shows the level of posterior support for the inclusion of a given variable in the GLM. The solid blue line and dotted blue line in the posterior probability panel represent the calculated BF support equal 100 and 3 respectively. The * denotes coefficient HPDs that are greater than 5 and less than -5.

When comparing across the different schema, this was a consistent result where the climate data associated with temperature and temperature indices was correlated with the diffusion process as well as the major demographic factors of population density and migration (Figure S2.17-S2.18). Previous studies have indicated that climate, particularly temperature and humidity, and flight connectivity play significant roles in the epidemic behavior of the virus and transmission across the country. The antigenic co-variates of antigenic distance to vaccine candidates and inter-region distances showed a generally negative conditional effect for the GLM of Louvain diffusion rates. This differed from both the HHS and Census Division results where the HHS GLM model shows half of the vaccine indices are positively correlated and half of indices are negatively correlated. The Census Division GLM model little effect (near zero) of the vaccine indices.

Discussion

This work demonstrates the need for approaches that estimate data-informed regions for use in the phylogeographic analysis. When trying to control for locales that are disproportionate in size and population density, it is important to use available information to create a cohesive network of infection, especially when looking at a given region's smaller divisions such as those found at the precinct district, county, and state level. Information on the place of collection often suffers from a lack of consistent reporting and the reporting of broad regions in administrative units. These quirks of reporting can affect the inferences about transmission made between man-made borders as opposed to a more continuous space of transmission. The case data approach used in this study can also be used for the partitioning of the United States during an epidemic or pandemic to study the progression and regional dynamics of the disease and allow for targeted resource distribution.

A limitation of phylogeographic discrete trait methods is the simplification of a large area that can have complex internal transmission patterns, by using case data and drawing finer divisions within a larger region, these complex diffusion patterns can be observed and accounted for in a wider geographical model. An important limitation of this approach is the availability of sequence data for a given region that is being used. As sequencing technology becomes cheaper and practices of routine clinical isolate sequencing become more regularized, especially regarding the COVID-19 pandemic, the lack of sequence data for a given region can be accounted for. As surveillance systems are developed to study the diversity and transmission of influenza and other pathogens, the joint estimation of phylogenetic histories allows us to study epidemics that are not directly linked over large periods of time and allow us to characterize important transmission patterns that occur frequently, these signals might otherwise be obscured by taking a season-by-season approach where sequence data may be lacking or less representative of a given population.

The United States is a diverse country geographically, climatically, and demographically and this can present major challenges in understanding how these features can affect the evolution and transmission of pathogens in a complex system. Previous studies in North America explore the factors that relate to the transmission of influenza viruses in the United States and have found that temperature and precipitation heavily influence transmission (60). The effects of other major climate variables such as humidity have also been implicated in the transmission and seasonality of influenza viruses (51)(61). A major focus in the study of seasonal influenza virus transmission and evolution is the development of predictive models to forecast the emergence/prevalence of a given lineage in future epidemic seasons. Forecasting efforts aid the development of effective vaccines for potentially emerging lineages and can inform non-pharmaceutical control strategies (62)(63). When attempting to create models that predict the emergence of virus variants and the

potential of a lineage to dominate in a coming epidemic season, it is important to identify and include the major co-variates of transmission in a holistic and effective predictive model. As more robust data are made available, especially in regard to climate and transportation data, important inferences can be made about the importance of these factors in the overall transmission and evolutionary dynamics of the virus.

The GLM has previously been used to identify important predictors of pathogen transmission such as air passenger flow in the study of global H3N2 influenza transmission and spatial distance between bat populations in the transmission of Rabies (49)(39). Our study reaffirms the previous findings of past research into drivers of influenza transmission, which found that climate variables such as temperature and humidity played a critical role in the epidemic behavior of the virus (64)(65)(54). Additionally, our results support previous studies which provide evidence of the importance of transportation in the transmission where air traffic has especially been seen as a major indicator (36)(35)(66). Other transportation data available from the U.S. federal government for Amtrak usage and highway miles driven also given important insight into less rapid but potentially important routes of transmission across state boundaries. We encountered difficulty in the somewhat-arbitrary-until-now choice of locations for discrete traits (primarily what scales are available and which are even feasible given sequence data). Sensitivity analyses when identifying which predictors are chosen (and their effect sizes) based on which regionalization schema are critical when using data-informed regions as opposed to administrative regions.

Future work would augment this regionalization by attempting to source data that has a finer scale to case data (county, city, district level) and use these smaller divisions to better define the regional variation in cases and the overall social network. A finer administrative scale would also allow researchers to group these regions by geographic centroids using latitude and longitude data for a

given locale to try to eliminate the geographic bias that arbitrary state borders might create. These models can be further applied to other influenza sub-types and respiratory viruses such as respiratory syncytial virus and SARS-CoV-2 where incidence data and sequence data are available.

CHAPTER 3: GENETIC AND ANTIGENIC CHARACTERIZATION OF GLOBAL SEASONAL INFLUENZA EVOLUTION, 2017-2022

Damodaran, Lambodhar et al. To be submitted to Virus Evolution

Abstract

The evolution and transmission of seasonal influenza viruses on a global scale has been characterized using genomic data from virus isolates, laboratory tests, and epidemiological models. These methods allow researchers to understand major drivers of virus diversity and the drivers of epidemic behavior. Despite the wide assortment of available data associated with influenza transmission, methods that effectively combine different types of data to create a holistic representation are still lacking. In this study the recent evolution of seasonal influenza subtypes between 2017 and 2022 is described and the results from studying genetic data through phylodynamic modelling and antigenic cartography are combined to utilize antigenic assay data. The combination of phylogenetic metrics and antigenic cartography coordinates is achieved using generalized additive models (GAM). From these analyses it was observed that recent evolution of seasonal influenza is marked by less genetic diversity but consistent rates of evolution. Additionally, the implementation of the GAMs to study to association of phylogenetic distance to antigenic cartography identified major clades and geographic place of isolation for each influenza virus subtype. This analysis presents a framework for the delineation of virus evolution that can provide major avenues for the prediction of evolutionary patterns.

Introduction

Seasonal influenza virus epidemics present an important public health challenge causing significant morbidity and mortality on a global scale (87)(6). The viruses that are recognized as seasonal influenza belong to the family Orthomyxoviridae and include two type A influenza viruses, subtypes H3N2 and H1N1pdm09, and two type B influenza viruses, subtypes B-Victoria and B-Yamagata (41). This respiratory pathogen is responsible for seasonal epidemics that oscillate between the southern and northern hemispheres of the globe, correlating with the timing

of the winter months (51)(88)(31). With the emergence of the SARS-COV-2 virus and the ongoing COVID-19 pandemic human behaviors resulting from pandemic response have disrupted the transmission of these viruses and have potentially created important bottlenecks for the evolution and transmission of these subtypes (89)(90)(91).

Characterization of the molecular epidemiology of seasonal influenza using genetic data has allowed for the inference of important evolutionary patterns and drivers of transmission. Phylodynamic methods which utilize a Bayesian framework for phylogenetic reconstruction and co-inference of different evolutionary models provide a conventional framework for molecular epidemiology of viral pathogens (44)(92). Studies using phylodynamic models have leveraging genomic data and associated metadata to make inferences about the evolution of seasonal influenza. These studies capture important facets of transmissions and evolution such as the boom-and-bust cycle of major epidemics and the how the transitions between major lineages of influenza viruses are affected by cluster specific immunity and the level of population immunity to a given lineage (30)(93). Phylodynamic models have also been used to understand the transmission of seasonal influenza on different geographic scales. Additionally, have been used to identify major locations that serve as source populations for epidemics as well as the drivers of global migration patterns (50)(32)(49).

In addition to genetic data, antigenic data in the form of hemagglutinin inhibition (HI) assay titer data has allowed for the estimation of antigenic space between different influenza isolates using traditional multidimensional scaling (MDS) of titer data for isolates against different antisera (22)(94). These traditional MDS methods allow for the definition of evolutionary space using biological measure of immunity to influenza infection and add important information about virus evolution when compared, contrasted, and added to genetic characterizations of isolates through

phylogenetic methods. The introduction of models such as the Bayesian MDS model, further aim to better utilize available HI data by modeling variability in testing conditions and virus immunology, providing comprehensive evolutionary analysis with paired genetic data (95).

Several studies have been conducted that pair different genetic data, associated metadata, and antigenic data with the aim of better representation of the evolutionary history of viruses and other organisms (96). The field of ecology has explored the use of pairing genetic data with landscape features through redundancy analysis which constrain geographic space in relation to genomic data (97)(98). Previous work characterizing dengue virus evolution has implemented the Generalized Additive Model (GAM) utilizing antigenic cartography data by using the distance between and within antigenic groups of as a function of time (99). Despite these studies, few studies have leveraged linear models, such as the generalized linear model and GAM, to associate phylogenetic measures and antigenic locations.

In this study we use the antigenic coordinate for each isolate as a functional term combining both dimensions against a phylogenetic distance metric, the root-to-tip distance, as a response variable to allow for the paired genomic and antigenic inference of evolutionary space. We find that the GAM allows for the identification of major geographic locations and clades that have previously been correlated with seasonal influenza transmission and evolution. Furthermore, we characterize the molecular epidemiology and evolution of landscape and interrogate important features of recent influenza transmission and evolution. We find that despite a reduction in incidence and number of distinct lineages for each subtype there is a constant rate of evolution over time.

Materials and Methods

Data

Sequence data for the hemagglutinin protein and associated metadata for all H3N2 isolates collected globally between 2017-06-01 and 2022-06-15 was downloaded from GISAID (18). Sequence data was aligned using MAFFT and sequences causing gaps were removed after visual inspection of sequence alignments (68). Sequences with metadata indicating variant H3N2, chimeric sequences, experimental sequences, poor sequence quality were removed. Sequence data was then merged with associated antigenic data provided from CDC. Antigenic data sets for titer data collected after 2005 were used (2022-06-15), filtering for all related data for each subtype and available HI assay data, for H3N2 viruses HI and HI+ Oseltamavir assay data were combined. Only Isolates with paired sequence and antigenic data were used for further analysis. For each sub-type this resulted in the following number of characterized isolates: H3N2 (n=2740), H1N1 (n= 4172), B-Victoria (n= 2006), B-Yamagata (n= 2388). The lineage and number of nucleotide and amino acid substitutions from reference were obtained using NextClade using the alignment for isolate of each subtype (72).

MDS

The RACMACS R program was used to create antigenic cartographies for associated antigenic data (22). Antigenic data in the form of HI assay titer data was filtered by date and duplicate pairs of antigen-sera pairs were removed, taking the first occurrence by test date of the given pair. All experimental antigen/sera were removed. The antigenic cartography was run for 100 optimizations in two dimensions. The Euclidean distance for each isolate to the vaccine strain was calculated using a provided in-house python script. The vaccine candidate strains were determined using the WHO vaccine candidate recommendation strain (100). The nomenclature

for vaccine strains were abbreviated to the place of origin as first two characters and year of isolate as last two characters e.g. A/Hong Kong/2019 = HK19.

Phylogenetics

The associated sequence data for each antigen used in the MDS analysis for each subtype were used for Bayesian phylogenetic reconstruction using BEAST v.1.10.4 (27). The following models, priors, and parameters for were used for each subtype. The SRD06 codon partition model, uncorrelated relaxed clock, and constant coalescent model were used (73)(74). Maximum likelihood phylogeny inferred using IQ-TREE v.1.6.12 for each subtype were analyzed in TempEst v1.5.5 to provide mean values for normally distributed root height and ucld.mean priors (70)(46). Six independent MCMC with a 50 million chain length were performed. The results of each independent chain were analyzed using Tracer v1.7.2, and the three best runs were combined to ensure adequate effective sampling size (ESS) of parameters greater than 200 (76). The root-to-tip distances of all taxa for each sub-type MCC tree was calculated using the program TreeStat v.1.10.4.

BMDS

HI assay titer data previously described for each subtype of influenza used in traditional MDS was formatted for use in BMDS (95). The BMDS model was implemented in BEAST v1.10.4. A posterior set of 500 trees from previously described phylogenetic reconstructions of each subtype were used as an empirical tree set for use in the tree inclusive BMDS model. For each model described an MCMC chain length of 500 million steps was used.

GAM

The “mgcv” R package was used to estimate the GAM models for each subtype (110). The estimated root-to-tip distances for the Bayesian phylogenetic reconstruction of each subtype

were used as the dependent response variable used in the GAM. In implementation of the GAM antigenic coordinate data was used as a smooth were the x and y coordinates were combined in the following function:

$$RTT = \boldsymbol{f}(x,y) + \boldsymbol{f}(\text{predictor 1}) + \boldsymbol{f}(\text{predictor 2}) + \ldots + \alpha$$

Where y is the root-to-tip distance response variable, $\boldsymbol{f}()$ represents the smoothing function, (x,y) represents the antigenic coordinates, and α is the model intercept.

To incorporate categorical variables into our GAM the variables were set as ordered factors and the smooth function was applied to the coordinate as the parametric term for the factor. An example implementation in pseudo code: ("gam(rtt ~ (x,y, by = factor))"). An interaction term between x and y coordinate is implemented in the model which allows the terms to be treated together and using a categorical-continuous interaction (i.e, the by= form) allows us to interrogate the different factors of a given categorical trait for isolates such as place of isolation or constituent clade.

All code and XMLs used in this analysis can be found in the following Github repository: https://github.com/ldamodaran/Influenza-antigenic-2017-2022

Results

Bayesian phylogenetic reconstruction and Multidimensional scaling

Nucleotide sequence data for the hemagglutinin protein and paired antigenic data for isolates collected between 2017-06-01 and 2022-06-01 were used to reconstruct the phylogeny and antigenic cartographies for each seasonal influenza subtype (Figure 3.1).

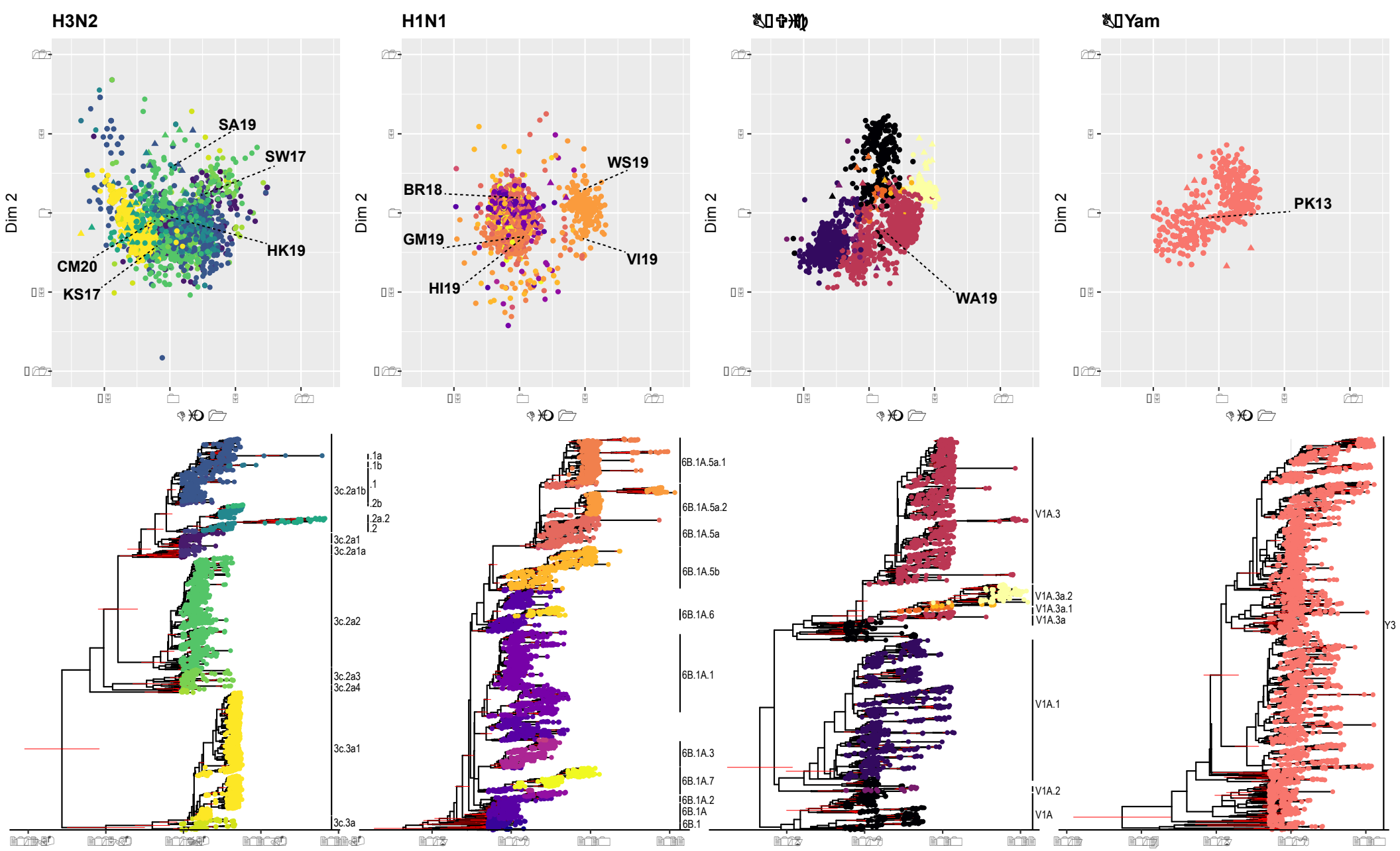


Figure 3.1. Antigenic and genomic characterization of seasonal influenza subtypes globally between 2017 and 2022. The results of multidimensional scaling for HI assay titer data are shown with the vaccine candidate isolates marked by dashed line and abbreviated name/year, isolates are colored by clade as designated by genomic data using NextClade. The corresponding time-scaled maximum clade credibility Bayesian phylogenetic reconstructions for the HA protein of each subtype is seen below the corresponding antigenic cartography with the lineages of each major clade colored with legend. The 95% BCI for nodes with a posterior support of at least 70% were visualized with red bars.

Bayesian phylogenetic reconstruction using BEAST for each subtype showed that the time to most recent common ancestor for each subtype was within 3-5 years of the oldest isolate consistent with seasonal influenza evolutionary pattern, H3N2: 2015.578 (BCI: 2012.421, 2014.6566), H1N1: 2015.317 (BCI: 2014.7136, 2015.9504), B-Victoria: 2015.164 (BCI: 2014.2814, 2015.9503), B-Yam: 2013.29 (BCI: 2012.0103, 2014.6456). Nucleotide substitution rates (substitutions/site/year) were also consistent for each subtype, H3N2: 5.139×10^{-3} (BCI: 4.7588×10^{-3}, 5.5204×10^{-3}), H1N1: 5.66×10^{-3} (BCI: 5.377×10^{-3}, 5.925×10^{-3}), B-Victoria: 2.626×10^{-3} (BCI: 2.417×10^{-3} , 2.846×10^{-3}), and B-Yamagata: 2.881 x 10-3 (BCI: 2.6491×10^{-3}, 3.1225×10^{-3}). When isolates were colored on the phylogeny by genetic lineage, as described by Nextclade, the clear separation of clades is seen for each subtype. Upon visual inspection of corresponding antigenic cartographies, constructed using traditional MDS, there is some separation in antigenic space between different genetic lineages for each of the subtypes. Of note, B-Yamagata only has one constituent genetic lineage, the Y3 lineage.

Bayesian Multidimensional scaling

To further characterize the molecular epidemiology of seasonal influenza, the paired genetic and antigenic data were integrated into Bayesian multidimensional scaling (BMDS) frameworks which allow for the modelling of different antigenic constraints (95). Using previously described HI titer data and an empirical set of trees from prior Bayesian phylogenetic reconstruction (Figure 3.1), multiple BMDS models were implemented to define the antigenic space and estimate different measures of antigenic variability. The BMDS method allows for models that account for the variability in the reactiveness of ferret antisera and the reactiveness of the antigens themselves taking an empirical Bayesian approach to specifying the mean and variance of titers for both serum potency and virus avidity. The models that account for this variation are henceforth referred to as

"effects" models. Additionally, BMDS it allows for the modelling of drift in the antigenic space through the inclusion of a diffuse prior on the relative location of viruses and sera in cartographic space due to the potential for separate cartographies to have equal likelihood values with differing positions of tested antigens and sera. The results of the BMDS models showed some variation in the scale of antigenic units between models and across subtypes (SI Appendix, Figure S3.1). The antigenic cartographies for the BMDS model that includes the phylogenetic history showed the widest range in antigenic units for each subtype. When cartographies were annotated by genetic lineage, separation of antigenic space by was less pronounced for BMDS models of H3N2, while H1N1 and B-Victoria models showed greater distinction between major lineages. The results of antigenic parameter estimates showed that serum potency for H3N2 and H1N1 models with phylogenetic history were higher than all other models for other subtypes (Figure 3.2). Measures of virus avidity were generally similar across models and subtypes with a slightly wider 95% Bayesian credible interval for the H3N2 model with phylogenetic history. The overall diversity of multiple competing lineages and continued epidemic occurrence might account for the higher avidity in H3N2 viruses compared to other subtypes. Lower values for location drift were observed for H1N1 and B-Yamagata subtypes, this is consistent with the reported lower rates of antigenic drift seen in lineages during global circulation (31).

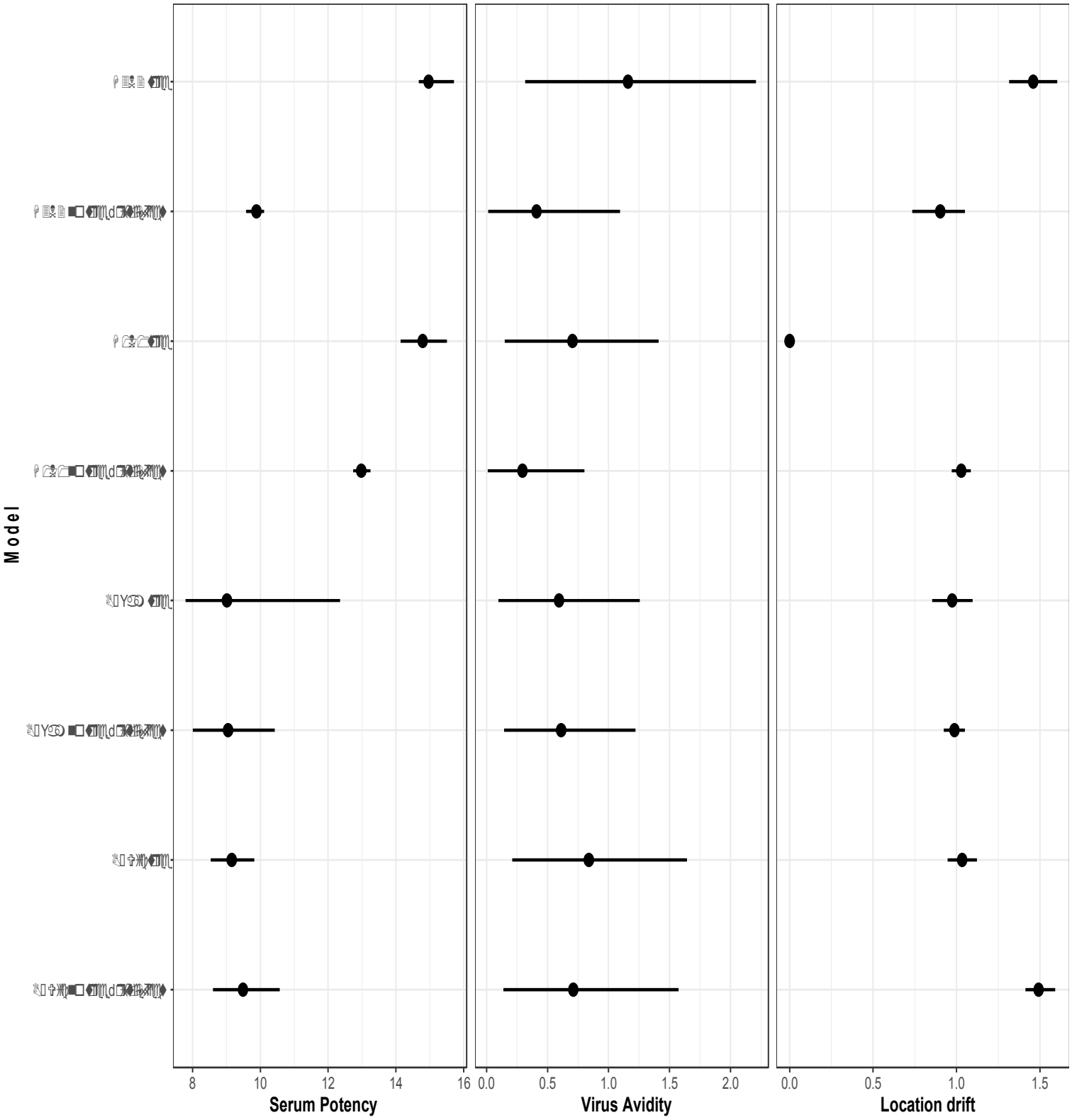


Figure 3.2. – **Results of BMDS for each subtype and evolutionary model**. The median value and the 95% BCI for serum potency, virus avidity, and location drift for each subtype and evolutionary model of BMDS implemented in BEAST.

Estimation antigenic drift

To further investigate the evolutionary trends and quantify antigenic drift within the study space, the traditional MDS antigenic cartographies for each subtype were used to estimate the Euclidean distance for each isolate to the WHO recommended vaccine candidates isolated between 2017 and 2022 (100). These results were used to estimate the antigenic distance of isolates from each vaccine candidate over time, annotating by the constituent clade and by the number of amino acid substitutions from reference (SI Appendix, Figure S3.2-S3.6). For H3N2 vaccine candidates the antigenic distance over time was positively correlated for the SW17, SA19, and HK19 strains (Pearson correlation: r 0.1 - 0.45) while the KS17 and CM20 showed negative correlations overtime (Pearson correlation: r = -0.13 and -0.041 respectively). For each of the vaccine candidates of H1N1 viruses there were significant positive correlations (Pearson correlation: r =0.42 - 0.6) with antigenic distance over time. B-Victoria isolates showed highest correlation of antigenic distance from vaccine candidate over time (Pearson correlation: r = 0.72). For each subtype there is considerable diversity in the number of clades represented before 2020, with only 2-4 lineages persisting post-2020. The number of amino acid substitutions was greater for each subtype after 2020, indicating consistent antigenic drift over time. To further characterize the molecular epidemiology of seasonal influenza viruses the antigenic distance to vaccine candidates can provide an important measure of isolate relatedness to implemented vaccines and other circulating viruses. An antigenic unit represents a two-fold dilution of antiserum in the HI assay (22). Isolated viruses are considered similar antigenically if there is less than or equal to 4-fold titer difference when compared to vaccine strains (101)(97). Using this information as a guide, the

proportion of isolates for a given epidemiological week that are antigenically distinct from the vaccine candidate can be shown for each subtype (Figure 3.3).

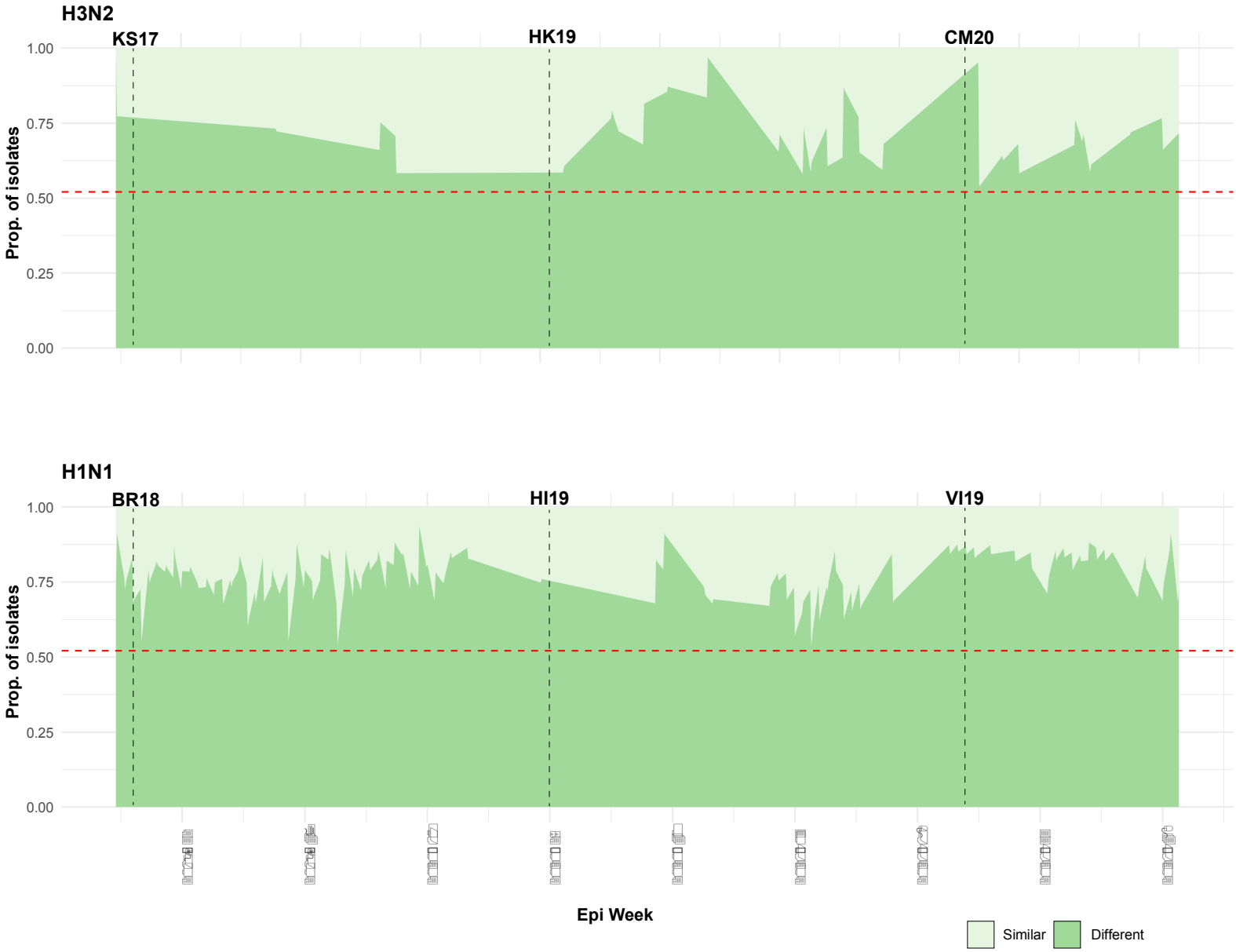


Figure 3.3. – Proportion of antigenically similar and antigenically different isolates to vaccine candidates by epidemic week. Proportion of global influenza isolates characterized antigenically that are antigenically different from the recommended vaccine strain, by CDC epidemic week for H3N2 and H1N1. Isolates that are antigenically different are defined as having a four-fold dilution in antisera (two antigenic units). The dashed red line represents the maximum proportion of the U.S. population (0.521) that is vaccinated against seasonal influenza (Data = CDC FluVaxView). The vertical dashed black line indicates when a given vaccine candidate was implemented.

Quantifying the level antigenic similarity for an isolate to given vaccine can indicate how robust sampling measures are for virus isolates and can show the breadth of diversity within the virus population. Importantly it can be observed that in most cases the vaccine greater than 50% of viruses characterized are antigenically distinct from the vaccine candidate. This is an important metric to consider as a wide diversity of circulating viruses allow for vaccine escaped lineages to persist and go on to seed future epidemics. Because this represents a global sampling of diversity it is important to note that in the inter-epidemic season period for the northern hemisphere, when the proportion of the population vaccinated is lowest, there are high proportions of virus isolates that are antigenically different from the vaccine, this naive population might allow for greater chance of genetically similar lineages to emerge. Additionally, it can be inferred that the increase in proportion of isolates that are antigenically different post-implementation of a vaccine candidate can be the result of the elimination of strains antigenically similar to the vaccine due to increase population immunity. The proportion of all isolates over the entirety of the study space that are antigenically different for a given vaccine candidates varied (SI Appendix, Figure S3.7). For influenza B-viruses all the samples characterized within the last 3 years were antigenically different from the vaccine candidate. Vaccine candidates for H3N2 showed a higher proportion of escape than H1N1 vaccine candidates, H1N1 vaccine strains WS19, HI19, and VI19, showed only ~50% of isolates being antigenically distinct.

Root-to-tip distances and antigenic drift

In addition to the antigenic distances to vaccine candidates, another metric to study antigenic drift is the root-to-tip distance, which measures the cumulative branch lengths from given taxa to the root of a phylogeny and estimates genetic distance over time. For each subtype the number of nucleotide and amino acid substitutions from reference vs the root-to-tip distance was plotted

showing a significant ($p < 0.05$) positive correlation (SI Appendix, Figure S3.8). This result is indicative that the constant antigenic drift and rate of evolution that occurs in seasonal influenza is persistent. There are a higher number of nucleotide and amino acid substitutions for H3N2 and H1N1 viruses compared to B-Victoria and B-Yamagata. When the relationship between root-to-tip distance and nucleotide substitutions was faceted by the constituent clade the Pearson correlation for all significantly supported lineages was positive (SI Appendix, Figure S3.9-S3.12). These different clades showed different ranges in correlation coefficients for each subtype, H3N2: Pearson Correlation R= 0.31-0.79, H1N1: Pearson Correlation R= 0.55-0.91, and B-Victoria: Pearson Correlation R= 0.42-0.71.

Generalized additive model

The root-to-tip distance can be further used as a measure of genetic evolution over time for each isolate in a phylogeny and can be paired with other data to make inferences about virus evolution. This allowed for the inference of the effects of different variables on the root-to-tip using the generalized additive model (GAM). The implemented GAM combined terms for the antigenic location of a given isolate to study the relationship between the root-to-tip distance and the antigenic cartography, different linear predictors, and metadata factors associated with the isolate. The results of the GAM showed that for each subtype different places of collection for the isolate, the constituent clade, and the number of nucleotide substitutions can be identified as major predictors of genetic evolution (Figure 3.4).

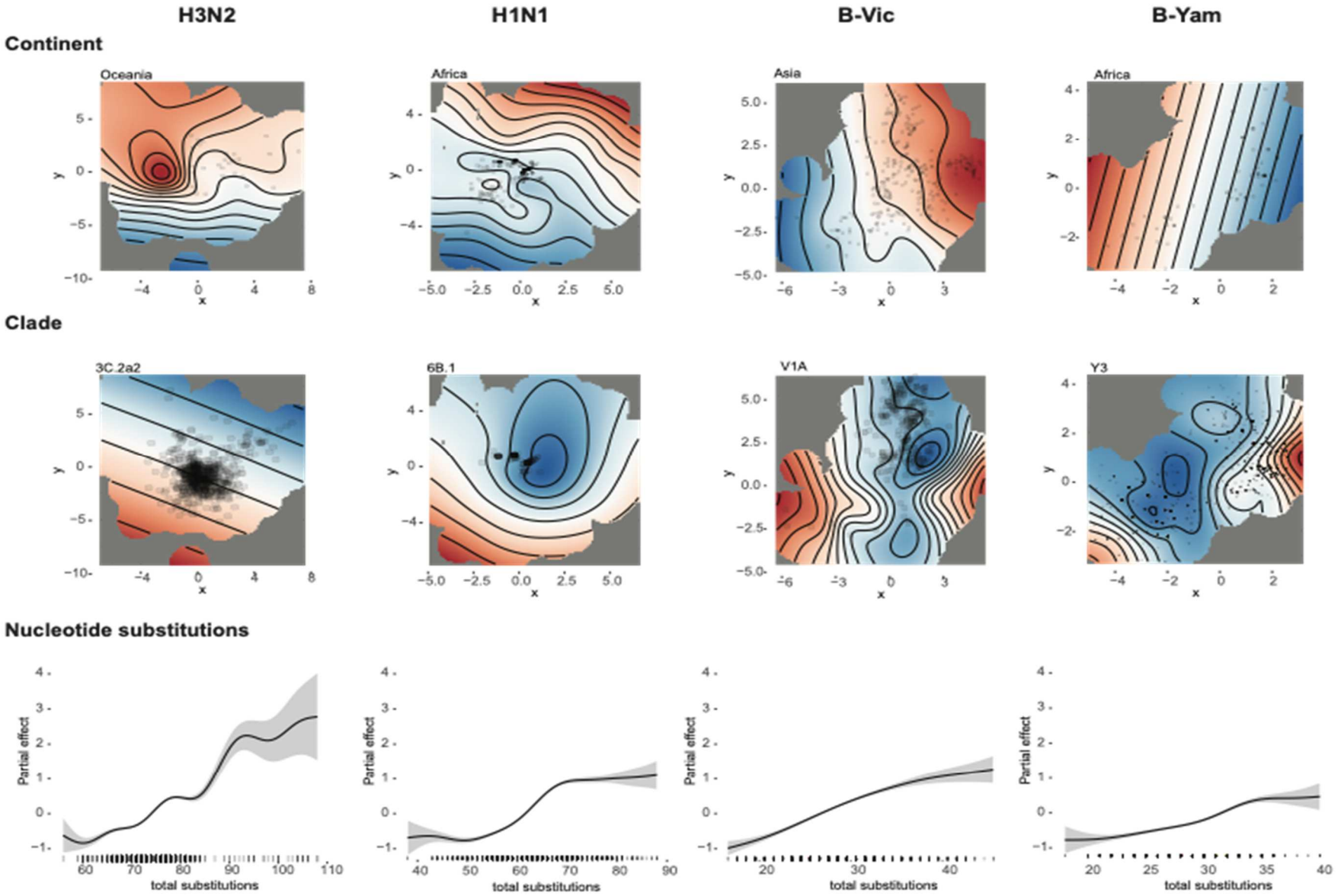


Figure 3.4. Results of GAM of root-to-tip distance to selected predictors. Predictors with highest F statistic value for place of isolation, constituent clade, and genetic distance metric for each isolate of each subtype are visualized. The color of across cartographic space represents the partial effect of the predictor where red indicates a positive partial effect and blue indicates a negative partial effect. For the total substitutions the confidence interval for estimated partial effect is shown.

The results for the GAM of each subtype identified predictors with strong support for the significance of smoothing terms and when preforming basis dimension checks, and identifies predictors that that have little to no support at all (Table S3.1-S3.8) The partial effects of the different predictors on each the response root-to-tip distance response variable for each subtype were visualized and indicate that certain genetic lineages, geographic locations, and metrics associated with have positive effects on areas to a given subtype cartography (SI Appendix, Figure S3.13-S3.28). The partial effects for the constituent clade of a given isolate of H3N2 were high for most clades, with the highest support for 3c.2a and 3c.2a1b.2a2 (F=67.22 and 60.34 respectively). The 3c.2a clade was a major clade that circulated early in the study space, the 3c.2a1b.2a2 clade arose pre-2020 and went on to predominate in subsequent seasons. Similarly, for H1N1 viruses, the 6B.1 clade was identified as the most supported clade within the subtype model (F=22.958), the second most supported clade was 6B.1A.5a.1 (F= 11.115) which was a major lineage that arose pre-2020 and persisted into subsequent seasons. Geographic partial effects showed that for H3N2 viruses Oceania and North America had the greatest support, while H1N1, B-Victoria and B-Yamagata indicated Asia, Oceania, and Africa. The highly supported partial effects of nucleotide substitution on the root-to-tip distance are consistent with previous results, of note for H1N1, B-Victoria and B-Yamagata subtypes there is a plateau in the partial effect of nucleotide substitutions as the number of substitution increases. Diagnostics of the model fit were preformed and showed that models had good fit for each subtype. Quantile-Quantile plots of residuals showed strong linear relationship (SI Appendix, Figure S3.29-S32).

Discussion

The use of the GAM with measures of genetic evolution and antigenic coordinate data can potentially be used to predict the divergence of isolates from vaccine isolates in the absence of

robust testing by allowing for the prediction of novel coordinates from genetic measures such as the root-to-tip regression. The GAM is a powerful tool which can allow for the more robust use of antigenic coordinate data in the future. To accurately analyze antigenic space it is critical that the GAM and other methods that account for and analyze differences in antigenic ordinal space, such as the Procrustes rotation test and Non metric multidimensional scaling, are investigated for their use in paired genetic analysis (102)(103). It is important to recognize the importance of changes in the efficacy of HI assays in distinguishing H3N2 viruses. Recent changes to receptor binding domain of the HA protein through glycoslyation have reduced the sensitivity of the HI assay and has resulted in the introduction of alternative neutralization assays such as the focus reduction assay and HINT assays (40). Estimating antigenic cartographies with these different assays can for H3N2 viruses can allow for more accurate characterization of antigenic space for subsequent paired genetic analyses.

The possible extinction of the B-Yamagata lineage is important consider in this analysis (90)(104)(105). Due to the lineages lack of genetic diversity as seen by it only having one clade, this might have played an important factor in its overall elimination when non pharmaceutical interventions (NPIs) were implemented. There is considerable evidence that NPIs play a major role in halting transmission of seasonal influenza which would allow for both restriction of available hosts and existing immunity and vaccination to effectively prevent further chains of transmission (106)(107)(108).

Taken together, this analysis shows that there is still a constant rate evolution occurring in seasonal influenza viruses despite global pressures from COVID-19 and NPIs and that constant characterization of the molecular epidemiology of seasonal influenza viruses is important in adequately capturing important trends in evolution and transmission. Recently in the beginning

of the 2022-2023 northern hemisphere influenza season there has been a significant increase in influenza like illness activity with H3N2 viruses accounting for the bulk of influenza infections (109). It is important that these paired genomic and antigenic approaches are utilized by public health practitioners to robustly characterize and respond to the ongoing and growing burden of seasonal influenza epidemics.

CHAPTER 4: COMPARISON OF ANTIGENIC CARTOGRAPHY BETWEEN DIFFERENT H3N2 INFLUENZA VIRUS ASSAY TYPES

Damodaran, Lambodhar et al. To be submitted to Influenza and Other Respiratory Diseases

Abstract

The antigenic characterization of seasonal H3N2 influenza viruses has been primarily achieved using the hemagglutination Inhibition Assay. In recent years changes in the binding affinity to surface proteins have due to changes in the receptor binding region have led to a lack of discerning power in these assays. To address this neutralization-based assays such as the focus reduction assay and the HINT assay have been adopted. These new assay types produce titer data that can be used to construct antigenic cartographies to assess the evolution of the virus over time. Using previously described multidimensional scaling methods for log-transformed titer data (Smith et al, 2004), the ordinations of antigenic cartographies and phylogenetic distance matrices for overlapping sets of isolates for H3N2 seasonal influenza viruses were compared. It was observed that there is strong correlation in the antigenic space for isolates for contemporary assays (FRA and HINT). Additionally, distance matrices for each assay cartography had low correlation with phylogenetic distance matrices. The HINT assay had the greatest linear relationship between phylogenetic distance and the Euclidean distance for isolates. These findings indicate the necessity of different methods to characterize the antigenic evolution of seasonal influenza viruses and the importance of finding methods for comparing their resulting antigenic cartography ordinations.

Introduction

Seasonal influenza virus epidemics are responsible for significant morbidity and mortality resulting in between 9.2-35.6 million cases and between 140 -710 thousand hospitalizations United States every season (2)(111). One of the primary influenza viruses that represent the bulk of infections during a seasonal influenza epidemic is the H3N2 influenza A virus. In the study of H3N2 influenza evolution different methods have been used to antigenically characterize viral

isolates and understand their relative transmission potential and their immunological properties. The Hemagglutination Inhibition (HI) assay has been the primary laboratory test to characterize the relative antigenic properties of different virus isolates for wide array of other pathogens over the past 60 years such as human influenza B viruses, H5N1 avian influenza viruses, and rubella virus (112)(113). The HI assay measures the level of agglutination in red blood cells (RBC) from hosts sera inoculated with a given reference strain of virus to measure the differential agglutination in the presence of different test antigen strain. The hemagglutinin (HA) surface protein of the influenza virus binds to the sialic acid receptors of host cells (114). The antibodies found in the host sera (typically ferret sera) inhibit erythrocyte agglutination by binding the receptor binding domain of the HA present in the host RBCs (115). Serial dilutions of virus are used to determine the amount of virus needed to visually measure the agglutination of RBCs resulting in a titer value associated with a given antigen and antisera. While HI assays typically use ferret sera but have also used sera from humans and other animals such as swine and guinea pigs (116)(117)(118).

Recent changes in the receptor binding domain (RBD) of the HA protein have led to the decreased efficacy of HI assays due to the inability to agglutinate RBCS. This is most likely due to glycosylation of the RBD which has been shown to modulate virus release and virulence (119). This change in sensitivity lead to the adoption of the Focus Reduction assay (FRA) and later the high content imaging-based neutralization test (HINT) assays (40). The FRA and HINT assays are virus neutralization assays which quantify virus kinetics and antigenicity through the imaging of infected host cells and the quantification of numbers of infected cells in the presence of antibodies. These assays determine serum neutralization titer which quantifies the titer at which a given antigen virus is cleared by measuring the infected number of cells post inoculum (120)(121)(122). These neutralization assays were implemented with the intent to address changes in agglutination

and virus adaptation to cell culture. The HINT assay is a strictly controlled neutralization assay that strictly controls the amount of inoculum used in each imaging well and uses an empirically determined an optimal ratio between the number of cells, dilution of virus, concentration of reagents, and incubation time.

Epidemic Identification and characterization of influenza isolates has primarily been performed via the HI assay and, as sequencing technology has become more widely available, genomic data with associated metadata (123). The addition antigenic assays like the FRA and HINT assays to HI assay data and genetic allow for a more robust modeling of evolution. Multidimensional scaling (MDS) of the log-transformed HI assay titer data has allowed for the estimation of distances between virus isolates and host sera (22). MDS methods are utilized to optimize the Euclidean distances between points and allow for a visual representation of antigenic evolution that can discern major antigenic clusters. These MDS methods have been implemented to create "antigenic cartographies" to represent evolution in seasonal influenza and changes in population immunity over time (124)(51). Despite the recent introduction of the FRA and HINT assays, comparison of the resultant antigenic cartographies to HI antigenic cartographies is lacking. In the following study the comparison of ordination through the Procrustes test and the comparison of distance matrices via the mantel test was performed to investigate the differences between assays and their resultant ordinations (102)(125). A greater correlation of antigenic space between the FRA and HINT assays as opposed to the HI assays was observed. Additionally, phylogenetic distance between isolates had low correlation with antigenic distances derived from corresponding antigenic maps.

Materials and Methods

Phylogenetics and antigenic cartography

Antigenic data in the form of HI, FRA, and HINT assay data for H3N2 seasonal influenza isolates collected globally between 2017-06-01 and 2022-06-01 was obtained from the CDC. The paired genomic data for the HA gene segment was downloaded from GISAID using the isolate identifiers and strain name (18). The genetic lineage of isolates was determined using the Nextclade program (72).

The Racmacs package v.1.0.14 was used to construct antigenic cartographies in 2-dimensions using 100 optimization for each assay data set (126). The paired genomic data for each assay data set was used to reconstruct Bayesian time-scaled phylogeny. Genomic data was aligned using MAFFT v7.453 and alignments were visually inspected using Geneious (68). To account for the region of the HA protein which is immunologically recognized in the HI and neutralization assays separate genomic data sets for the HA1 region of the HA segment were constructed by trimming genomic sequences to the HA1 region and were used for each of the following analyses. Initial phylogenetic reconstruction to diagnose temporal signal was performed for each data set using IQtree v1.6.12 and TempEst v1.5.3 (127)(46). Bayesian phylogenetic reconstruction was performed using BEAST v1.10.4. (27). A constant coalescent model, relaxed log normal clock model, and SRD06 codon partition model were utilized (27)(74)(73)(128). A UPGMA starting tree was used and normal priors on the root-height with a mean value corresponding to the root-height of previously constructed maximum likelihood phylogeny, as well as a uniform prior on the ucld.mean from the substitution rate of ML trees were used. Each genomic dataset was run in triplicate for an MCMC chain-length 50 million states, sampling every 5000 states. Diagnosis of runs was preformed using Tracer v.1.7.2 to ensure adequate effective sampling size (ESS) for all

parameters (76). Runs were combined removing 10-20% burn-in and a maximum clade credibility (MCC) tree was constructed using the LogCombiner v1.10.4 and TreeAnnotator v1.10.4 programs respectively (129).

Procrustean randomization test

To statistically compare different ordinations of assays, the coordinate data for isolates that were shared between two given assays were used in the Procrustean randomization test. This randomization test is a statistical test that randomly varies the distance between coordinates to assess the significance of the Procrustes distance between any two given sets of coordinates. The non-randomness (significance) between two ordinations was determined using the 'protest' function which performs a symmetric Procrustes analysis multiple times to estimate the significance of the Procrustes statistic by creating a null distribution of Procrustes distances based on random permutations of the coordinates (130). The Procrustean randomization test was preformed using the 'vegan' R package (103)(131).

Distance matrix comparisons

The phylogenetic genetic distance matrices for each Bayesian phylogeny were calculated using the cophenetic pairwise distance between taxa of the MCC phylogeny for each genomic data set using the 'ape' R package (132). The comparison of distance matrices was performed using the Mantel test implemented using the 'vegan' R package (131). The Mantel test calculates the correlation between corresponding positions of dissimilarity or distance matrices. This allows for the estimation of a measure of correlation between two matrices (133).

Results

Phylogenetic analysis and MDS

The Bayesian phylogenetic reconstructions for each antigenic assay and corresponding antigenic cartographies show different levels of separation between major lineages of H3N2 influenza (Figure 4.1).

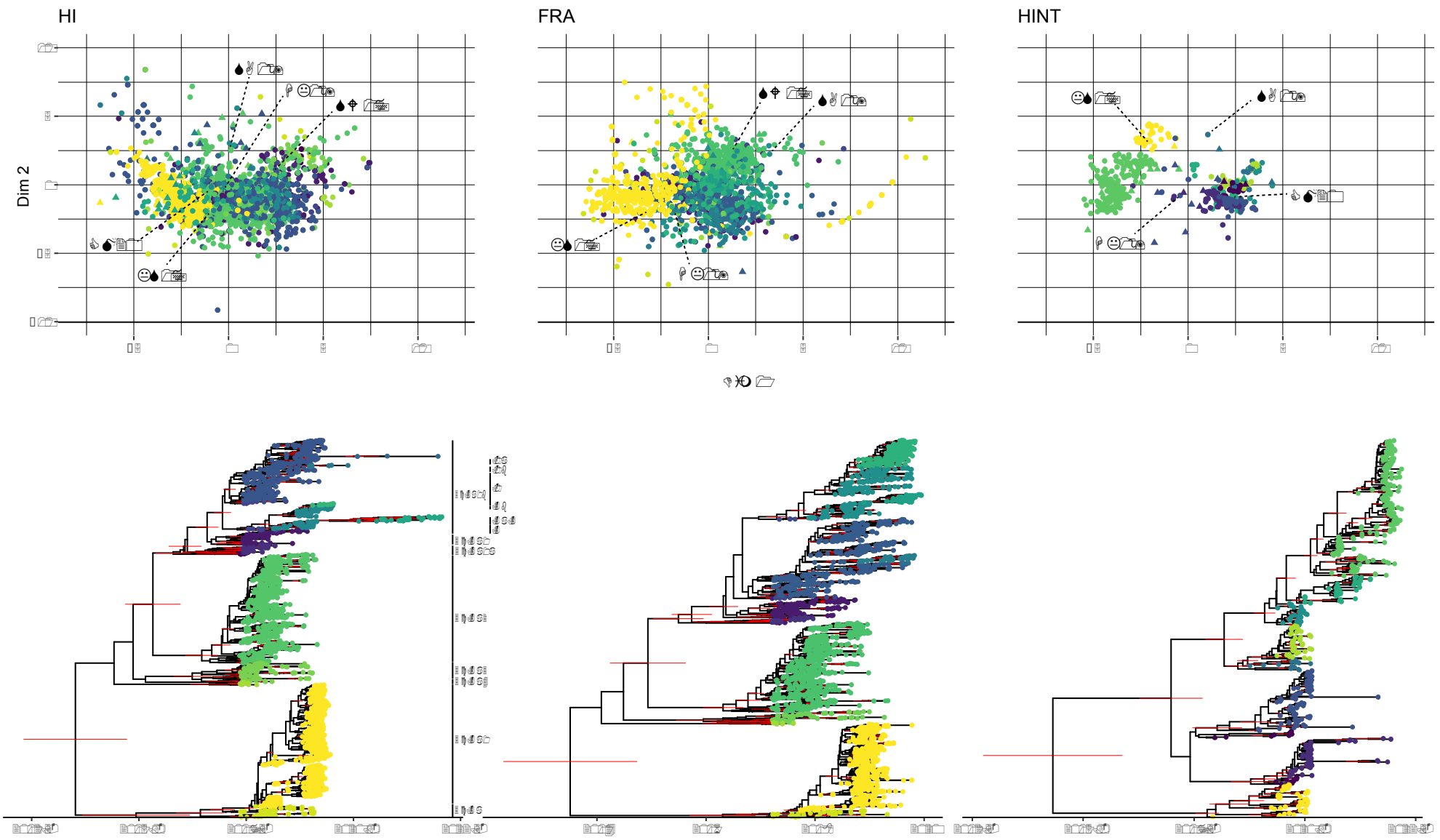


Figure 4.1 Results of MDS and Bayesian phylogenetic reconstruction for isolates collected between 2017 and 2022 across different antigenic assays. The antigenic cartography corresponds to the phylogeny directly below. The color of coordinate points and the tips of the phylogeny correspond to the lineage determined using NextClade. The 95% BCI is labeled with a red bar for nodes with at least 70% posterior support

When labeling antigenic cartography points using discrete values for the genetic lineage of the isolate, visual inspection shows that there is greater segregation in the antigenic space for the FRA and HINT assay cartographies. The position of vaccine candidate isolates, as indicated by the abbreviated place and year shorthand notation, shows that the placement of vaccine candidates in antigenic space is similar with differing space between candidates across assays. The HINT assay cartography showed the greatest separation among vaccine candidates in Cartesian space while HI and FRA had a more condensed grouping of isolates. Using available metadata for isolates associated with sequence data the cartographies were labeled accordingly with different metadata for the year and continent of isolation (Figure S4.1-S4.4.3). There was no apparent geographic structuring for each assay with the majority of isolates characterized originating from North America. The temporal structuring of antigenic space was most pronounced for the HINT assay data while the HI and FRA data showed some heterogeneity.

The phylogenetic reconstructions for each assay in BEAST allowed for estimation of time to the most recent common ancestor (TMRCA) and substitution rate. The TMCRA estimates for HI, FRA, and HINT assay data sets respectively were as follows: 2013.578 (95% BCI: 2012.421, 2014.657), 2013.633 (95% BCI: 2012.333, 2014.766), and 2014.262 (95% BCI: 2012.647, 2015.777). Nucleotide substitution rates (substitutions/site/year) for the HI, FRA, and HINT assay data sets respectively were as follows: 5.139E-3 (95% BCI: 4.759E-3, 5.52E-3), 4.786E-3 (95% BCI: 4.465E-3, 5.106E-3), and 3.842-3 (95% BCI: 3.372E-3, 4.33E-3).

Ordination stress was evaluated using scree plots which allow for the inference of the minimal number of dimensions for appropriate ordination. Using the heuristic breakpoints of the scree plot for each assay it was observed that the ordination stress for at least 2 dimensions were appropriate to visualize the ordinations of antigenic data (Figure S4.4-S4.6). Additionally, the validity of

ordinations was diagnosed using the Shepard stress plot. These plots show the relationship between data matrices and the ordination distances. These plots allow for the inference of stress in the ordination and identification of potential outliers and to diagnose the validity of a given ordination. The stress pots for each assay showed a linear relationship between map distance and the table distances (Figure S4.7-S4.9). The FRA stress plot showed a slight drop for higher map/table values. The grouping of points in the HI and FRA stress plots reflects the different graduated values for titer values (i.e. 180,360,720 etc) whereas the stress plot for HINT data more diffuse grouping of points reflects the neutralization tests ability to quantify a specific value associated with the number of infected cells in a given assay well.

Procrustean Randomization Test

To compare the different ordinations of the antigenic data the Procrustes test was used and showed some varying differences in position between assays. The Racmacs Procrustes map function visualization of the shared isolates between and their relative positioning between ordinations (Figure 4.2

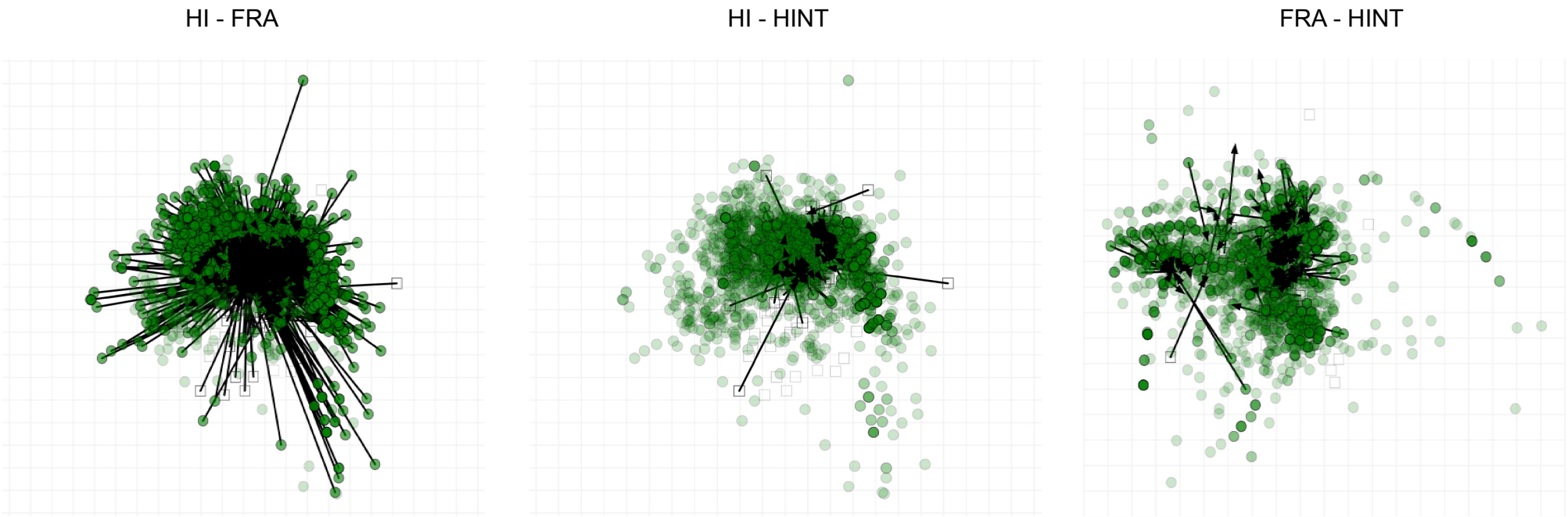


Figure 4.2. Procrustes projections for each antigenic assay comparison. Arrows indicate the relative positioning of a given isolate in the comparison assay ordination. Points without errors are isolates that do not have overlap between the two antigenic datasets.

The visual inspection of the Procrustes positioning of the assays shows that, in relation to the HI coordinates, isolates in both the FRA and HINT assay points were closer together. This relationship is seen also for FRA and HINT, where coordinates associated with HINT assay were situated closely in the map. This is reflected in the results of the Procrustes randomization test were the range of Procrustes errors for the FRA/HINT assay [Procrustes comparison is much smaller than the HI/FRA and HI/HINT assay Procrustes comparisons (Figure S4.10-S4.12).

There was a higher correlation between the FRA and HINT assays than the other assay comparisons (Table 4.1).

Assay Comparison	Sum of Squares	Correlation (m^2)	Significance
HI,FRA	0.8839	0.3408	0.001
HI,HINT	0.7262	0.5233	0.001
FRA,HINT	0.4265	0.7573	0.001

Table 4.1. Results of Procrustean randomization test for each pair of assay coordinate data. The sum of squares, correlation in a symmetric Procrustes rotation and the significance of the test are reported for Procrustean randomization for 999 permutations.

There was a greater sum of squares value for the HI/FRA assay comparison than the HI/HINT and FRA/HINT comparisons. The higher sum of squares indicates a higher variability in Procrustes distance values towards the mean. The combination of high sum of squares and low correlation indicate that the HI/FRA comparison are not identical in the scale and rotational transformations for the coordinate locations. The residual errors for the different assay comparisons showed that

there were larger residual errors in the HI/FRA assay comparison than the HI/HINT and FRA/HINT comparisons (Figure S4.13-S4.15).

Mantel test

The result of the comparison of phylogenetic distance matrices of isolates to the Euclidean distances estimated between isolates as determined by MDS coordinates shows that the FRA and HINT assay distances were slightly more correlated with phylogenetic distance than the HI assay distances. The Mantel statistic r values, which are based on the Pearson product-moment correlation, for comparisons against phylogenetic trees constructed for the whole HA gene segment the r = 0.3719 for the HI assay, r = 0.4523 for the FRA assay, and r = 0.4189 for the HINT assay. The Mantel statistic values were similar for the mantel test of phylogenies constructed for the HA1 region of the HA segment showing a lower value for the FRA assay comparison, r = 0.3802, the lowest value for HI assay comparison, r = 0.3533, and the HINT assay have the highest value r = 0.4175. All comparisons showed a low correlation with phylogenetic distance, had a statistical significance of p=0.001, and were based on 999 permutations.

Mantel correlograms for each assay were estimated to show the spatial correlation of the data sets. with the resulting shape of the correlograms representing the spatial gradient by which the two matrices are related (Figure 4.3).

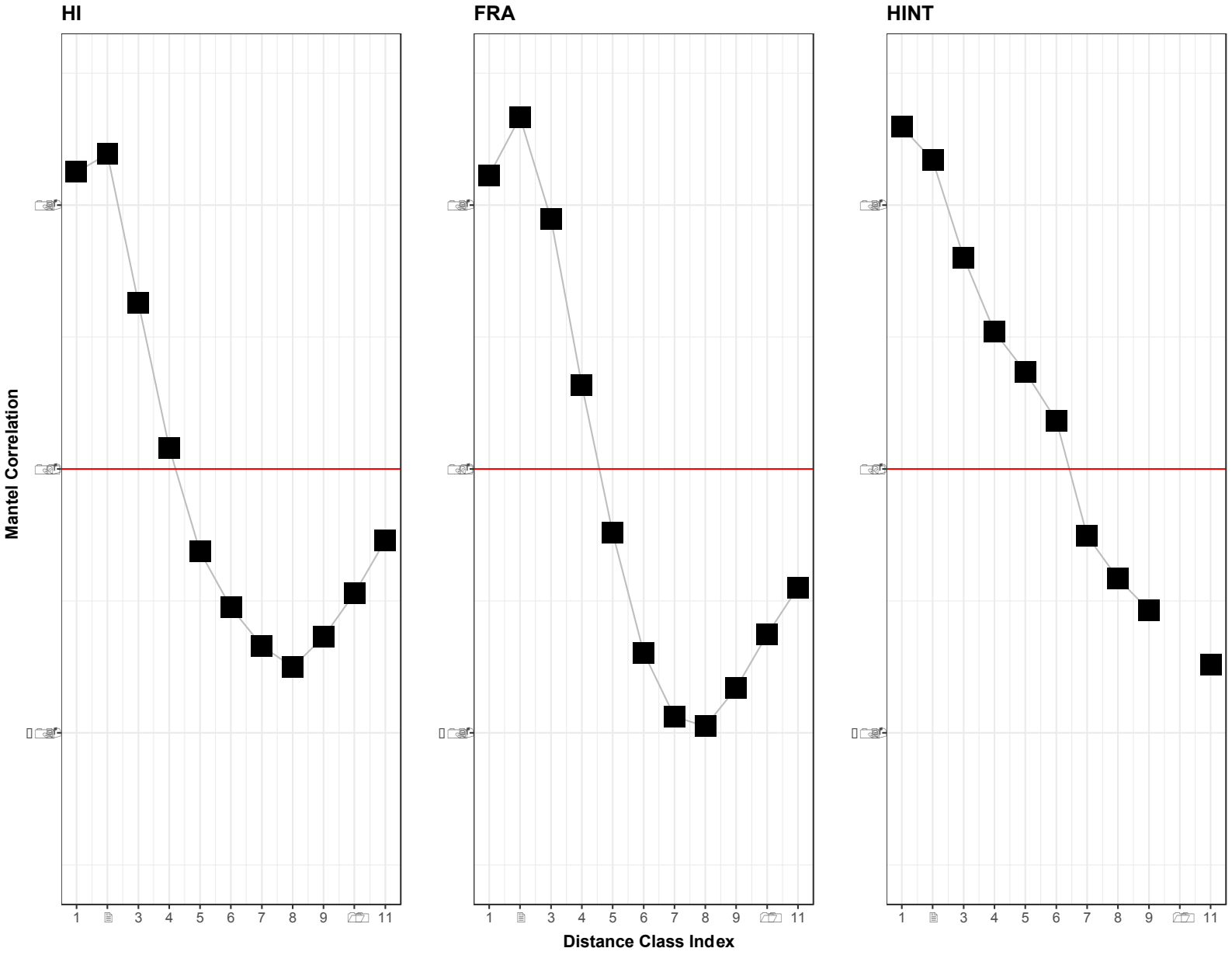


Figure 4.3. Mantel correlogram for antigenic assays tested against corresponding phylogenies for the complete HA sequence. The non-significant correlation values were not plotted for a given distance class index.

A linear gradient is observed for the HINT assay correlation indicating that the distances of both the phylogenetic distance matrix and the HINT antigenic distance matrix were directly correlated. For both the HI and FRA assay correlations there is a single "bump" in the gradient for the lowest distance which indicates a significant positive correlation for the lower values followed by a significant negative spatial correlation gradient which reverses at distance class of 8 to a positive

spatial correlation (133). This behavior in the HI and FRA correlograms can be interpreted as certain regions of the cartography being more correlated with phylogenetic distance than other regions. In the case of the HI assay this might be due to the relative lack of sensitivity and ability to segregate antigenic space. Similar trends in the correlograms for the mantel test between antigenic distances and phylogenies constructed using the HA1 genomic region of the HA protein were observed (Figure S4.16).

Discussion

Understanding the similarities in discerning antigenic space between these different assays is important in assessing the virulence and evolution of novel strains of influenza viruses. The recent reduction of discerning power of the HI assay is partly what lead to the adoption of the FRA and HINT assays due to glycosylation and changes at the RBD of the HA protein. As the relative sensitivity of these compared to both other assays and genetic data, it is important to provide a quantification of the correlation between these assays to inform the use of these assays when determining the emergence of novel strains and when studying different pathogens. There is evidence that single amino acid changes near the RBD of the HA protein can result in major antigenic changes, this warrants robust and sensitive methods to detect these changes antigenically across different assays (134)(135). In addition to this attention should be paid to the NA protein and the antigenic landscape that is potentially under characterized due to the lack of utilization of other assays on a wider scale. These NA protein-based assays such as the Micro-neuraminidase inhibition assay and the Enzyme-linked lectin assay provide more data on antibody neutralization that can be used in antigenic cartography (136).

The availability of FRA data over time is not as robust as HI data and is currently not a major antigenic assay in use. This results in a smaller data for use in comparison, warranting further antigenic characterization with the assay for more robust comparison. The relatively recent adoption of the HINT assay has resulted there being less data available for isolates collected before 2019, leading to contemporary data representing the bulk of cross analyzed isolates. The interruption of the typical seasonal influenza epidemic patterns by the COVID-19 pandemic further reduced the sampling of isolates between after March of 2020 (90)(119). The Global scientific community's ability to implement assays other than the HI assay is another major point of concern. The cost of materials and access to appropriate lab facilities remains a challenge in the wider usage of assays like the HINT assay which is why HI assays, which are relatively cheap to conduct, remain a major comparative tool today.

The use of spatial analysis methods for different datatypes is important to consider for further inquiry into the significance of cartographies between assays and how they relate genetic data. Euclidean distances be augmented by different multi-scale analysis methods such as Trend-surface analysis, Moran's eigenvector maps, and multivariate ordination methods (137). The analysis conducted in this study and future spatial techniques should also be further expanded to other pathogens such as the B-Victoria and H1N1 pdm seasonal influenza viruses.

The antigenic similarity based on distance in antigenic units between isolates can be different between assays the threshold for similarity is typically recognized as a 2-fold titer dilution (1 antigenic units) for HI assay data, 4-fold titer dilution (2 antigenic units) for FRA assay data, and 8-fold titer dilution (4 antigenic units) for HINT assay data (97). This consideration is important when looking at the separation and emergence of major lineages. Further study into the inter-laboratory variation between assays and for the same for the same set of isolates as there is

evidence of variation that could potentially impact the diagnostic power of these assays (138). This analysis which attempts to quantify the differences between antigenic cartographies of antigenic assays is a preliminary attempt to study this variation, further research into the different methods of antigenic cartography such as the Bayesian MDS and other assays is critical in comprehensively characterizing influenza viruses for effective and actionable surveillance and prevention.

CHAPTER 5: CONCLUSION

Thesis Summary

This thesis has presented methodological advances in the field of phylogeography and has identified and utilized new and important methods to pair different forms of data to comprehensively study the evolution and transmission of human seasonal influenzas viruses. Additionally, this work provides important comparisons of major characterization tools and provides a level of correlation between these different tools to evaluate their diagnostic capability.

Chapter 2 of this dissertation describes a novel methodology for the partitioning of geographic space for use in phylogeographic analysis. This chapter focused on the influenza dynamics of H3N2 influenzas in the United States between 2011 and 2020. The state level ILI based network that was analyzed using the Louvain community detection algorithm provided discrete trait partitions for the United States which had a greater marginal likelihood support than the U.S. Census Divisions and administrative regions. This data-driven discrete trait partitioning was used in further phylogeographic analyses, jointly estimating the discrete trait diffusion across a 10-year period. The jointly estimated discrete trait diffusion matrix was finally used to study the influence of different climate, demographic, and antigenic predictors on the viral diffusion process via the GLM. The results of the GLM show that the diffusion process was affected by antigenic, climate, and demographic factors relating to the source/sink regions. This work underscores the importance of studying separate seasonal epidemics together to make large scale inferences about general

source/sink diffusion dynamics and associating different environmental and demographic factors with the viral diffusion process.

Chapter 3 of this thesis describes the global phylodynamic analysis of H3N2, H1N1, B-Yamagata, and B-Victoria seasonal influenza viruses between 2017 and 2020. This work combines the results from studying genetic data through phylodynamic modelling and antigenic cartography to utilize antigenic assay data. The combination of phylogenetic metrics and antigenic cartography coordinates is achieved using generalized additive models (GAM), treating the root-to-tip distance of taxa from a Bayesian phylogenetic reconstruction as the response variable in relation to the placement of isolates in an antigenic cartography. From these analyses we find that recent evolution of seasonal influenza is marked by less genetic diversity but consistent rates of evolutions. Additionally, the implementation of the GAMs to study to association of phylogenetic distance to antigenic cartography identified major clades and geographic place of isolation for each influenza virus subtype. This analysis presents a framework for the delineation of virus evolution that can provide major avenues for the prediction of evolutionary patterns. This work describes the recent evolution and transmission of seasonal influenza on a global scale that leverages different paired datatypes with existing and new models. The methods used in this study provide a more comprehensive investigation of contemporary influenza evolution and a framework for the use of paired genetic and antigenic data in the study of seasonal influenza and other pathogens. Furthermore, it offers an investigation into influenza dynamics under the pressures associated with the COVI-19 pandemic.

Chapter 4 described the comparison of antigenic cartographies resulting from different antigenic assay types. The ordinations of isolates in relation to each other can be different depending on the assay used in antigenic characterization. This analysis compares the ordination of isolates that were

tested for the HI assay and two neutralization-based assays, the FRA and HINT assays. The antigenic characterization of seasonal H3N2 influenza viruses has been primarily achieved using the hemagglutination Inhibition Assay. In recent years changes in the binding affinity to surface proteins have due to changes in the receptor binding region have led to a lack of discerning power in these assays. To address this neutralization-based assays such as the focus reduction assay and the HINT assay have been adopted. These new assay types produce titer data that can be used to construct antigenic cartographies to assess the evolution of the virus over time. Using the procrustean randomization test it was found that there is strong correlation in the antigenic space for isolates for contemporary neutralization assays (FRA and HINT). Additionally, each of the antigenic assays were compared with the phylogenetic distance matrices for the taxa in the cartographies paired Bayesian phylogeny. Utilizing the mantel test to see the correlation between distance matrices for each assay cartography showed that there was low correlation with phylogenetic distance matrices for each antigenic assay. The HINT assay had the greatest linear relationship between phylogenetic distance and the Euclidean distance for isolates. These findings indicate the necessity of different methods to characterize the antigenic evolution of seasonal influenza viruses and the importance of finding methods for comparing their resulting antigenic cartography ordinations.

Challenges

There are important challenges that need to be addressed when characterizing the transmission and evolution of seasonal influenza viruses. Chief among these issues is sampling bias. The sampling of genomic data has grown substantially overtime but still has important issues that need to be addressed. One such issue is the lack of comprehensive metadata associated with the timing and place of isolation. Geographic place of isolation associated with genomic data in major repositories lacks fine granularity and typically the lowest geographic resolution provided for sequence data is at the State level. Information at different finer scales such as on the city, municipality, and district of isolation is seldom found and may only be derived from descriptive strain names that include the city name for the isolate. Furthermore, longitude and latitude data are virtually unavailable for most publicly available isolates. Health privacy is a chief reason for a lack in depth for geographic place of isolation information, the concern that having in-depth data on place of isolation can identify were individuals may life or work can be a major concern for patient safety (139). These safety concerns can be addressed by taking relative centroids in a grid-like fashion across geographic space and associating sequence data with these centroids to provide a deeper level of geographic spread information. Geographic place of isolation can allow for the collection of several different associated datatypes which can be used in subsequent analysis such as the climatic data, demographic data and transportation data which have been used to study influenzas transmission and can be used in a similar fashion using GLM models as described in Chapter 2. Climate data associated with place of isolation can take the form of local temperature, humidity, and precipitation. Demographic data associated with place of isolate can take the form of local population size, age structure, and socio-economic background. Transportation data can take several different forms when associated with the place of isolation, a chief example is human

mobility to workplace which is collected by the American community survey of the U.S. Census bureau (140). Additionally, flight data for origin and destination from the FAA can more accurately associated with place of isolation if geographic metadata has greater level of resolution. It is a challenging to adequately integrate these different transportation data without fine scales and as metadata for place of isolate becomes finer the network models that can be created from mobility and flight data can be efficiently integrated. Another important metadata characteristic for strains is comprehensive data about the host. This includes but is not limited to age, gender, race, socio-economic background, and infection severity. It is imperative that stricter measures for data quality are adhered to by major genetic data repositories in order to better aid in comprehensive analyses of outbreaks and transmission. In many cases for submission to repositories such as NCBI's Genbank and GISAID, there are very few required data labels for submission of sequence data. Countries in the global north represent the bulk of available isolate data with North America being the major contributor of sequence data to public repositories. These patterns for influenza data have been observed for data relating to COVID-19 sequence data where the vast majority of sequence data has strong biases to North America and Europe. In Chapter 3 of this thesis sequence data is used for influenza viruses that circulated during the COVID-19 pandemic, this is during a time when transmission of influenzas viruses was greatly depressed due to non-pharmaceutical interventions by the human population such as quarantine, social distancing, and mask wearing (90). This lack of data is important to consider when attempting to make observations about the overall trajectory of influenza virus evolution. As COVID-19 continues to circulate it is important that the global surveillance apparatus is actively testing clinical isolates for respiratory cases robustly to detect influenza viruses in addition to suspected COVID-19 and other respiratory pathogens.

In the study of phylogeography the use of state defined administrative regions is the basis of determining viral diffusion rates between major geographic locations. These administrative regions and borders can be arbitrary and based on socio-political boundaries that do not necessarily reflect the demographics of a region. This arbitrary nature in defining administrative boundaries can translate to artificial grouping of populations in discrete traits used in diffusion models that can obscure important internal transmission signals and may lead to misrepresentations in the identity of the most probable source and sink populations. Continuous space phylogeographic diffusion models offer one administrative border agnostic method to study viral diffusion but is reliant on accurate data regarding the location of isolation for a sample (141). These methods allow for a diffuse view in space and time of viral dispersion. The pairing of continuous space phylogeographic methods and fine-scale discrete location data for isolates can potentially identify major sources of spread within a given community or between major metropolitan districts.

In Chapter 2 the adjacency matrix representing the network of transmission for the United States was studied using a Louvain community detection algorithm. This is considered a "greedy" algorithm, meaning that it chooses a heuristic threshold to use in the optimization of cluster detections, in the case of community detection the heuristic threshold which maximizes the modularity of the network is used. The use of greedy algorithms for cluster detection such as the k-means clustering algorithm method offer strong benefits for computational tractability but can be extremely sensitive to outliers (142). This can be a challenge if data sources used for adjacency matrix construction are poorly or erratically sampled. Another important challenge in dealing with greedy algorithms is its inability to consider the veracity of previous steps in the clustering process, this can mean that if the researcher does not robustly and repetitively test the algorithm implementation randomly the algorithm can choose a sub-optimal clustering to base successive

modularity steps on (143). While this iterative process is simple to implement via recursion, it can be computationally expensive. It is important to test a wide host of algorithms and assess not only the differences in resultant clusters and the computational tractability but a level of statistical support for a given partitioning schema.

Antigenic cartography is utilized as an important tool in the study of influenza evolution and is used in each experimental chapter of this thesis. The accuracy and variability in antigenic assay data, particularly HI data is an important challenge that needs to be addressed especially in regard to the construction of antigenic cartographies. The challenges to antigenic cartography presented by the changes in the RBD of the HA protein in the past decade are important to study as the HI assay has become less reliable. In addition to the problem of identifiability the nature of immunity exhibited by the host is important to consider as ferrets produce monoclonal antibodies versus in human populations where individuals produce polyclonal antibodies and their respective immune pressures are potentially more variable (144)(43). The BMDS method attempts to account for variability in serum potency and virus avidity and is an important step in trying to address the variability in testing conditions and host biology (95). Additional parameters such as the scaling of antigenic distances between isolates based on two different antigenic assays can potentially aid in better identifying antigenic space. Neutralization assays are important tools that are continuously being developed and utilized, it is important that differences between neutralization assays are studied, and their utility is gauged as a cost-effective and accurate tool for influenzas characterization. In addition to differences in assays is important to immunologically and antigenically characterize other surface proteins for the influenza virion, namely the Neuraminidase (NA) protein. The NA protein which facilitates the viral release after virus replication has emerged as a major target for humoral immune responses and evidence shows that

they play an important role overall in the recognition and elimination of influenza from the host (145). The methods used for antigenic cartography can be used with titer data for inhibition and neutralization assays studying the NA protein and can give some insight in the antigenic landscape for this other major protein which is critically important in attempting a more holistic and accurate immunological understanding of influenzas viruses in human populations (146).

Another major challenge presented is the availability of computational resources and the computational tractability of the analyses laid out in this work. Phylodynamic models require considerable resources to perform complex computation this necessitates access to high performance computing infrastructure to robustly test hypothesis. As computational resources become more widely available and cheaper this challenge has been lessened but it is important to be aware of, especially when considering the effects on public health surveillance and response. Computational resource sharing via online platforms such as CIPRES are powerful open-source tools that allow for the distribution of computational resources for public use. As the demands for phylodynamic models and their inferences increases in addressing major public health problems it is important that these public resources are bolstered. It is critical that a robust computational infrastructure is maintained for groups that are utilizing phylodynamic methods to answer serious real-time epidemics and future pandemics.

Future Work

Each chapter of this thesis represents important contributions to the study influenza viruses and their molecular epidemiology. There are several important avenues for future work that are important to consider for each experimental chapter. In the study of the phylogeography the use of a data driven framework for the partitioning of discrete geographic space for use in discrete trait phylogenetic analysis is lacking. In this study, ILI data was used to create adjacency matrices that were analyzed using community detection algorithms. The data input of the proportion of ILI visits for each state is only one such possible data that can help to elucidate networks of transmissions. The analysis of different data such as human mobility data from different sources such as cell phone location data and transportation surveys have been used in different studies of disease dispersion but have not been used in phylogenetic studies (147). By studying and comparing difference network constructions for influenza viruses, and other respiratory viruses, important inferences can be made about potential sites for public health interventions and the deployment of resources for transmission mitigation.

The future implementation of different linear models in a Bayesian framework is important to consider as they provide methods for the testing of associated metadata that allow for deeper study of influenza evolution. Co-variates of the discrete trait diffusion process have been studied using GLM models, allowing for the association of different datatypes phylogenetic diffusion process directly. The Generalized Additive Model is one such framework that might allow for a similar inference with greater variable flexibility, allowing for the inference of partial effects in a robust framework that, while potentially more computationally complex, will allow for the more accurate representation of predictor data. Additionally, the framework adopted in Chapter 3 of this thesis,

treating the root-top-tip distance for isolates as an independent variable of different isolate associated meta offers another potentially useful framework for Bayesian phylodynamic inference This would allow for the use of GLMs to study a taxa associated metrics such as the root-to-tip distance with different co-variates predictors as opposed to studying the discrete trait diffusion process.

Realtime tracking of virus evolution can provide stronger insights into the transmission dynamics of seasonal influenzas and can help to evaluate and identify a novel strains potential to cause epidemics. The right size approach to sampling can help to capture some of the circulating diversity of influenza viruses and allow for a real-time tracking system (20). Platforms like the Nextstrain platform have allowed for the real-time tracking of epidemics and the evolutionary history of pathogens can be studied as data becomes available (148). Another major platform to study the real-time evolutionary history of pathogens is the UShER program which curates a large-scale phylogeny for SARS-COV-2 sequences to aid in the placement of isolates in major evolutionary clades. The Nextstrain platform utilizes a maximum likelihood phylogenetic framework while the UShER program utilizes a parsimony based phylogenetic reconstruction method. Both phylogenetic reconstruction methods have certain pros and con, for example computational efficiency and tractability is much higher for these methods than Bayesian phylogenetic reconstructions, but this might come at the expense of accuracy. Methodologies have been developed to allow for the addition of sequence data to ongoing Bayesian phylogenetic reconstructions in BEAST as new isolate data is made available (149). It is important that future phylogenetic work in real-time characterization considers the use of Bayesian reconstruction as a viable and tractable methodology as technological demands such as computational power become

less burdensome, and this provides a public health practitioner and public facing GUI for use in real-time tracking of epidemics.

An important and less discussed aspect in the field of Phylodynamics, and more broadly the field of computational biology, is the environmental impact of computation. It is imperative that future work focuses on computational efficiency to reduce the carbon footprint of these analyses. The onus on scientists to take active steps to make their computation more efficient and reduce the amount of unneeded or redundant analysis and computation has never been so salient as our society battles climate change. This can in part be aided by preliminary analyses which allow for the selection of informative priors and providing analyses highly supported maximum likelihood starting trees to reduced probability space.

Conclusion

This body of work was attempted with a goal producing accurate molecular epidemiology analyses that provide actionable results for public health interventions. This work aimed to study seasonal influenzas viruses at varying geographic scales and utilized a diverse set of data to answer important questions about the recent evolution of seasonal influenzas. In addition to the molecular epidemiology characterizations of recent seasonal influenza transmission the comparison of antigenic cartographies and their originating data is important in ensuring a more accurate characterization of influenza evolution. The use of Bayesian phylodynamic frameworks is a powerful tool used in this body of work and has demonstrated its power provide key insights into influenza transmission and evolution. In this thesis these methods elucidated major geographic source and sink regions for H3N2 seasonal influenza and provided a framework to study the phylogeography of data-driven discrete traits. These methods also aided in the identification of

major co-variates of the diffusion process and allowed for the paired inference of influenza evolution with antigenic assay data in the form of antigenic cartographies.

The methods and research that are described in this body of work are not restricted in their utility to the study of human seasonal influenza. The described phylodynamic methods can be utilized for the study of other RNA viruses as well as other pathogens. It is important that as the global public health community looks to address ongoing and future epidemics and pandemics, methods such as those used in this body of work are considered to make important decisions that can ultimately lead to a reduction in morbidity and mortality to influenza viruses.

REFERENCES

1. K. Stöhr, Influenza—WHO cares. *The Lancet Infectious Diseases* **2**, 517 (2002).

2. M. A. Rolfes, *et al.*, Annual estimates of the burden of seasonal influenza in the United States: A tool for strengthening influenza surveillance and preparedness. *Influenza Other Respi Viruses* **12**, 132–137 (2018).

3. N. M. Bouvier, P. Palese, The biology of influenza viruses. *Vaccine* **26**, D49–D53 (2008).

4. M. Banning, Influenza: incidence, symptoms and treatment. *Br J Nurs* **14**, 1192–1197 (2005).

5. K. E. Lafond, *et al.*, Global Role and Burden of Influenza in Pediatric Respiratory Hospitalizations, 1982–2012: A Systematic Analysis. *PLoS Med* **13**, e1001977 (2016).

6. H. Nair, *et al.*, Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet* **378**, 1917–1930 (2011).

7. A. D. Iuliano, *et al.*, Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *The Lancet* **391**, 1285–1300 (2018).

8. C. J. Russell, M. Hu, F. A. Okda, Influenza Hemagglutinin Protein Stability, Activation, and Pandemic Risk. *Trends in Microbiology* **26**, 841–853 (2018).

9. M. Cohen, *et al.*, Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase. *Virol J* **10**, 321 (2013).

10. A. W. Hampson, "Influenza virus antigens and 'antigenic drift'" in *Perspectives in Medical Virology*, (Elsevier, 2002), pp. 49–85.

11. E. Kirkpatrick, X. Qiu, P. C. Wilson, J. Bahl, F. Krammer, The influenza virus hemagglutinin head evolves faster than the stalk domain. *Sci Rep* **8**, 10432 (2018).

12. K. B. Westgeest, *et al.*, Genomewide Analysis of Reassortment and Evolution of Human Influenza A(H3N2) Viruses Circulating between 1968 and 2011. *J Virol* **88**, 2844–2857 (2014).

13. S.-W. Yoon, R. J. Webby, R. G. Webster, "Evolution and Ecology of Influenza A Viruses" in *Influenza Pathogenesis and Control - Volume I*, Current Topics in Microbiology and Immunology., R. W. Compans, M. B. A. Oldstone, Eds. (Springer International Publishing, 2014), pp. 359–375.

14. G. J. D. Smith, *et al.*, Dating the emergence of pandemic influenza viruses. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 11709–11712 (2009).

15. F. Carrat, A. Flahault, Influenza vaccine: The challenge of antigenic drift. *Vaccine* **25**, 6852–6862 (2007).

16. C. J. Burrell, C. R. Howard, F. A. Murphy, “Epidemiology of Viral Infections” in *Fenner and White’s Medical Virology*, (Elsevier, 2017), pp. 185–203.

17. M. Biggerstaff, S. Cauchemez, C. Reed, M. Gambhir, L. Finelli, Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infect Dis* **14**, 480 (2014).

18. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22** (2017).

19. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID’s innovative contribution to global health: Data, Disease and Diplomacy. *Global Challenges* **1**, 33–46 (2017).

20. M. Rosenthal, *et al.*, Evaluation of Sampling Recommendations From the Influenza Virologic Surveillance Right Size Roadmap for Idaho. *JMIR Public Health Surveill* **3**, e57 (2017).

21. A. J. Drummond, O. G. Pybus, A. Rambaut, R. Forsberg, A. G. Rodrigo, Measurably evolving populations. *Trends in Ecology & Evolution* **18**, 481–488 (2003).

22. D. J. Smith, *et al.*, Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **305**, 371–376 (2004).

23. T. Bedford, *et al.*, Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3**, e01914 (2014).

24. B. T. Grenfell, *et al.*, Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* **303**, 327–332 (2004).

25. O. G. Pybus, A. Rambaut, GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics* **18**, 1404–1405 (2002).

26. A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, W. Solomon, Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics* **161**, 1307–1320 (2002).

27. M. A. Suchard, *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4** (2018).

28. R. Bouckaert, *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **15**, e1006650 (2019).

29. B. F. Finkenstädt, A. Morton, D. A. Rand, Modelling antigenic drift in weekly flu incidence. *Statist. Med.* **24**, 3447–3461 (2005).

30. K. Koelle, S. Cobey, B. Grenfell, M. Pascual, Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans. *Science* **314**, 1898–1903 (2006).

31. T. Bedford, *et al.*, Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217–220 (2015).

32. J. Bahl, *et al.*, Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proceedings of the National Academy of Sciences* **108**, 19359–19364 (2011).

33. J. Wallinga, M. Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B.* **274**, 599–604 (2007).

34. M. Prosperi, *et al.*, Molecular Epidemiology of Community-Associated Methicillin-resistant Staphylococcus aureus in the genomic era: a Cross-Sectional Study. *Sci Rep* **3**, 1902 (2013).

35. C. Viboud, M. A. Miller, B. T. Grenfell, O. N. Bjørnstad, L. Simonsen, Air Travel and the Spread of Influenza: Important Caveats. *PLoS Med* **3**, e503 (2006).

36. R. F. Grais, J. H. Ellis, A. Kress, G. E. Glass, Modeling the Spread of Annual Influenza Epidemics in the U.S.: The Potential Role of Air Travel. *Health Care Management Science* **7**, 127–134 (2004).

37. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol* **5**, e1000520 (2009).

38. , 2010 Census Regions and Divisions of the United States. *2010 Census Regions and Divisions of the United States* (2021).

39. N. R. Faria, M. A. Suchard, A. Rambaut, D. G. Streicker, P. Lemey, Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Phil. Trans. R. Soc. B* **368**, 20120196 (2013).

40. P. A. Jorquera, *et al.*, Insights into the antigenic advancement of influenza A(H3N2) viruses, 2011–2018. *Sci Rep* **9**, 2676 (2019).

41. V. N. Petrova, C. A. Russell, The evolution of seasonal influenza viruses. *Nat Rev Microbiol* **16**, 47–60 (2018).

42. M. Wille, E. C. Holmes, The Ecology and Evolution of Influenza Viruses. *Cold Spring Harb Perspect Med* **10**, a038489 (2020).

43. S. E. Hensley, *et al.*, Hemagglutinin Receptor Binding Avidity Drives Influenza A Virus Antigenic Drift. *Science* **326**, 734–736 (2009).

44. E. M. Volz, K. Koelle, T. Bedford, Viral Phylodynamics. *PLoS Comput Biol* **9**, e1002947 (2013).

45. G. Baele, M. A. Suchard, A. Rambaut, P. Lemey, Emerging Concepts of Data Integration in Pathogen Phylodynamics. *Syst Biol*, syw054 (2016).

46. A. Rambaut, T. T. Lam, L. Max Carvalho, O. G. Pybus, Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* **2**, vew007 (2016).

47. G. B. Cybis, *et al.*, Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *Ann. Appl. Stat.* **9** (2015).

48. N. R. Faria, M. A. Suchard, A. Rambaut, P. Lemey, Toward a quantitative understanding of viral phylogeography. *Current Opinion in Virology* **1**, 423–429 (2011).

49. P. Lemey, *et al.*, Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLoS Pathog* **10**, e1003932 (2014).

50. T. Bedford, S. Cobey, P. Beerli, M. Pascual, Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLoS Pathog* **6**, e1000918 (2010).

51. C. A. Russell, *et al.*, Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine* **26**, D31–D34 (2008).

52. L. B. Beheregaray, Twenty years of phylogeography: the state of the field and the challenges for the Southern Hemisphere. *Molecular Ecology*, ???-??? (2008).

53. L. Lu, A. J. Leigh Brown, S. J. Lycett, Quantifying predictors for the spatial diffusion of avian influenza virus in China. *BMC Evol Biol* **17**, 16 (2017).

54. B. D. Dalziel, *et al.*, Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. *Science* **362**, 75–79 (2018).

55. C. Viboud, *et al.*, Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science* **312**, 447–451 (2006).

56. K. M. Rich, A. Winter-Nelson, N. Brozović, Regionalization and foot-and-mouth disease control in South America: Lessons from spatial models of coordination and interactions. *The Quarterly Review of Economics and Finance* **45**, 526–540 (2005).

57. Y.-L. Nguyen, J. M. Kahn, D. C. Angus, Reorganizing Adult Critical Care Delivery: The Role of Regionalization, Telemedicine, and Community Outreach. *Am J Respir Crit Care Med* **181**, 1164–1169 (2010).

58. A. H. Seitzinger, P. L. Paarlberg, others, Regionalization of the 2014 and 2015 highly pathogenic Avian influenza outbreaks. *Choices* **31**, 1–8 (2016).

59. R. E. Kass, A. E. Raftery, Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795 (1995).

60. D. Magee, M. A. Suchard, M. Scotch, Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. *PLoS Comput Biol* **13**, e1005389 (2017).

61. K. Jaakkola, *et al.*, Decline in temperature and humidity increases the occurrence of influenza in cold climate. *Environ Health* **13**, 22 (2014).

62. C. S. Lutz, *et al.*, Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples. *BMC Public Health* **19**, 1659 (2019).

63. E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, M. V. Marathe, A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza Other Respi Viruses* **8**, 309–316 (2014).

64. C. Viboud, *et al.*, Association of influenza epidemics with global climate variability. *Eur J Epidemiol* **19**, 1055–1059 (2004).

65. J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, M. Lipsitch, Absolute Humidity and the Seasonal Onset of Influenza in the Continental United States. *PLoS Biol* **8**, e1000316 (2010).

66. A. Browne, S. St-Onge Ahmad, C. R. Beck, J. S. Nguyen-Van-Tam, The roles of transportation and transportation hubs in the propagation of influenza and coronaviruses: a systematic review. *Journal of Travel Medicine* **23**, tav002 (2016).

67. P. Bogner, I. Capua, D. J. Lipman, N. J. Cox, A global initiative on sharing avian flu data. *Nature* **442**, 981–981 (2006).

68. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

69. M. Kearse, *et al.*, Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).

70. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

71. O. Chernomor, *et al.*, Split diversity in constrained conservation prioritization using integer linear programming. *Methods Ecol Evol* **6**, 83–91 (2015).

72. I. Aksamentov, C. Roemer, E. Hodcroft, R. Neher, Nextclade: clade assignment, mutation calling and quality control for viral genomes. *JOSS* **6**, 3773 (2021).

73. B. Shapiro, A. Rambaut, A. J. Drummond, Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. *Molecular Biology and Evolution* **23**, 7–9 (2006).

74. A. J. Drummond, Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution* **22**, 1185–1192 (2005).

75. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution* **25**, 1459–1471 (2008).

76. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67**, 901–904 (2018).

77. M. Kulldorff, SaTScan—software for the spatial, temporal, and space-time scan statistics. 2016. *Boston: Harvard Medical School and Harvard Pilgrim Health Care* (2015).

78. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).

79. , HHS Regional Offices. *HHS Regional Offices* (2021).

80. F. Bielejec, *et al.*, SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol Biol Evol* **33**, 2167–2169 (2016).

81. A. Patil, D. Huard, C. J. Fonnesbeck, PyMC: Bayesian Stochastic Modelling in Python. *J Stat Softw* **35**, 1–81 (2010).

82. , Monthly Transportation Statistics. U.S. Department of Transportation's Bureau of Transportation Statistics. *Monthly Transportation Statistics* (2021).

83. U. C. Bureau, State Population Totals and Components of Change: 2010-2019 (2020).

84. G. Baele, P. Lemey, M. A. Suchard, Genealogical Working Distributions for Bayesian Model Testing with Phylogenetic Uncertainty. *Syst Biol* **65**, 250–264 (2016).

85. G. Baele, W. L. S. Li, A. J. Drummond, M. A. Suchard, P. Lemey, Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics. *Molecular Biology and Evolution* **30**, 239–243 (2012).

86. J. Parker, A. Rambaut, O. G. Pybus, Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution* **8**, 239–246 (2008).

87. E. Goldstein, C. Viboud, V. Charu, M. Lipsitch, Improving the Estimation of Influenza-Related Mortality Over a Seasonal Baseline. *Epidemiology* **23**, 829–838 (2012).

88. E. Azziz Baumgartner, *et al.*, Seasonality, Timing, and Climate Drivers of Influenza Activity Worldwide. *The Journal of Infectious Diseases* **206**, 838–846 (2012).

89. S. S. Lee, C. Viboud, E. Petersen, Understanding the rebound of influenza in the post COVID-19 pandemic period holds important clues for epidemiology and control. *International Journal of Infectious Diseases* **122**, 1002–1004 (2022).

90. V. Dhanasekaran, *et al.*, Human seasonal influenza under COVID-19 and the potential consequences of influenza lineage elimination. *Nat Commun* **13**, 1721 (2022).

91. T. C. Williams, I. Sinha, I. G. Barr, M. Zambon, Transmission of paediatric respiratory syncytial virus and influenza in the wake of the COVID-19 pandemic. *Eurosurveillance* **26** (2021).

92. E. C. Holmes, B. T. Grenfell, Discovering the Phylodynamics of RNA Viruses. *PLoS Comput Biol* **5**, e1000505 (2009).

93. Y. C. F. Su, *et al.*, Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun* **6**, 7952 (2015).

94. Z. Cai, T. Zhang, X.-F. Wan, A Computational Framework for Influenza Antigenic Cartography. *PLoS Comput Biol* **6**, e1000949 (2010).

95. T. Bedford, *et al.*, Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3**, e01914 (2014).

96. G. B. Cybis, *et al.*, Bayesian nonparametric clustering in phylogenetics: modeling antigenic evolution in influenza. *Statist. Med.* **37**, 195–206 (2018).

97. , CDC - Antigenic Characterization (2022).

98. T. Capblancq, B. R. Forester, Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods Ecol Evol* **12**, 2298–2309 (2021).

99. L. C. Katzelnick, *et al.*, Antigenic evolution of dengue viruses over 20 years. *Science* **374**, 999–1004 (2021).

100. , Candidate vaccine viruses. *World Health Organization*.

101. L. Kaufmann, *et al.*, An Optimized Hemagglutination Inhibition (HI) Assay to Quantify Influenza-specific Antibody Titers. *JoVE*, 55833 (2017).

102. J. C. Gower, Generalized procrustes analysis. *Psychometrika* **40**, 33–51 (1975).

103. P. Legendre, D. Borcard, P. R. Peres-Neto, ANALYZING BETA DIVERSITY: PARTITIONING THE SPATIAL VARIATION OF COMMUNITY COMPOSITION DATA. *Ecological Monographs* **75**, 435–450 (2005).

104. M. Koutsakos, A. K. Wheatley, K. Laurie, S. J. Kent, S. Rockman, Influenza lineage extinction during the COVID-19 pandemic? *Nat Rev Microbiol* **19**, 741–742 (2021).

105. D. Vijaykrishna, *et al.*, The contrasting phylodynamics of human influenza B viruses. *eLife* **4**, e05055 (2015).

106. B. J. Cowling, *et al.*, Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. *The Lancet Public Health* **5**, e279–e288 (2020).

107. W. W. Davis, J. A. Mott, S. J. Olsen, The role of non□pharmaceutical interventions on influenza circulation during the COVID□19 pandemic in nine tropical Asian countries. *Influenza Resp Viruses* **16**, 568–576 (2022).

108. Z. Qiu, *et al.*, The effectiveness of governmental nonpharmaceutical interventions against COVID-19 at controlling seasonal influenza transmission: an ecological study. *BMC Infect Dis* **22**, 331 (2022).

109. , Weekly U.S. Influenza Surveillance Report. *Centers for Disease Control and Prevention* (2022).

110. S. Wood, M. S. Wood, Package 'mgcv.' *R package version* **1**, 729 (2015).

111. C. Reed, *et al.*, Estimating Influenza Disease Burden from Population-Based Surveillance Data in the United States. *PLoS ONE* **10**, e0118369 (2015).

112. D. L. Noah, H. Hill, D. Hines, E. L. White, M. C. Wolff, Qualification of the Hemagglutination Inhibition Assay in Support of Pandemic Influenza Vaccine Licensure. *Clin Vaccine Immunol* **16**, 558–566 (2009).

113. G. L. Stewart, *et al.*, Rubella-Virus Hemagglutination-Inhibition Test. *N Engl J Med* **276**, 554–557 (1967).

114. W. Weis, *et al.*, Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* **333**, 426–431 (1988).

115. J. C. Pedersen, "Hemagglutination-Inhibition Assay for Influenza Virus Subtype Identification and the Detection and Quantitation of Serum Antibodies to Influenza Virus" in *Animal Influenza Virus*, Methods in Molecular Biology., E. Spackman, Ed. (Springer New York, 2014), pp. 11–25.

116. T. Rowe, *et al.*, Detection of Antibody to Avian Influenza A (H5N1) Virus in Human Serum by Using a Combination of Serologic Assays. *J Clin Microbiol* **37**, 937–943 (1999).

117. P. Kitikoon, P. C. Gauger, A. L. Vincent, “Hemagglutinin Inhibition Assay with Swine Sera” in *Animal Influenza Virus*, Methods in Molecular Biology., E. Spackman, Ed. (Springer New York, 2014), pp. 295–301.

118. A. C. Lowen, *et al.*, Blocking Interhost Transmission of Influenza Virus by Vaccination in the Guinea Pig Model. *J Virol* **83**, 2803–2818 (2009).

119. C.-Y. Wu, *et al.*, Influenza A surface glycosylation and vaccine design. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 280–285 (2017).

120. S. Truelove, *et al.*, A comparison of hemagglutination inhibition and neutralization assays for characterizing immunity to seasonal influenza A. *Influenza Other Respi Viruses* **10**, 518–524 (2016).

121. C. A. van Baalen, *et al.*, ViroSpot microneutralization assay for antigenic characterization of human influenza viruses. *Vaccine* **35**, 46–52 (2017).

122. Y. Lin, *et al.*, Optimisation of a micro-neutralisation assay and its application in antigenic characterisation of influenza viruses. *Influenza Other Respi Viruses* **9**, 331–340 (2015).

123. E. Spackman, I. Sitaras, “Hemagglutination Inhibition Assay” in *Animal Influenza Virus*, Methods in Molecular Biology., E. Spackman, Ed. (Springer US, 2020), pp. 11–28.

124. J. M. Fonville, *et al.*, Antibody landscapes after influenza virus infection or vaccination. *Science* **346**, 996–1000 (2014).

125. N. Mantel, The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research* **27**, 209–220 (1967).

126. S. Wilks, Racmacs: R Antigenic Cartography Macros (2022).

127. B. Q. Minh, *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530–1534 (2020).

128. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol* **4**, e88 (2006).

129. M. A. Suchard, *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4** (2018).

130. D. A. Jackson, PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Écoscience* **2**, 297–303 (1995).

131. O. Jari, *et al.*, vegan: Community Ecology Package. R package version 2.5-6 (2019).

132. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

133. P. Legendre, L. Legendre, Numerical Ecology 3rd edn, Vol. 24. *Developments in Environmental Modelling. Oxford, UK: Elsevier* (2012).

134. B. F. Koel, *et al.*, Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science* **342**, 976–979 (2013).

135. L. Glaser, *et al.*, A Single Amino Acid Substitution in 1918 Influenza Virus Hemagglutinin Changes Receptor Binding Specificity. *J Virol* **79**, 11533–11536 (2005).

136. Y. Wang, C. Y. Tang, X.-F. Wan, Antigenic characterization of influenza and SARS-CoV-2 viruses. *Anal Bioanal Chem* **414**, 2841–2881 (2022).

137. P. Legendre, M.-J. Fortin, Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data: SPATIAL ANALYSIS OF GENETIC DATA. *Molecular Ecology Resources* **10**, 831–844 (2010).

138. Ravina, *et al.*, A changing trend in diagnostic methods of Influenza A (H3N2) virus in human: a review. *3 Biotech* **11**, 87 (2021).

139. S. R. Mehta, S. A. Vinterbo, S. J. Little, Ensuring privacy in the study of pathogen genetics. *The Lancet Infectious Diseases* **14**, 773–777 (2014).

140. D. S. Johnson, C. Massey, A. O'Hara, The Opportunities and Challenges of Using Administrative Data Linkages to Evaluate Mobility. *The ANNALS of the American Academy of Political and Social Science* **657**, 247–264 (2015).

141. N. R. Faria, *et al.*, Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* **361**, 894–899 (2018).

142. P. Arora, Deepali, S. Varshney, Analysis of K-Means and K-Medoids Algorithm For Big Data. *Procedia Computer Science* **78**, 507–512 (2016).

143. J. Bang-Jensen, G. Gutin, A. Yeo, When the greedy algorithm fails. *Discrete Optimization* **1**, 121–127 (2004).

144. J. A. Rutigliano, *et al.*, Screening monoclonal antibodies for cross-reactivity in the ferret model of influenza infection. *Journal of Immunological Methods* **336**, 71–77 (2008).

145. S. E. Hensley, *et al.*, Influenza A Virus Hemagglutinin Antibody Escape Promotes Neuraminidase Antigenic Variation and Drug Resistance. *PLoS ONE* **6**, e15190 (2011).

146. M. R. Sandbulte, *et al.*, Discordant antigenic drift of neuraminidase and hemagglutinin in H1N1 and H3N2 influenza viruses. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20748–20753 (2011).

147. A. Wesolowski, *et al.*, Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11887–11892 (2015).

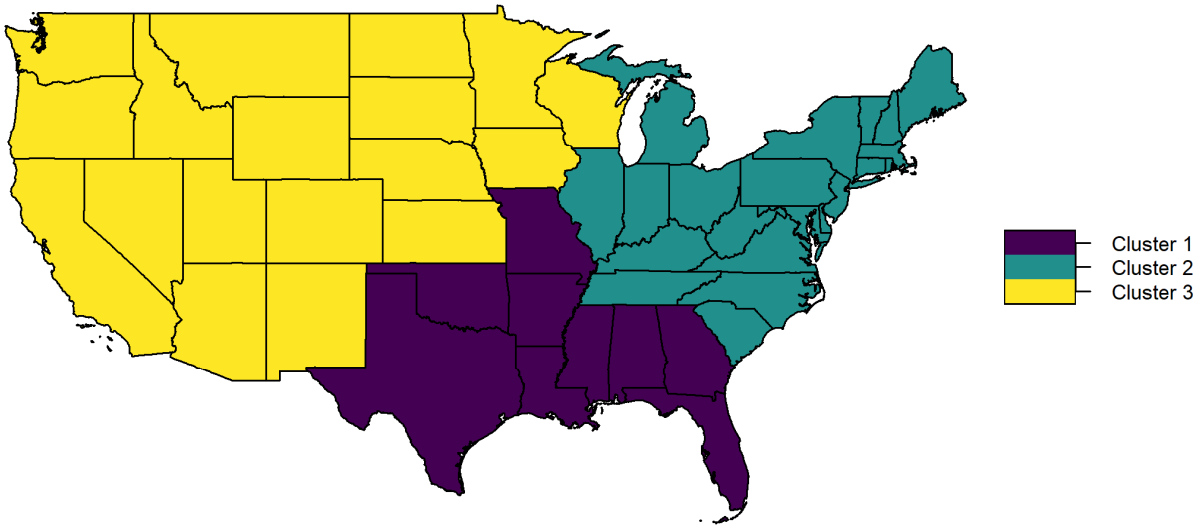
148. J. Hadfield, *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

149. M. S. Gill, P. Lemey, M. A. Suchard, A. Rambaut, G. Baele, Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction. *Molecular Biology and Evolution* **37**, 1832–1842 (2020).

Supplemental information Chapter 2: Analysis of seasonal H3N2 influenza diffusion in the United States

Tables

1. Table S2.1 - Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the HHS regional schema.
2. Table S2.2 - Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the U.S. Census Division regional schema.
3. Table S2.3 - Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the Louvain regional schema.
4. Table S2.4 - Multi-season phylogeny datasets, a sub-sample of 150 sequences were taken for each season using the PDA as well as two independent random samples of 150 sequences each season.
5. Table S2.5 - List of predictors for the GLM and their descriptions. See TableS5.csv.



Cluster 1 : Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Missouri, Oklahoma, Texas
Cluster 2 : Connecticut, Delaware, District of Columbia, Illinois, Indiana, Kentucky, Maine, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Rhode Island, South Carolina, Tennessee, Vermont, Virginia, West Virginia
Cluster 3 : Arizona, California, Colorado, Idaho, Iowa, Kansas, Minnesota, Montana, Nebraska, Nevada, New Mexico, North Dakota, Oregon, South Dakota, Utah, Washington, Wisconsin, Wyoming

Figure S2.1. Initial pass of community detection algorithm resulting in three distinct regions across the continental United States.

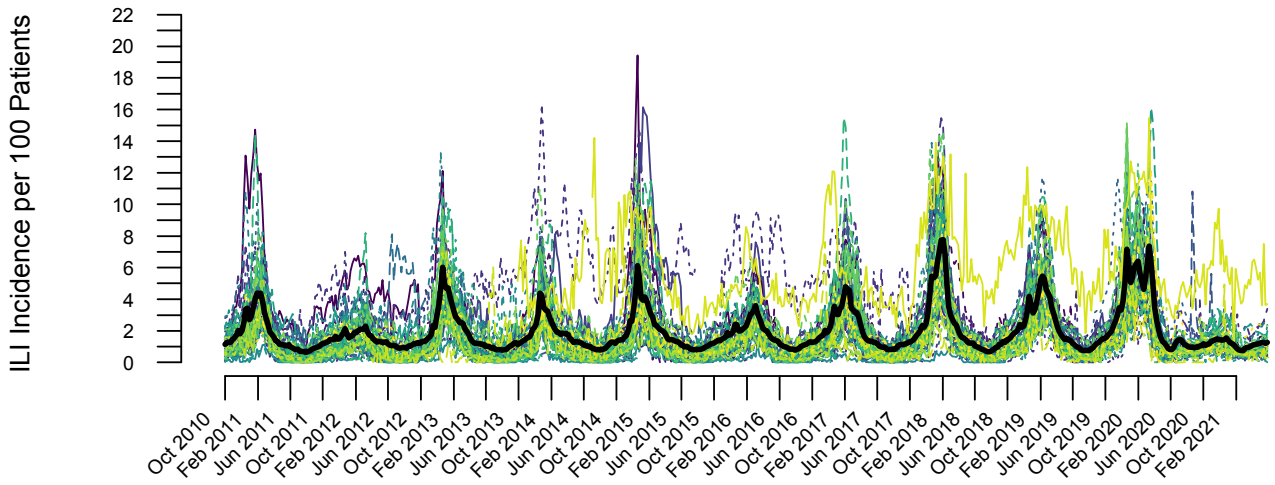


Figure S2.2. ILI incidence per 100 patients by state from October 2010 to February 2021. Each line represents the ILI data for a single U.S. State with the average of all states shown by the solid black line.

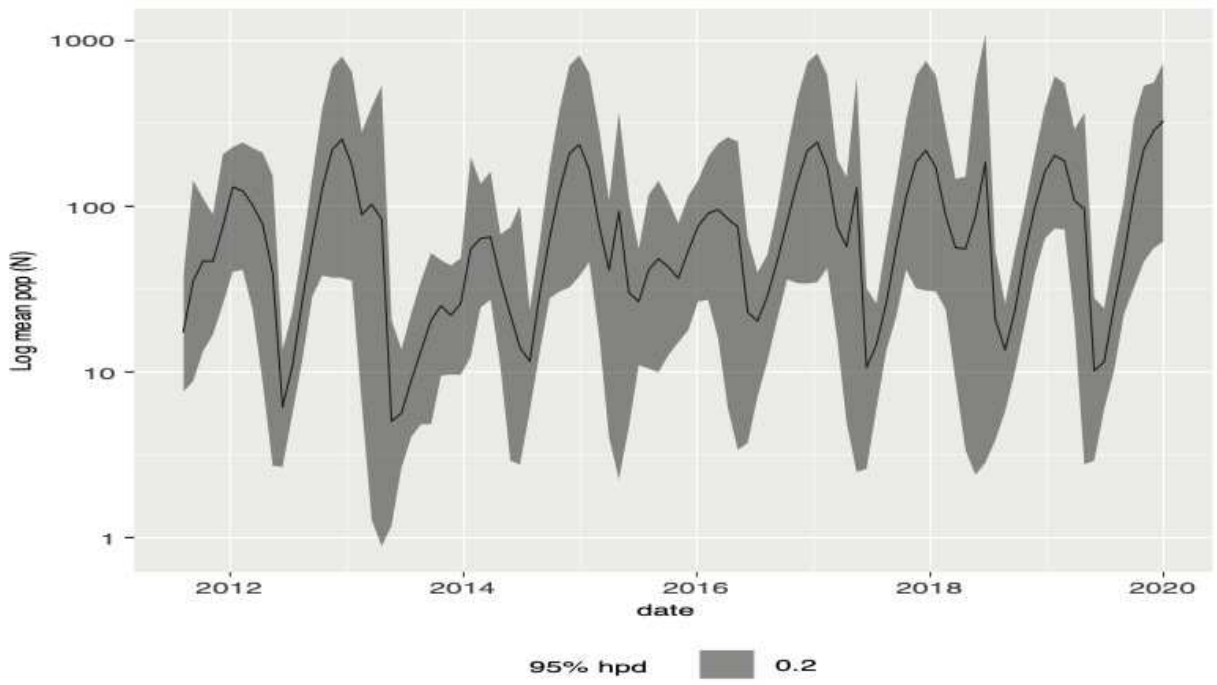


Figure S2.3. Effective population size estimated using the GMRF skyride reconstruction in BEAST. The solid black line represents the mean effective population size estimate, and the gray boarders indicate the 95% BCI.

Figure S2.4. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States. Isolates are colored by associated geographic metadata for U.S. Census Division region the isolate was collected in.

Figure S2.5. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States. Isolates are colored by associated geographic metadata for the HHS region the isolate was collected in.

Figure S2.6. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States. Isolates are colored by associated geographic metadata for the Louvain region the isolate was collected in.

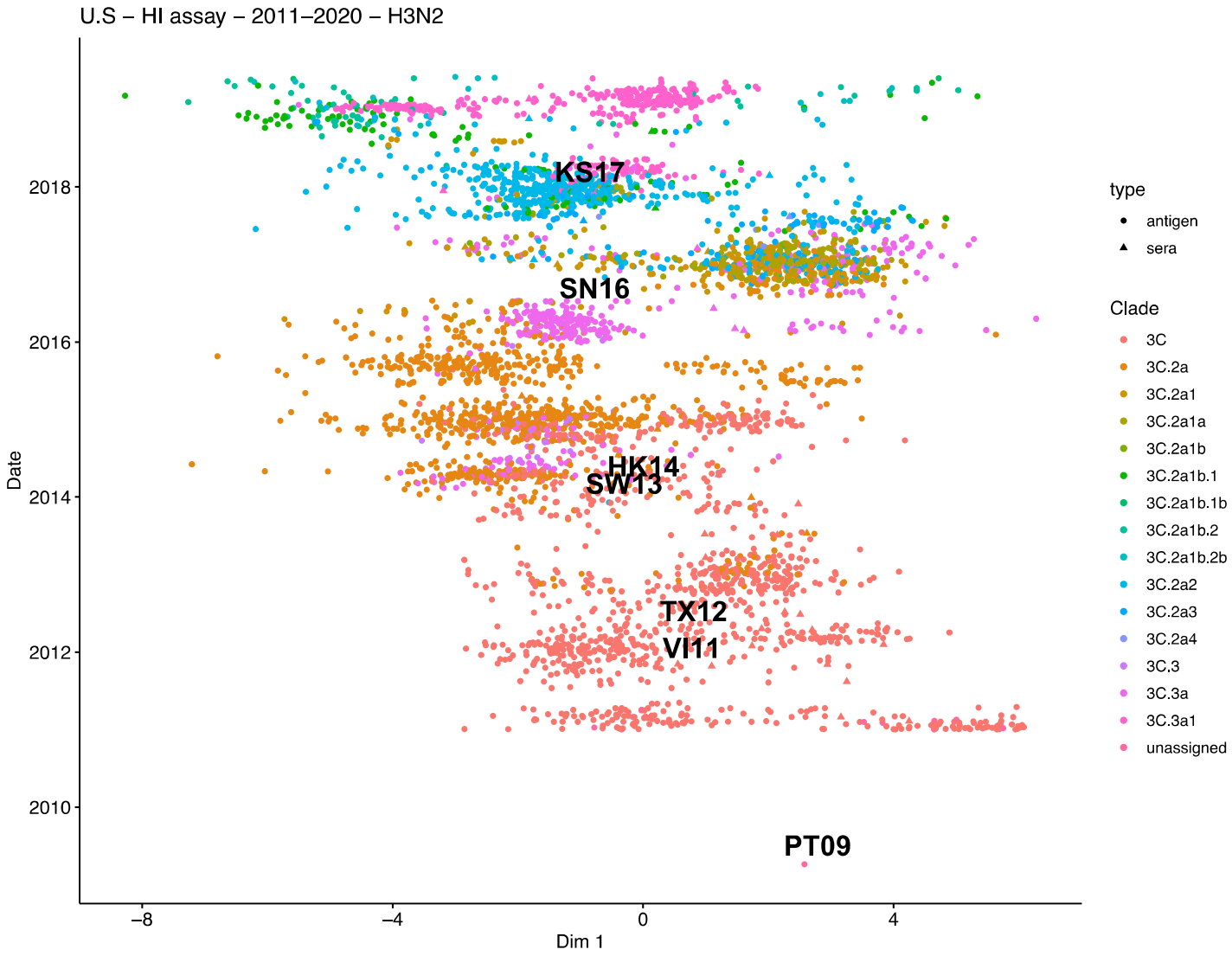


Figure S2.7. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States with isolates colored by respective clade determined using paired nucleotide data with the Nextclade program. The results of the 2D antigenic coordinates are used by plotting antigenic coordinate dimension 1 against time.

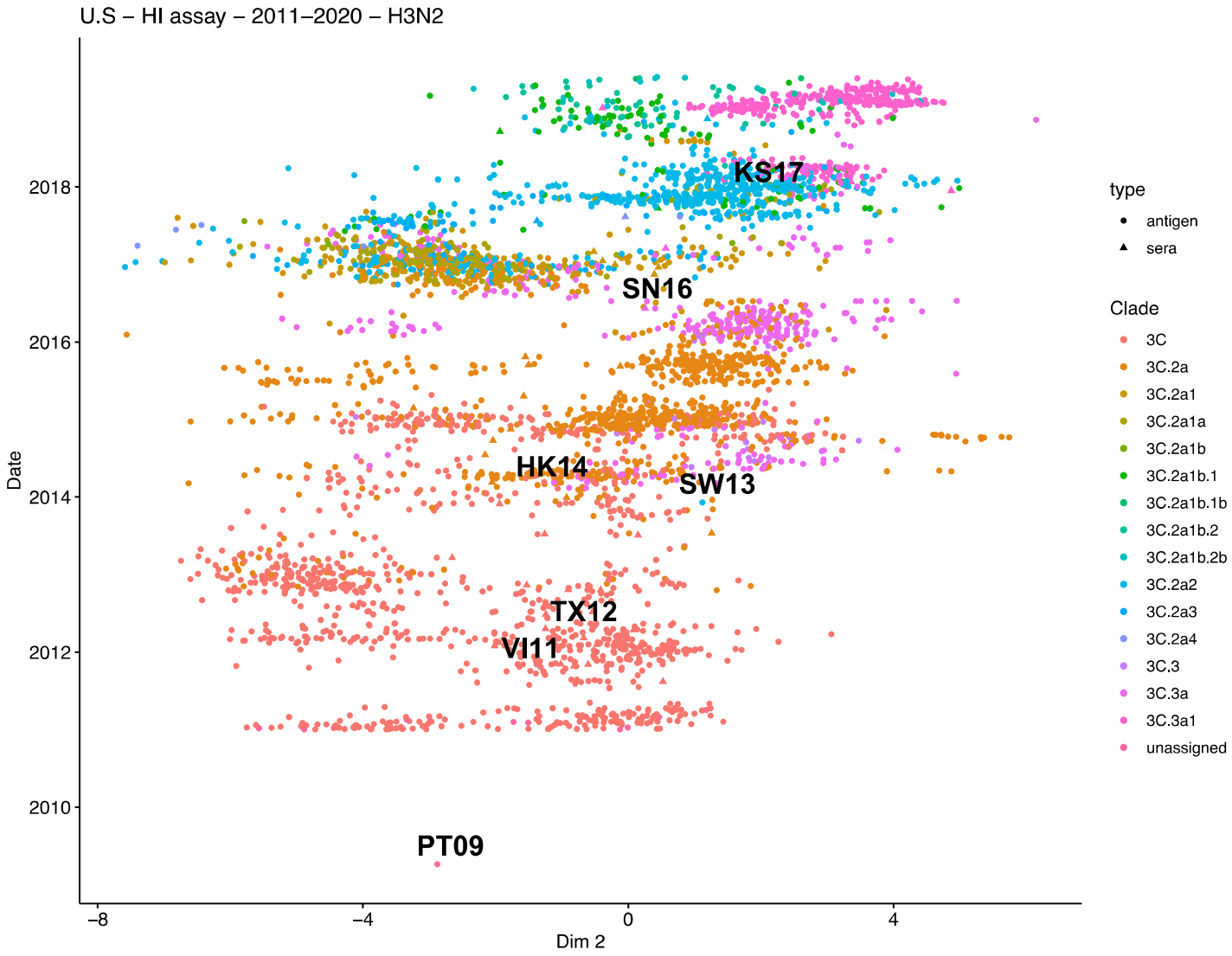


Figure S2.8. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States with isolates colored by respective clade determined using paired nucleotide data with the Nextclade program. The results of the 2D antigenic coordinates are used by plotting antigenic coordinate dimension 2 against time.

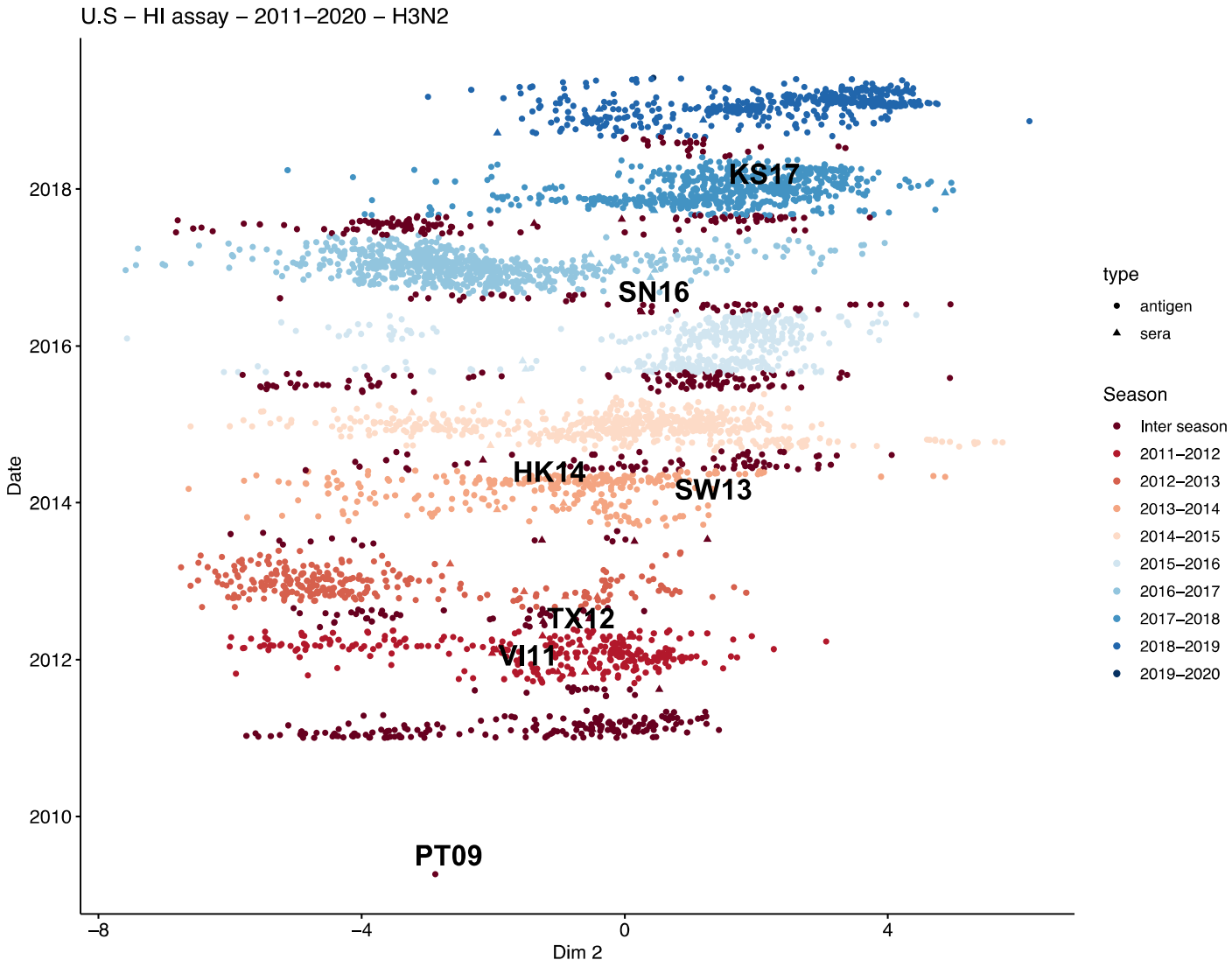


Figure S2.9. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States with isolates colored by the epidemic season the isolate was collected in. The results of the 2D antigenic coordinates are used by plotting antigenic coordinate dimension 2 against time.

Figure S2.10. Antigenic Cartography of H3N2 isolates collected between 2011-2020 in the United States with isolates colored by the epidemic season the isolate was collected in. The results of the 2D antigenic coordinates are used by plotting antigenic coordinate dimension 2 against time.

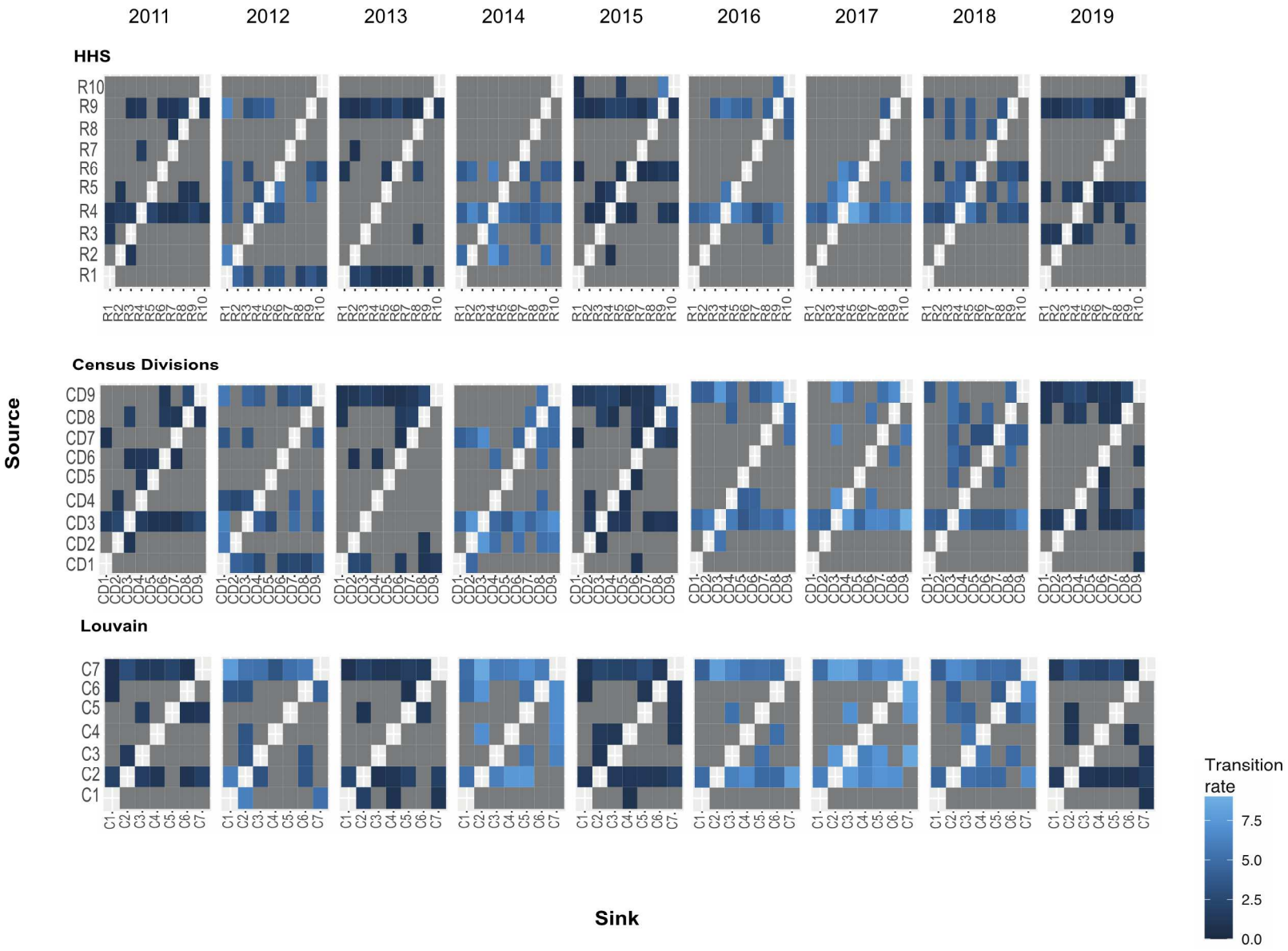


Figure S2.11. BSSVS for each epidemic season and regional schema from 2011 to 2020. The mean transition rate is colored for all rates with at least 50% posterior probability.

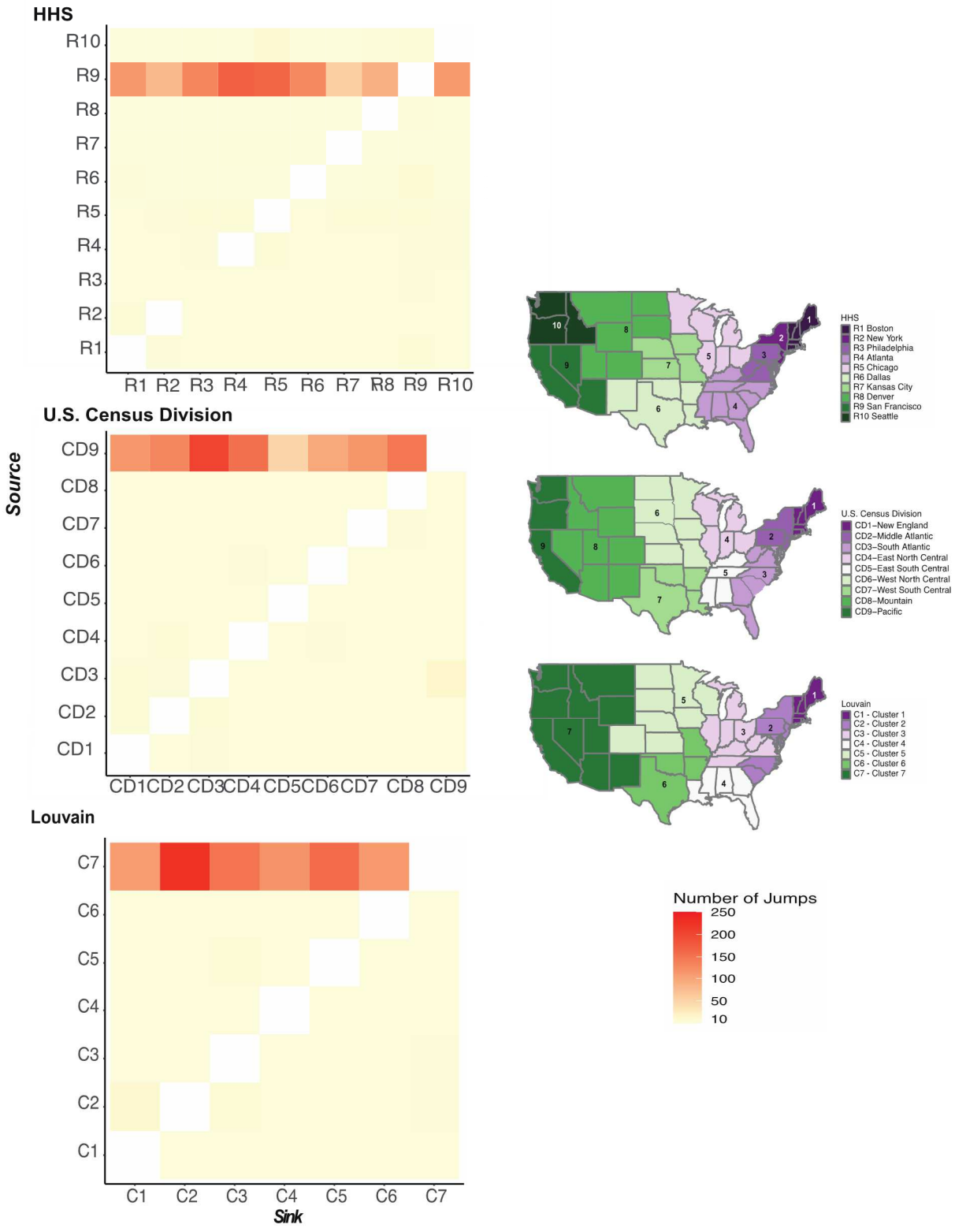


Figure S2.12 Jointly estimated Markov Jumps between geographic discrete trait schema for a subsample of H3N2 isolates in the United States between 2011-2020

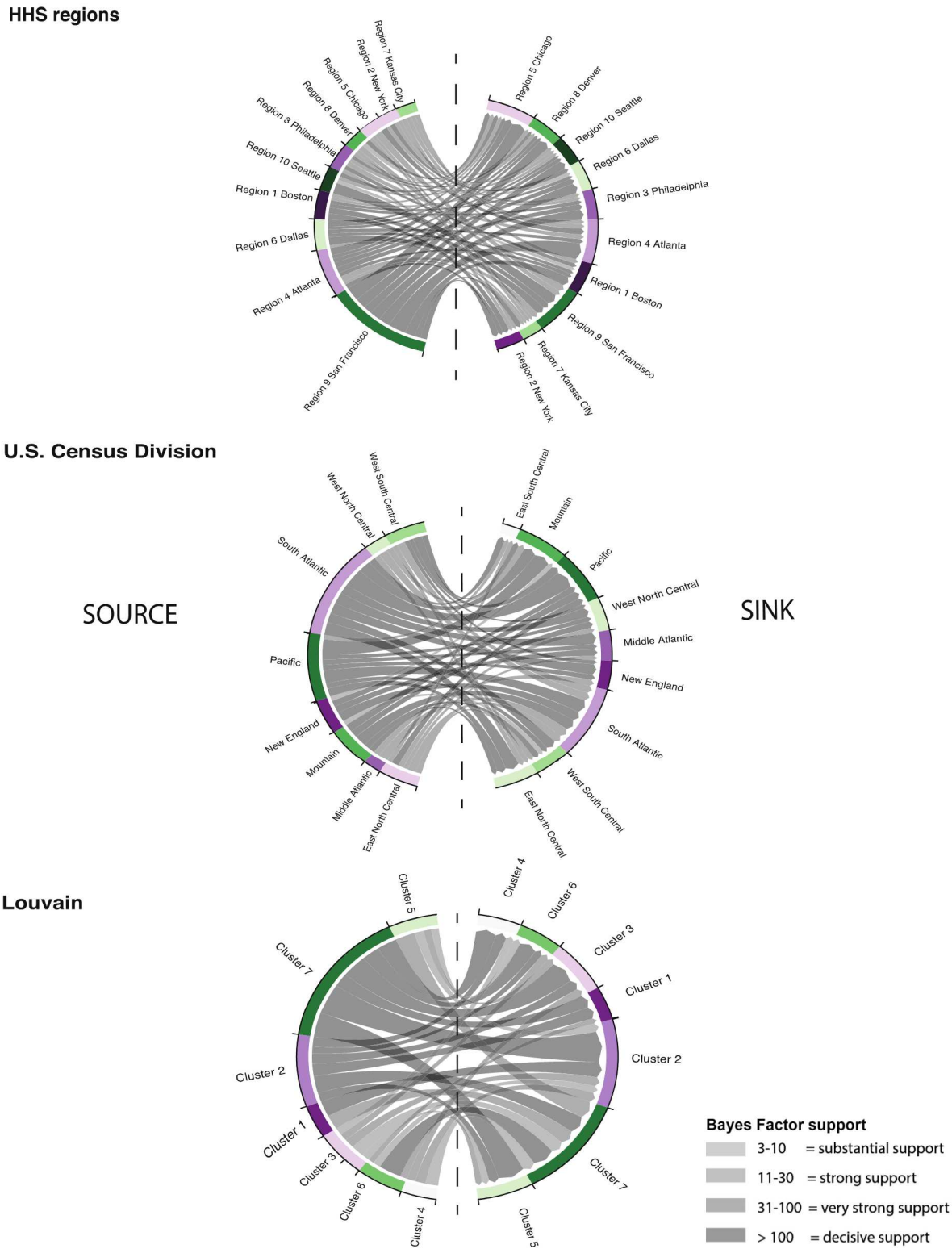


Figure S2.13. Results of jointly estimated BSSVS of epidemic seasons between 2011 and 2020 for each regional schema. The thickness of the chord corresponds to the discrete trait transition rate and the color corresponds to the Bayes Factor support. Color of source and sink bins correspond to the colors seen in Figure 1.

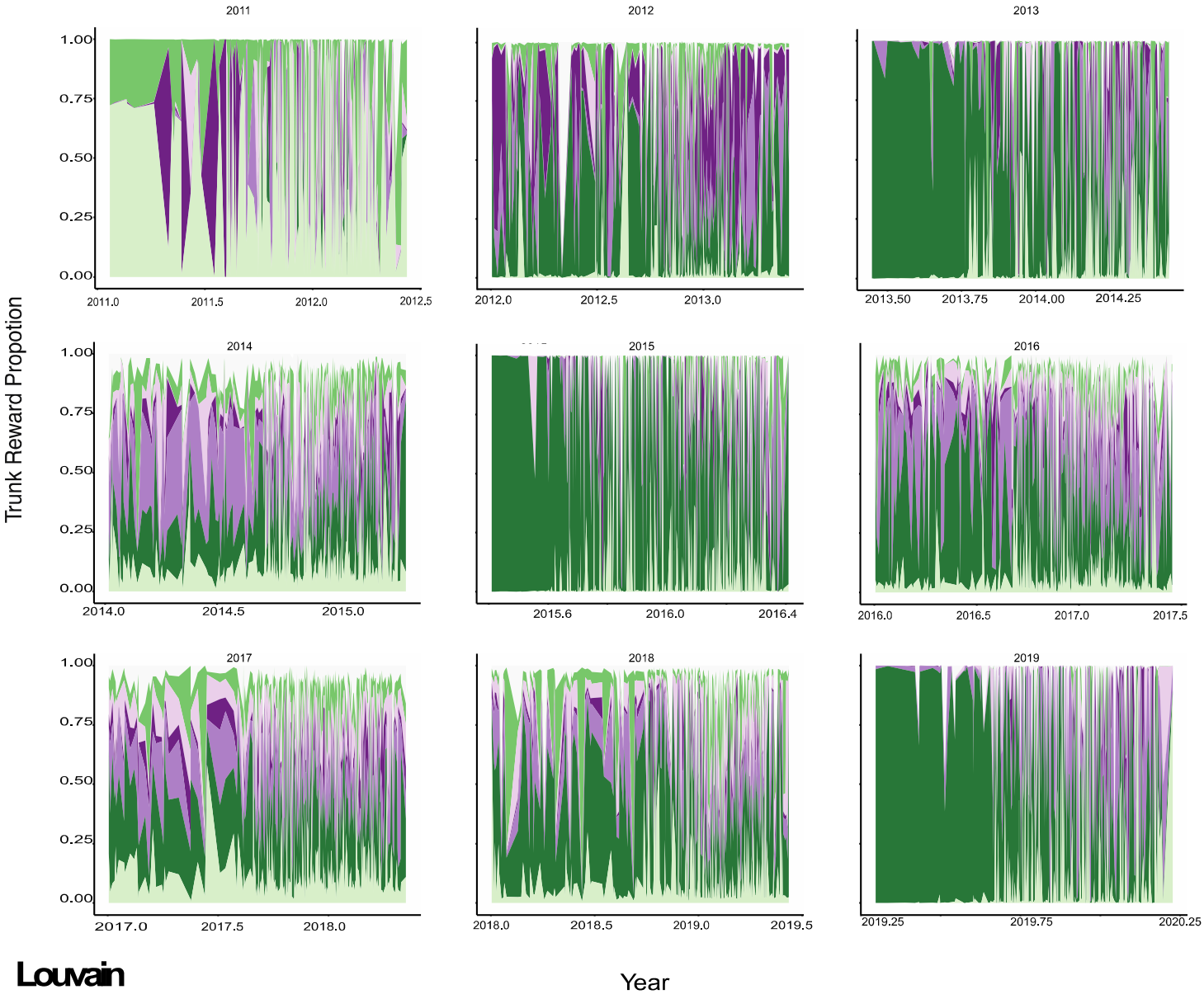


Figure S2.14 Markov reward trunk proportions for the Louvain regional schema for each epidemic season phylogeny estimated in BEAST. The color of chart area corresponds to the color of the Louvain regions as seen in Figure 1.

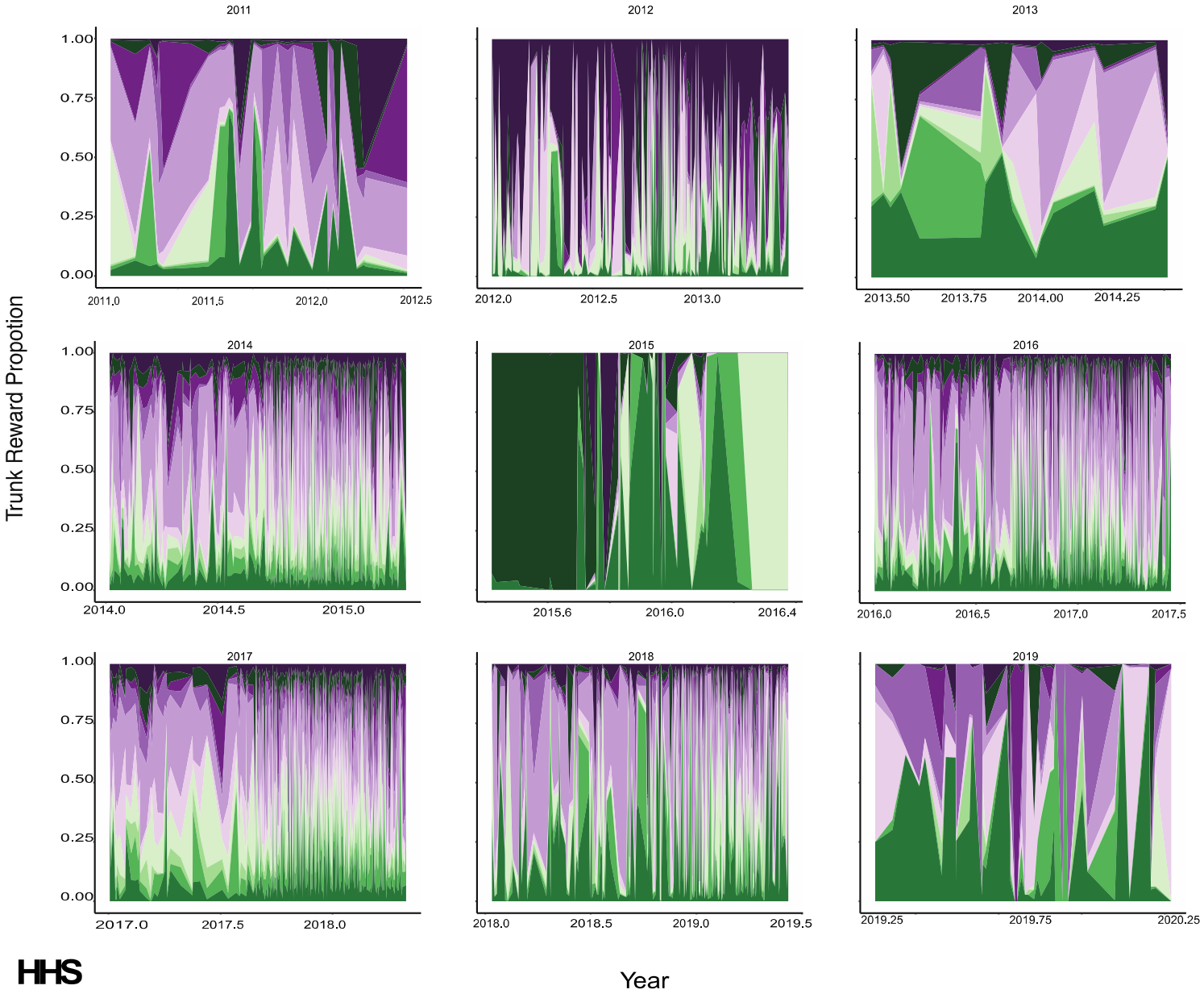


Figure S2.15 Markov reward trunk proportions for the HHS regional schema for each epidemic season phylogeny estimated in BEAST. The color of chart area corresponds to the color of the HHS regions as seen in Figure 1.

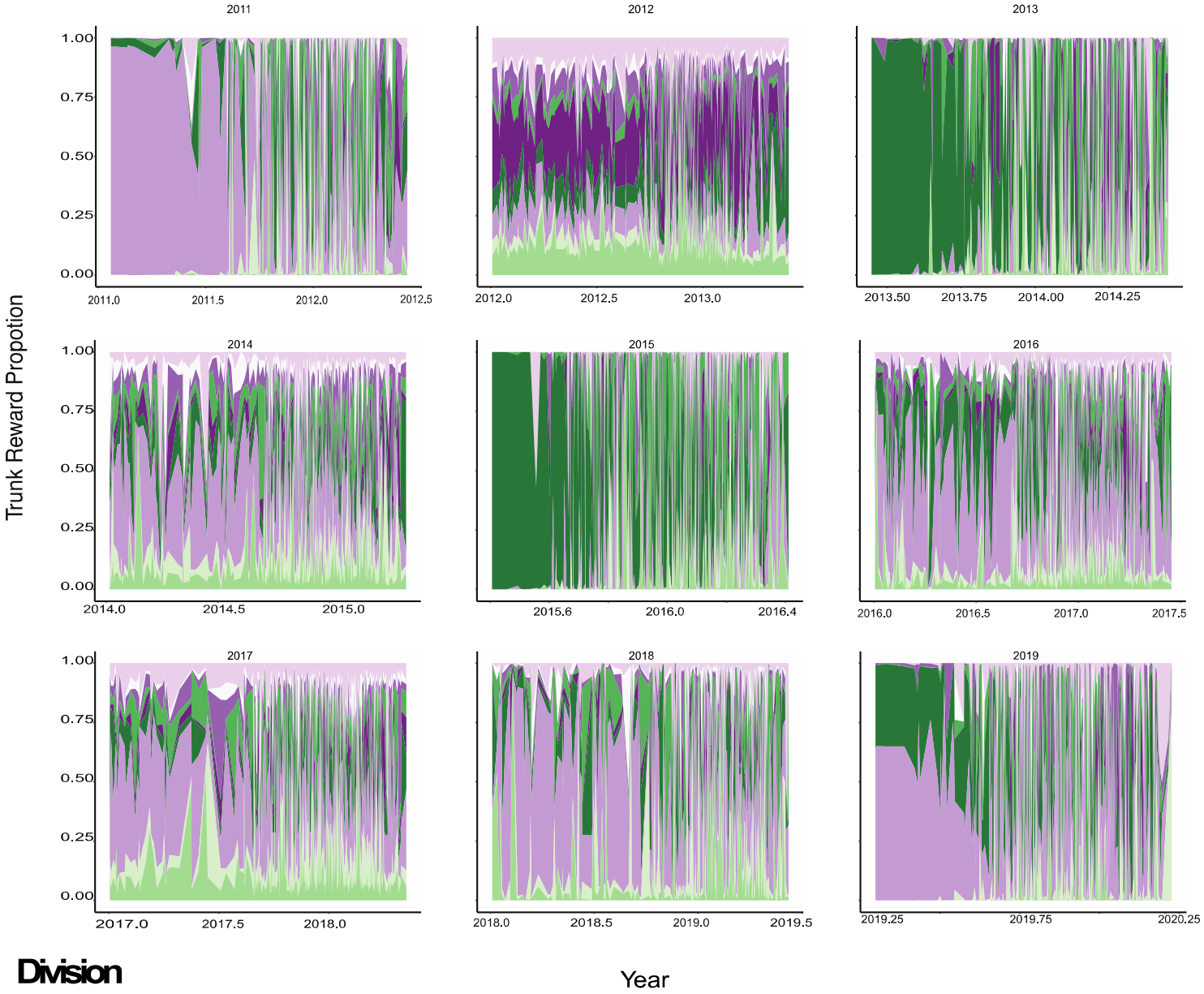


Figure S2.16 Markov reward trunk proportions for the US Census Division regional schema for each epidemic season phylogeny estimated in BEAST. The color of chart area corresponds to the color of the U.S. Census Divisions as seen in Figure 1.

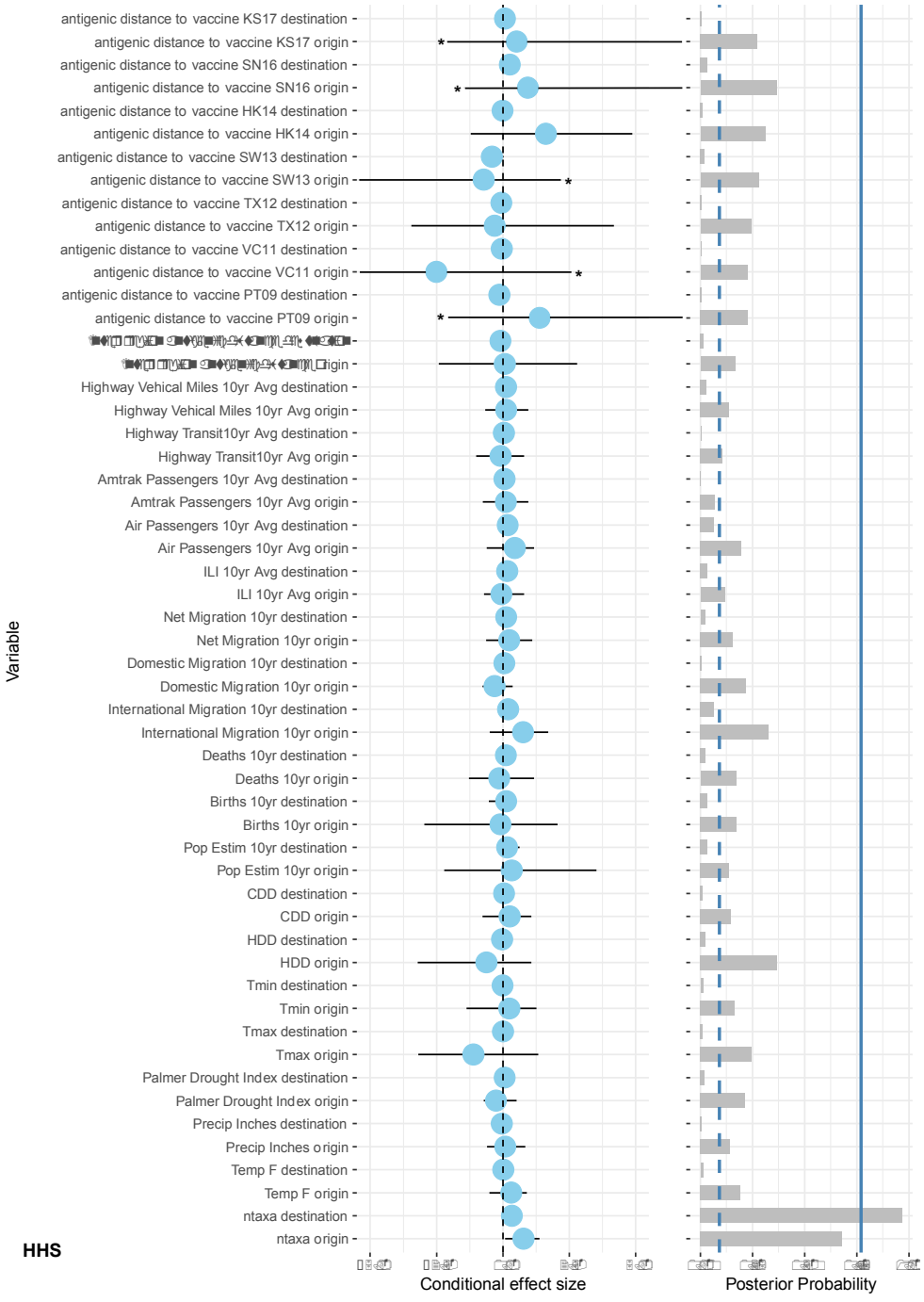


Figure S2.17 Results from Generalized Linear Model implemented in BEAST v 1.10.4 for co-variates of the diffusion process between regions of the HHS regional schema. The conditional effect size panel on the right indicates the level of inclusion for a given variable as a covariate for the diffusion process of the jointly estimated diffusion matrix. The posterior probability panel shows the level of posterior support for the inclusion of a given variable in the GLM. The solid blue line and dotted blue line in the posterior probability panel represent the calculated BF support equal 100 and 3 respectively. The * denotes coefficient HPDs that are greater than 5 and less than -5.

Figure S2.18 Results from Generalized Linear Model implemented in BEAST v 1.10.4 for co-variates of the diffusion process between regions of the U.S. Census Division regional schema. The conditional effect size panel on the right indicates the level of inclusion for a given variable as a covariate for the diffusion process of the jointly estimated diffusion matrix. The posterior probability panel shows the level of posterior support for the inclusion of a given variable in the GLM. The solid blue line and dotted blue line in the posterior probability panel represent the calculated BF support equal 100 and 3 respectively. The * denotes coefficient HPDs that are greater than 5 and less than -5.

Supplemental Tables

Statistic	observed.mean	null.mean	significance
AI	119.9642792 [116.4762039, 123.4885941]	132.282959 [130.907135, 133.9510498]	0.0
PS	888.5540161 [879, 898]	963.864624 [960.9660034, 969.2540283]	0.0
Region1-Boston	3.625999928 [3, 5]	2.289599895 [2.042000055, 2.951999903]	0.1
Region2-New_York	2.967999935 [2, 3]	1.981199861 [1.590000033, 2.937999964]	0.1
Region3-Philadelphia	3.052000046 [3, 3]	2.139400005 [1.988000035, 2.420000076]	0.1
Region4-Atlanta	3.154000044 [3, 4]	2.444000006 [2.140000105, 3.14199996]	0.2
Region5-Chicago	2.940000057 [2, 4]	2.296000004 [2.184000015, 2.444000006]	0.1
Region6-Dallas	2.364000082 [2, 3]	2.264199972 [2.069999933, 2.963999987]	1.0
Region7-Kansas_City	2.075999975 [2, 3]	1.593000174 [1.366000056, 1.95599997]	0.1
Region8-Denver	2.036000013 [2, 2]	1.826599836 [1.64199996, 2.036000013]	0.3
Region9-San_Francisco	3.977999926 [3, 4]	2.521000147 [2.313999891, 3.023999929]	0.1
Region10-Seattle	2.674000025 [2, 3]	2.300600052 [2.104000092, 3.503999949]	0.2

Table S2.1 Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the HHS regional schema.

Statistic	observed.mean	null.mean	significance
AI	117.8134308 [114.1977158, 121.5040436]	129.285202 [128.7859955, 129.8312225]	0.0
PS	862.5059814 [852, 872]	928.2871094 [923.1339722, 935.7700195]	0.0
Mountain	3.252000093 [3, 4]	2.253400087 [2.079999924, 2.700000048]	0.1
South_Atlantic	3.128000021 [3, 4]	2.606200218 [2.411999941, 3.186000109]	0.3
East_North_Central	2.436000109 [2, 4]	2.368599892 [2.138000011, 3]	1.0
West_North_Central	2.130000114 [2, 3]	2.087800026 [1.758000016, 2.971999884]	0.8
Pacific	3.170000076 [3, 4]	2.966399908 [2.713999987, 3.319999933]	0.7
Middle_Atlantic	3.002000093 [2, 4]	2.308399916 [2.053999901, 3]	0.2
New_England	3.625999928 [3, 5]	2.02519989 [1.925999999, 2.21600008]	0.1
East_South_Central	2.007999897 [2, 2]	1.565799952 [1.212000012, 2]	0.2
West_South_Central	2.278000116 [2, 3]	2.091599941 [1.972000003, 2.746000051]	0.9

Table S2.2 Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the U.S. Census Division regional schema.

Statistic	observed.mean	null.mean	significance
AI	108.191185 [104.5929718, 111.5136337]	120.5653534 [119.4020233, 121.8480148]	0.0
PS	773.8179932 [764, 783]	837.5040283 [834.2800293, 842.2160034]	0.0
Cluster1	3.088000059 [3, 4]	2.087799788 [2, 2.292000055]	0.1
Cluster2	2.25 [2, 3]	2.259799957 [2.036000013, 4]	1.0
Cluster3	3.247999907 [3, 4]	2.319400072 [2.174000025, 2.736000061]	0.1
Cluster4	3.625999928 [3, 5]	2.128599644 [1.993999958, 2.395999908]	0.1
Cluster5	4.096000195 [3, 5]	3.112999916 [2.49000001, 4.188000202]	0.3
Cluster6	4.602000237 [4, 6]	3.905199766 [3.467999935, 5.18599987]	0.4
Cluster7	2.565999985 [2, 4]	2.258599758 [2.052000046, 2.526000023]	1.0

Table S2.3 Results of BaTS. AI: Association Index, FP: Fitch Parsimony score. The maximum exclusive single state clade size is shown for each region in the Louvain regional schema.

X	tmcra	mrsd	sub.rate
Pda	2010.392	2020.227	0.00459
Rand1	2010.785	2020.227	0.00366
Rand2	2010.793	2020.216	0.00383

Table S2.4 Multi-season phylogeny datasets, a sub-sample of 150 sequences were taken for each season using the PDA as well as two independent random samples of 150 sequences each season.

Predictor	Type	Description
Mean Temp F	Climate	Average mean temperature (F) for the influenza season time period
Mean Precip Inches	Climate	Average precipitation in inches for the influenza season time period
Mean Palmer Drought Index	Climate	Average Palmer drought index for the influenza season time period
Mean Tmax	Climate	Average max temperature (F) for the influenza season time period
Mean Tmin	Climate	Average min temperature (F) for the influenza season time period
Mean HDD	Climate	Average Heating degree days (HDD) for influenza season time period. HDD is a measure of how cold the temperature was on a given day or during a period of days (ela.gov)
Mean CDD	Climate	Average Cooling degree days (CDD) for influenza season time period. CDD is a measure of how hot the temperature was on a given day or during a period of days.(ela.gov)
Pop Estim 10yr	Demographic	Annual Population Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
Births 10yr	Demographic	Birth Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
Deaths 10yr	Demographic	Death Estimates, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
International Migration 10yr	Demographic	Net international migration in period, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
Domestic Migration 10yr	Demographic	Net domestic migration in period, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
Net Migration 10yr	Demographic	Net migration in period, Estimated Components of Resident Population Change, and Rates of the Components of Resident Population Change for the United States, States, and Puerto Rico: April 1, 2010 to July 1, 2019, File: 7/1/2019 National and State Population Estimates, Source: U.S. Census Bureau, Population Division, Release Date: December 2019
ILI 10yr Avg	Epidemiologic	Influenza like illness data downloaded from CDC Fluview.Epidemiological data for cases within the United States.

Air Passengers 10yr Avg	Transportation	Number of air passengers in the United States (Domestic flights)
Amtrak Passengers 10yr Avg	Transportation	Number of air passengers in the United States (Domestic flights)
Highway Transit 10yr Avg	Transportation	Number of air passengers in the United States (Domestic flights)
Highway Vehicle Miles 10y rAvg	Transportation	Number of air passengers in the United States (Domestic flights)
Inter-region Antigenic distance	Antigenic	Average Euclidean distance from an antigenic cartography between all antigens from a given geographic region to another.
antigenic distance to vaccine VC11	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Victoria 2011
antigenic distance to vaccine TX12	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Texas 2012
antigenic distance to vaccine SW13	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Switzerland 2013
antigenic distance to vaccine SN16	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Singapore 2016
antigenic distance to vaccine PT09	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Perth 2009
antigenic distance to vaccine KS17	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Kansas 2017
antigenic distance to vaccine HK14	Antigenic	Average Euclidean distance from an antigenic cartography for all antigens in a given geographic region to the vaccine candidate Hong Kong 2014

Table S2.4 List of predictors for the GLM and their descriptions

Supplemental information Chapter 3: Genetic and antigenic characterization of global seasonal influenza evolution, 2017-2022.

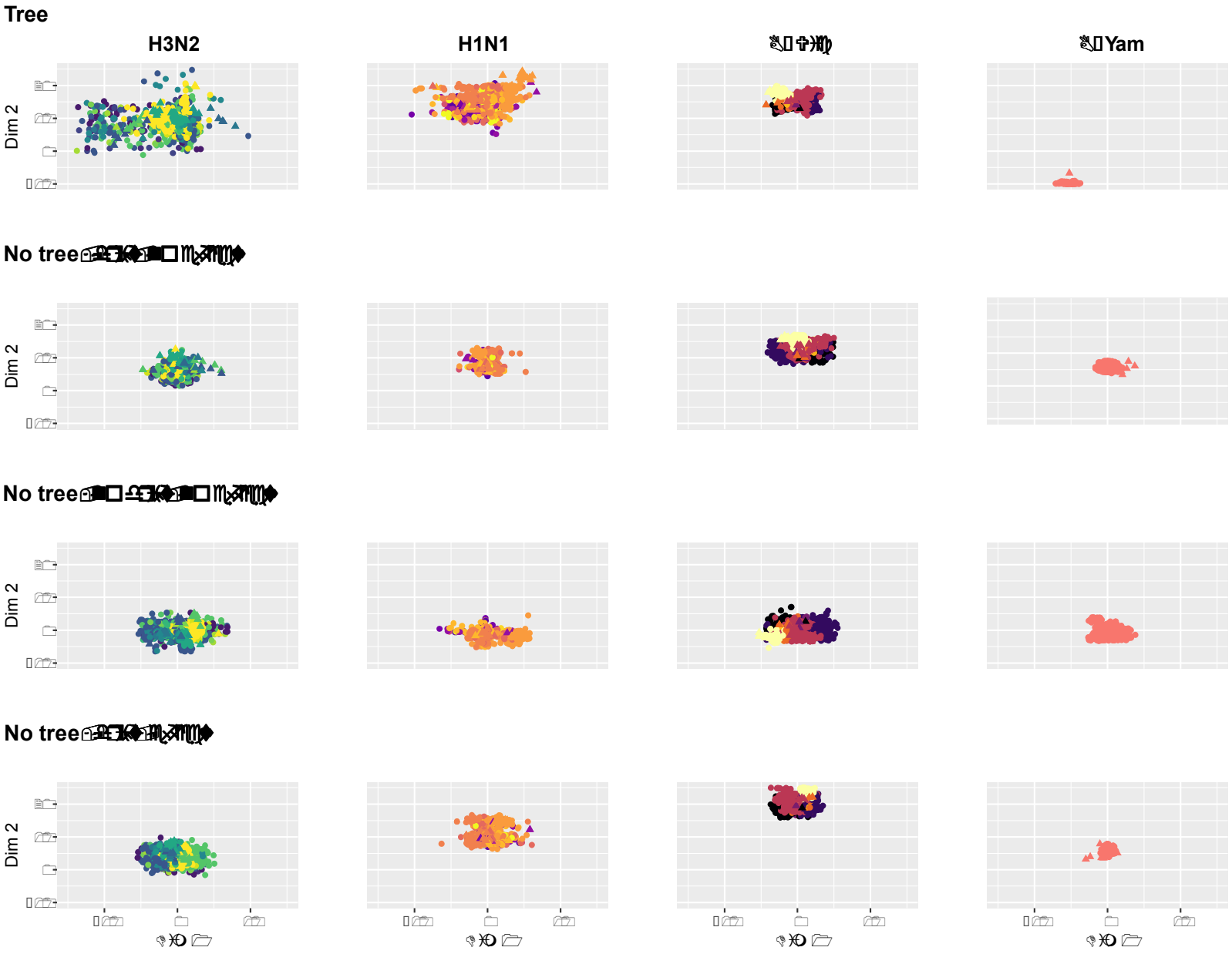


Figure S3.1. BMDS antigenic cartographies estimated for each subtype and evolutionary model. The color of isolates corresponds to the clade determined through NextClade. The shape of points, circle, and triangle, corresponds to viruses and sera respectively.

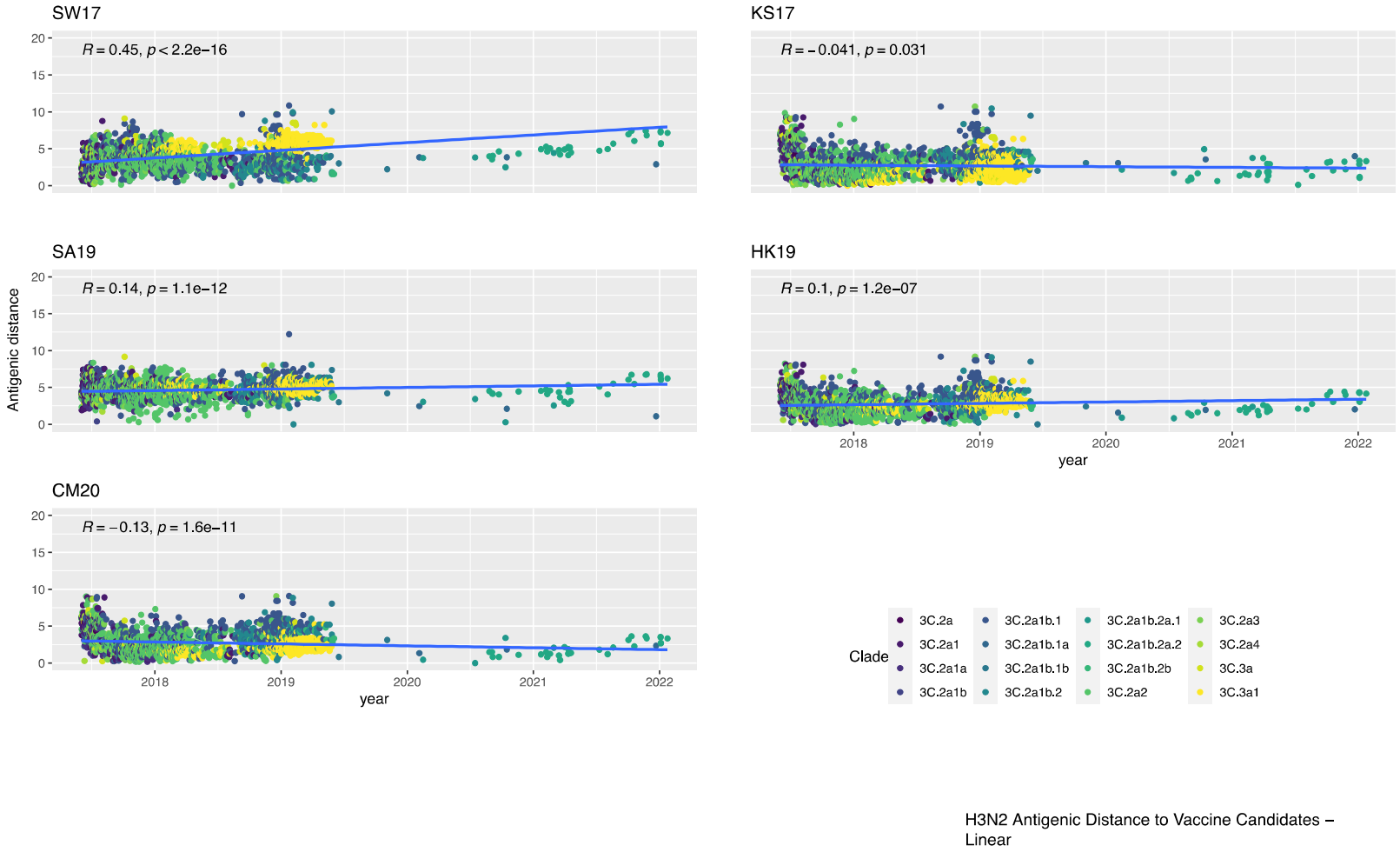


Figure S3.2. Antigenic distance of H3N2 isolates from vaccine candidates over time. The Euclidean distance of each isolate in antigenic units calculated from the corresponding subtype cartography was used to estimate distances. The coloring of isolates corresponds to the NextClade designation based on isolate genomic data.

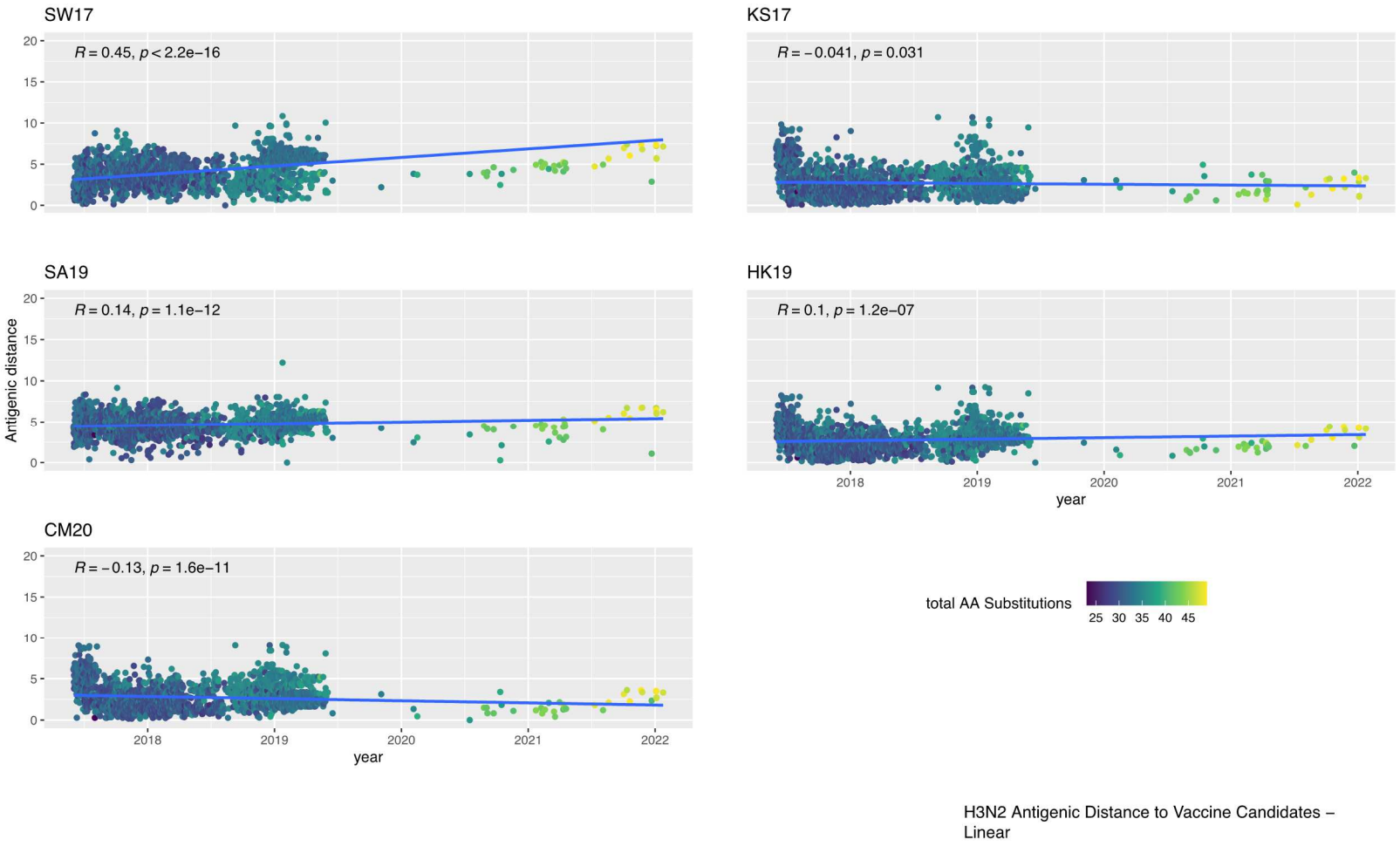


Figure S3.3. Antigenic distance of H3N2 isolates from vaccine candidates over time. The Euclidean distance of each isolate in antigenic units calculated from the corresponding subtype cartography was used to estimate distances. The coloring of isolates corresponds the number of amino acid substitutions from NextClade reference.

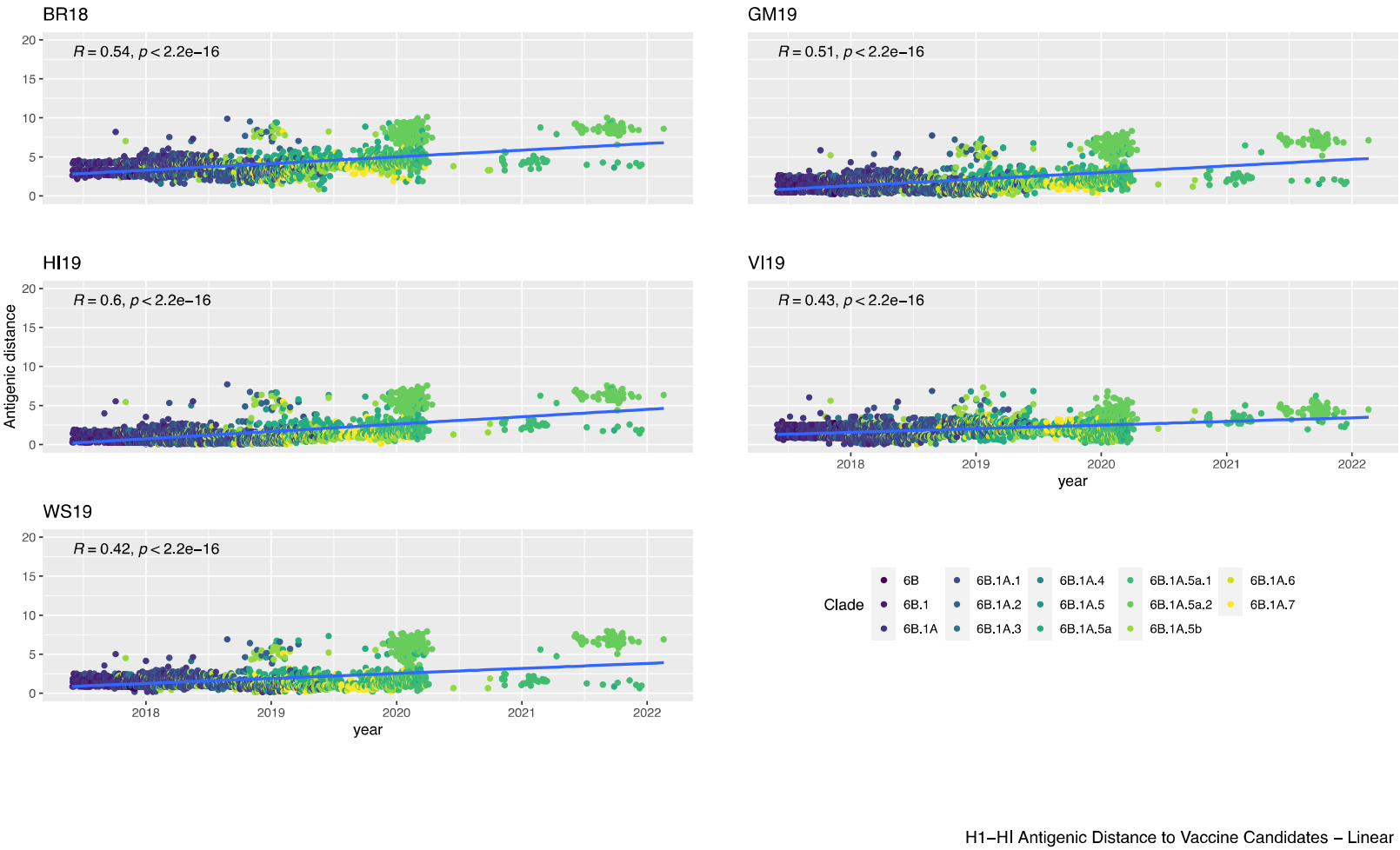


Figure S3.4. Antigenic distance of H1N1 isolates from vaccine candidates over time. The Euclidean distance of each isolate in antigenic units calculated from the corresponding subtype cartography was used to estimate distances. The coloring of isolates corresponds to the NextClade designation based on isolate genomic data.

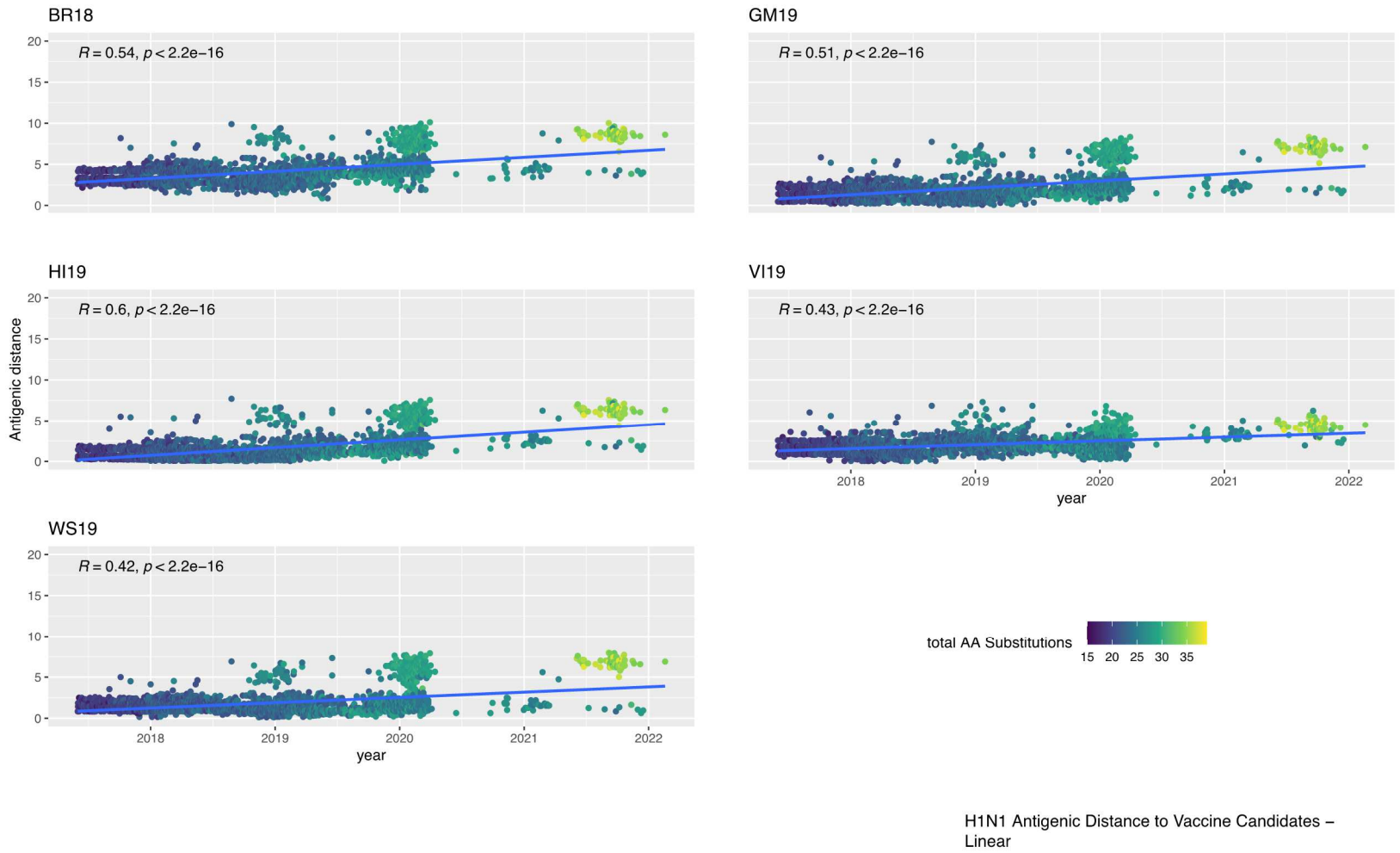


Figure S3.5. Antigenic distance of H1N1 isolates from vaccine candidates over time. The Euclidean distance of each isolate in antigenic units calculated from the corresponding subtype cartography was used to estimate distances. The coloring of isolates corresponds the number of amino acid substitutions from NextClade reference.

Figure S3.6. Antigenic distance of B-Vic isolates from vaccine candidate WA19 over time. The Euclidean distance of each isolate in antigenic units calculated from the corresponding subtype cartography was used to estimate distances. The coloring of isolates in the left panel corresponds to the NextClade designation based on isolate genomic data. The coloring of isolates in the right panel corresponds the number of amino acid substitutions from NextClade reference.

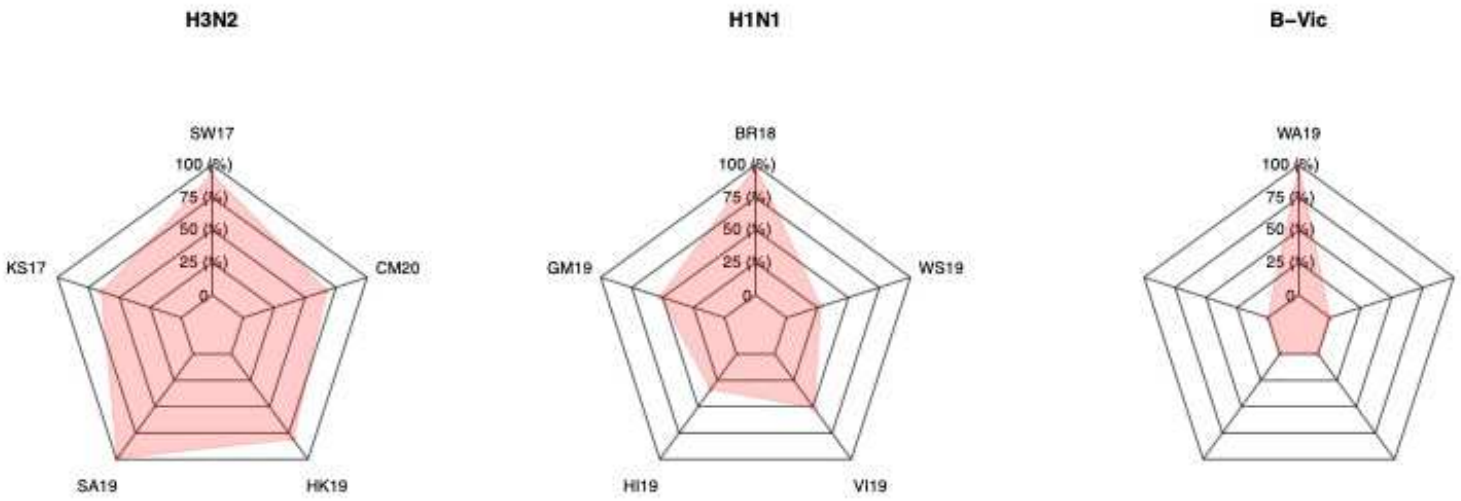


Figure S3.7. Proportion of all isolates that are vaccine escapes to a given vaccine candidate over study space. The proportion of isolates with an antigenic distance greater than 2 units from the vaccine strain is reported for all isolates with paired genetic data characterized between 2017 – 2022.

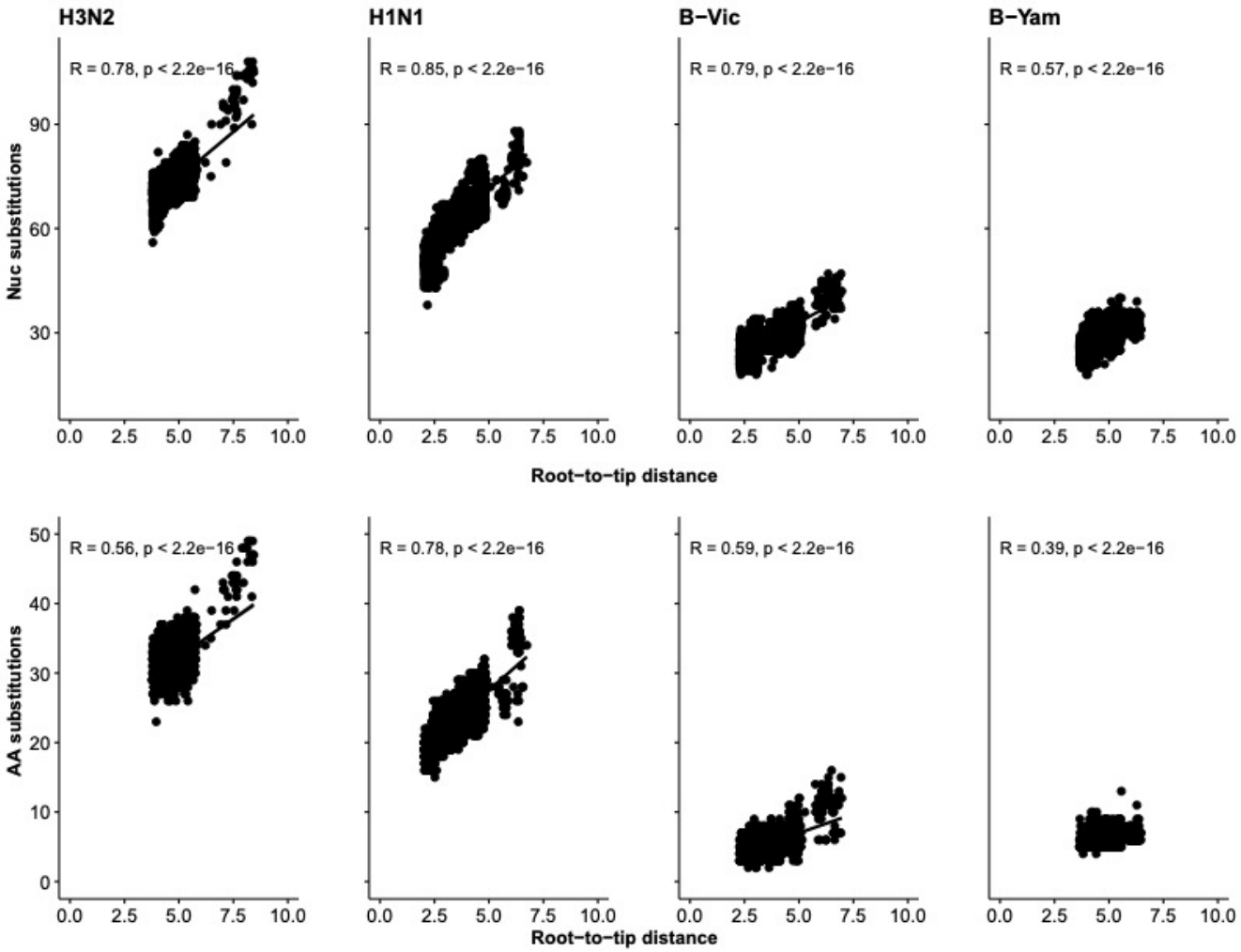


Figure S3.8. The number of nucleotide substitutions vs the root to tip distance for isolates of each subtype in the study space with paired antigenic/genomic data in the top panel. The number of amino acid substitutions vs the root to tip distance for isolates can be seen in the bottom panel.

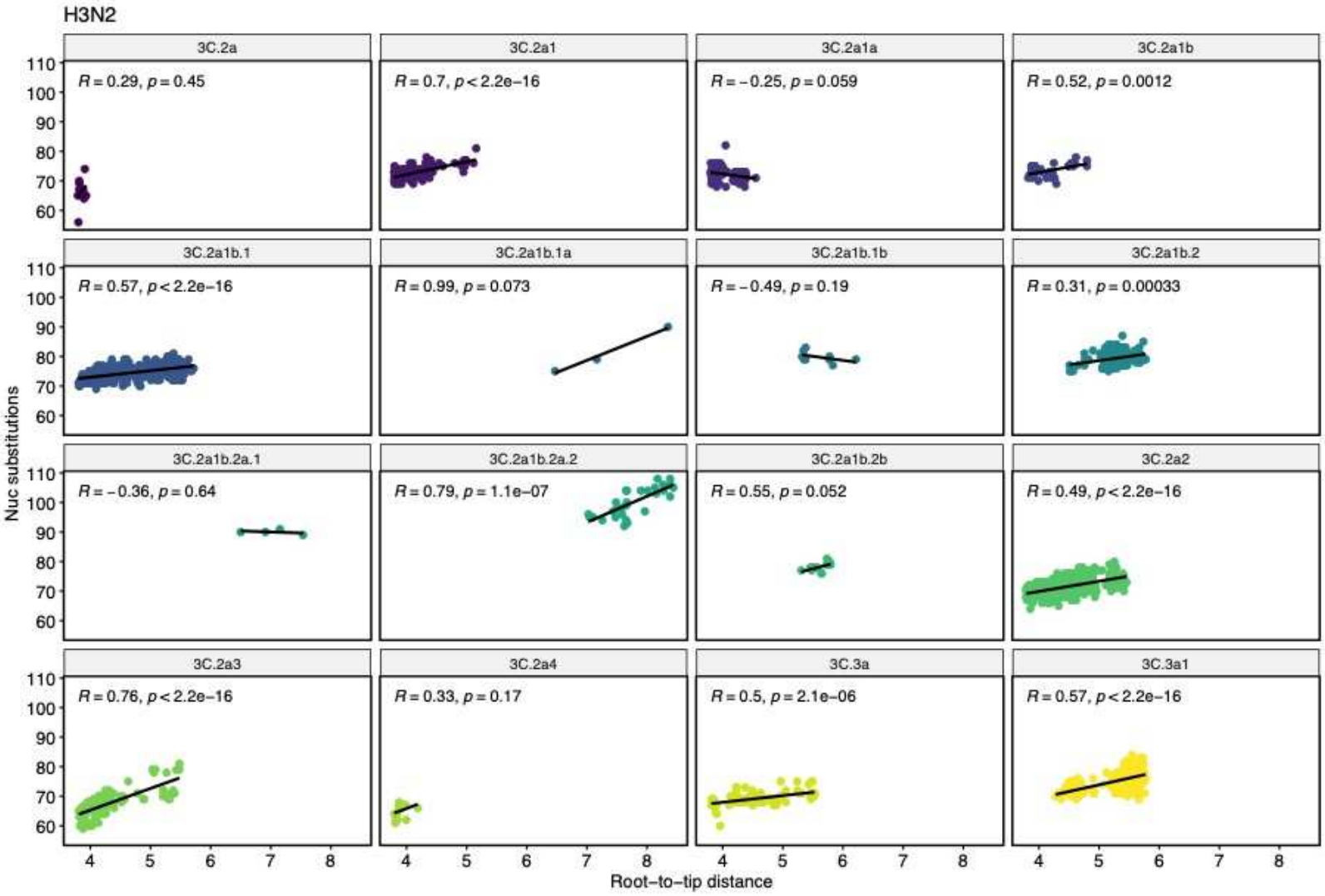


Figure S3.9. The number of nucleotide substitutions vs the root to tip distance for isolates of H3N2 with paired antigenic/genomic data for each NextClade lineage in the study space.

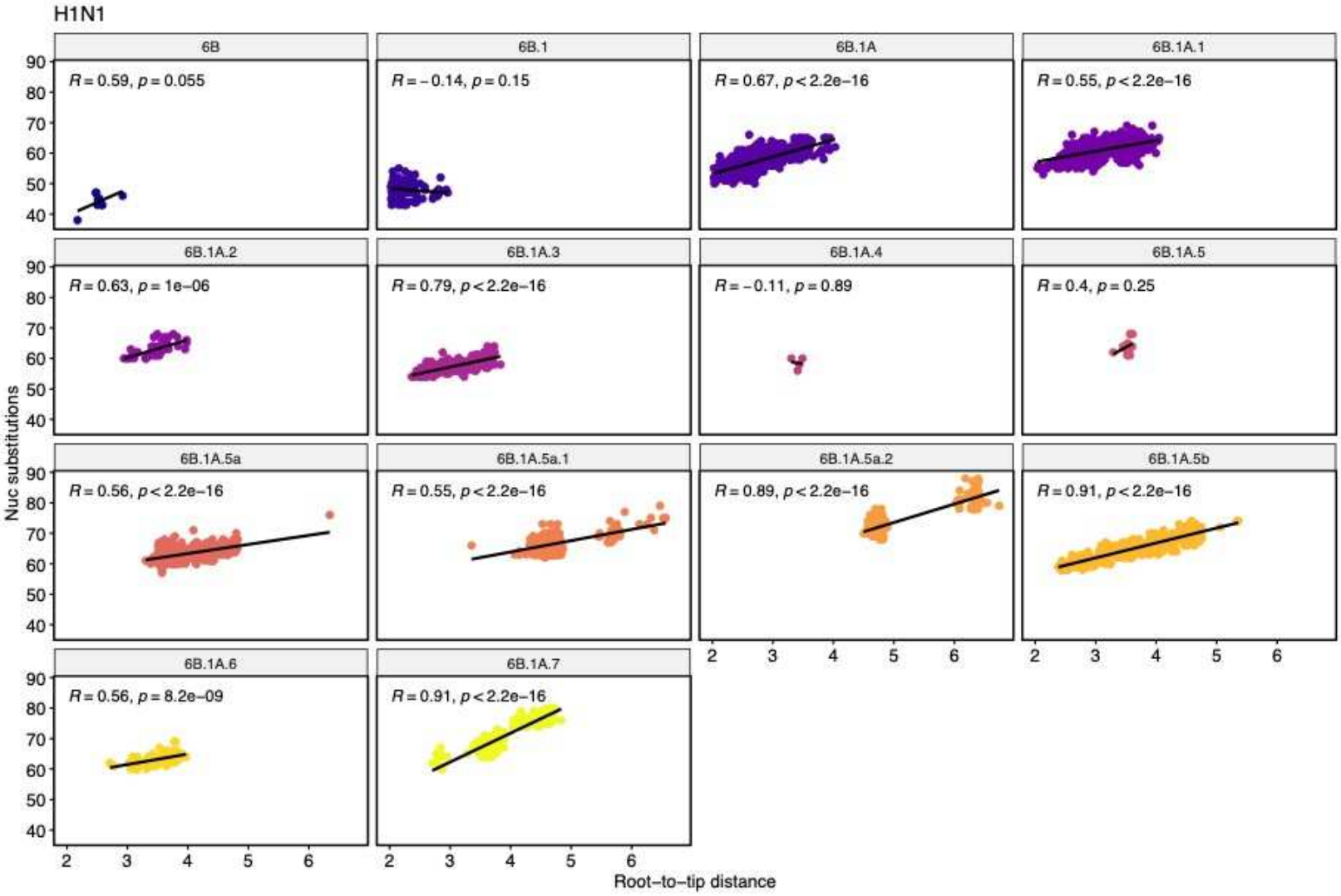


Figure S3.10. The number of nucleotide substitutions vs the root to tip distance for isolates of H1N1 with paired antigenic/genomic data for each NextClade lineage in the study space.

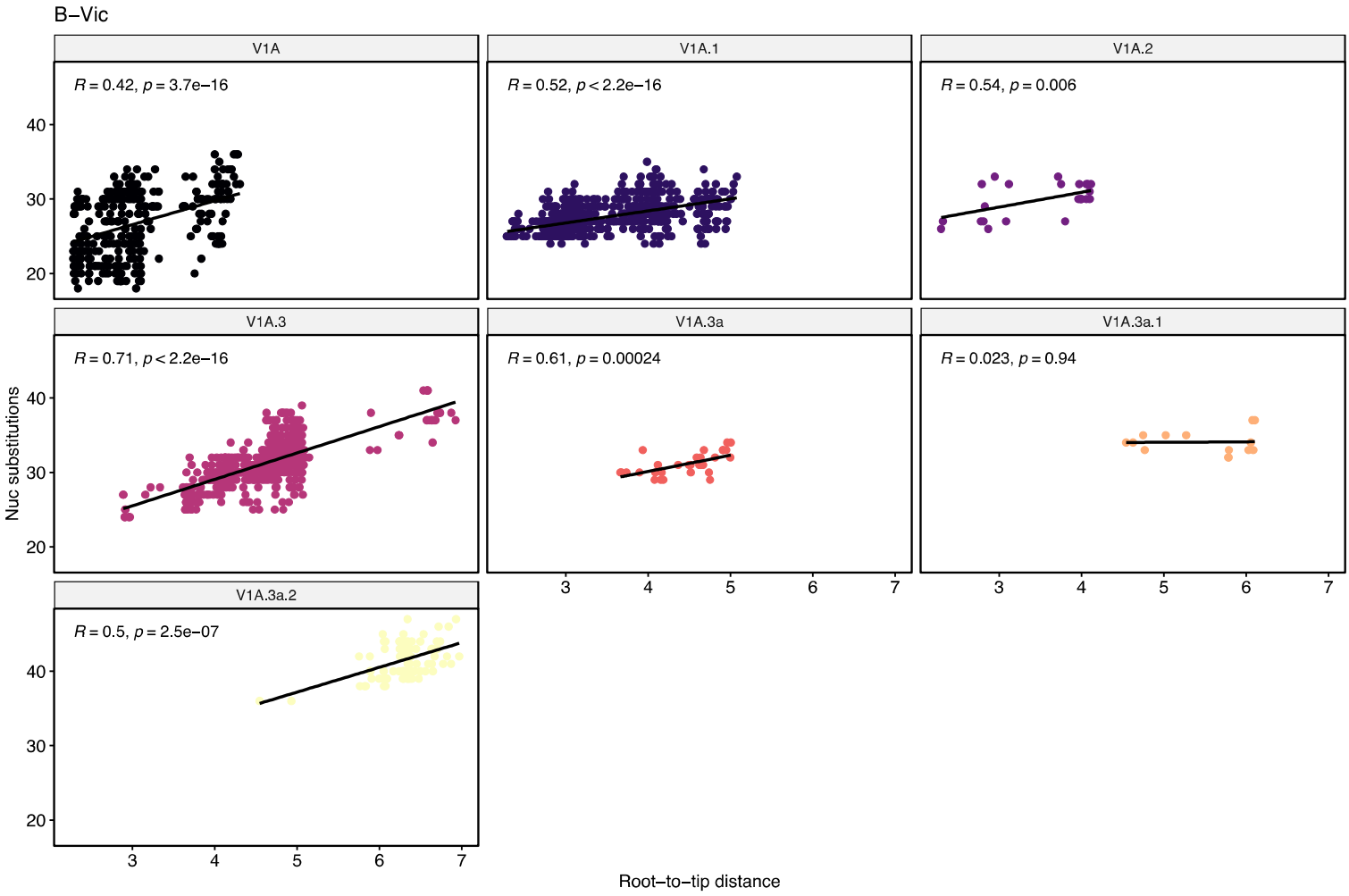


Figure S3.11. The number of nucleotide substitutions vs the root to tip distance for isolates of B-Vic with paired antigenic/genomic data for each NextClade lineage in the study space.

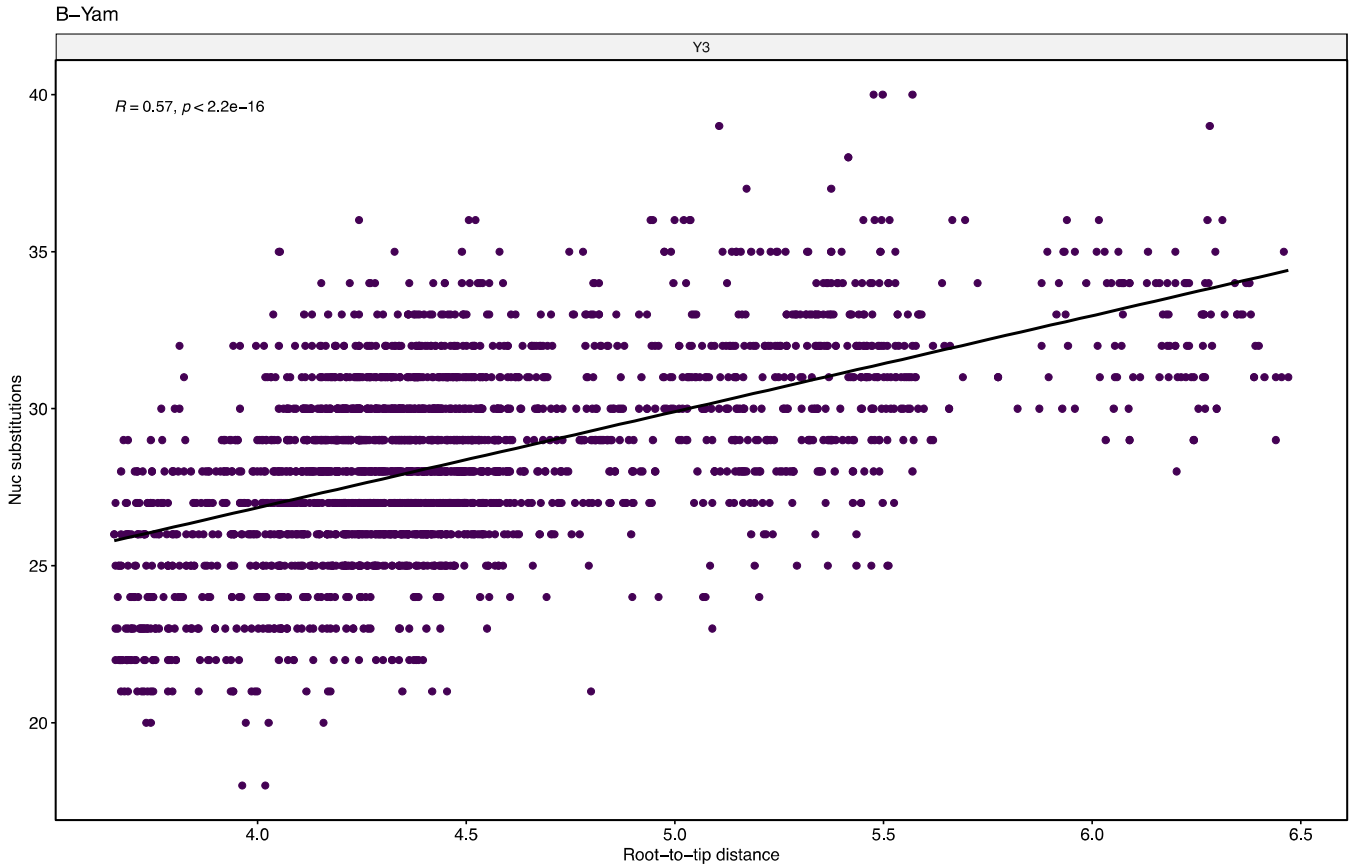


Figure S3.12. The number of nucleotide substitutions vs the root to tip distance for isolates of B-Yam with paired antigenic/genomic in the study space.

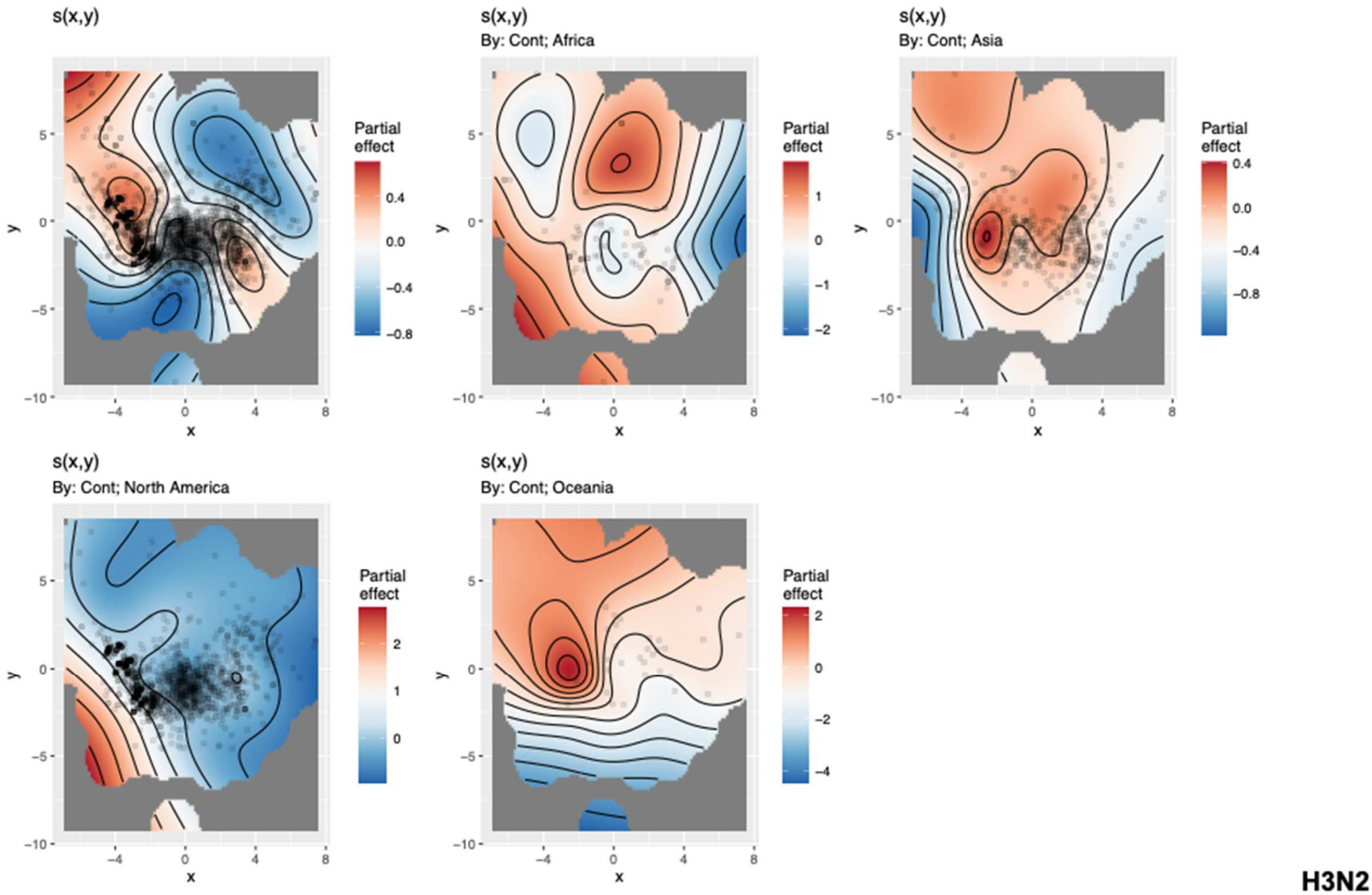


Figure S3.13. Visualization of GAM predictors for place of isolation for H3N2 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

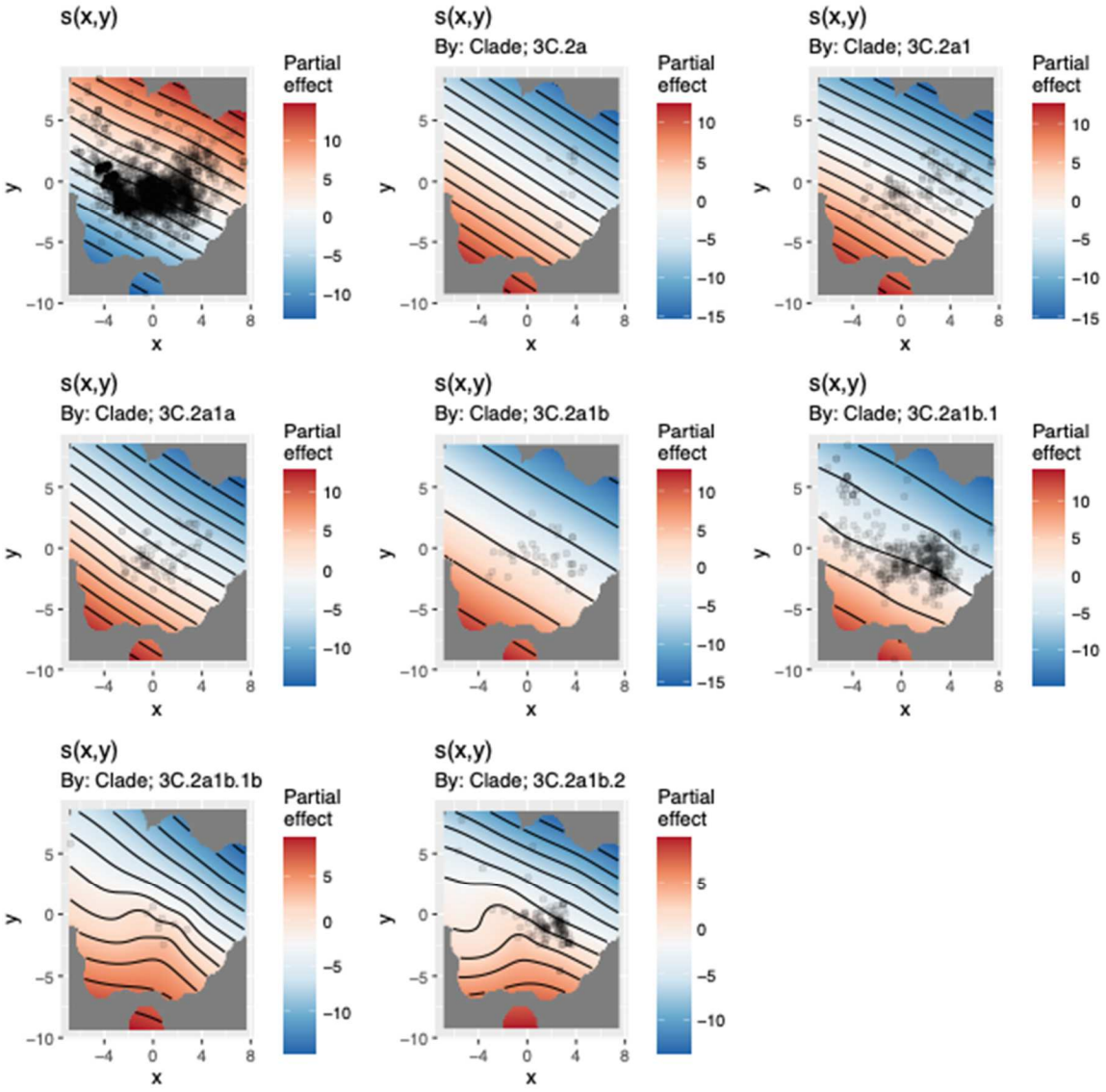
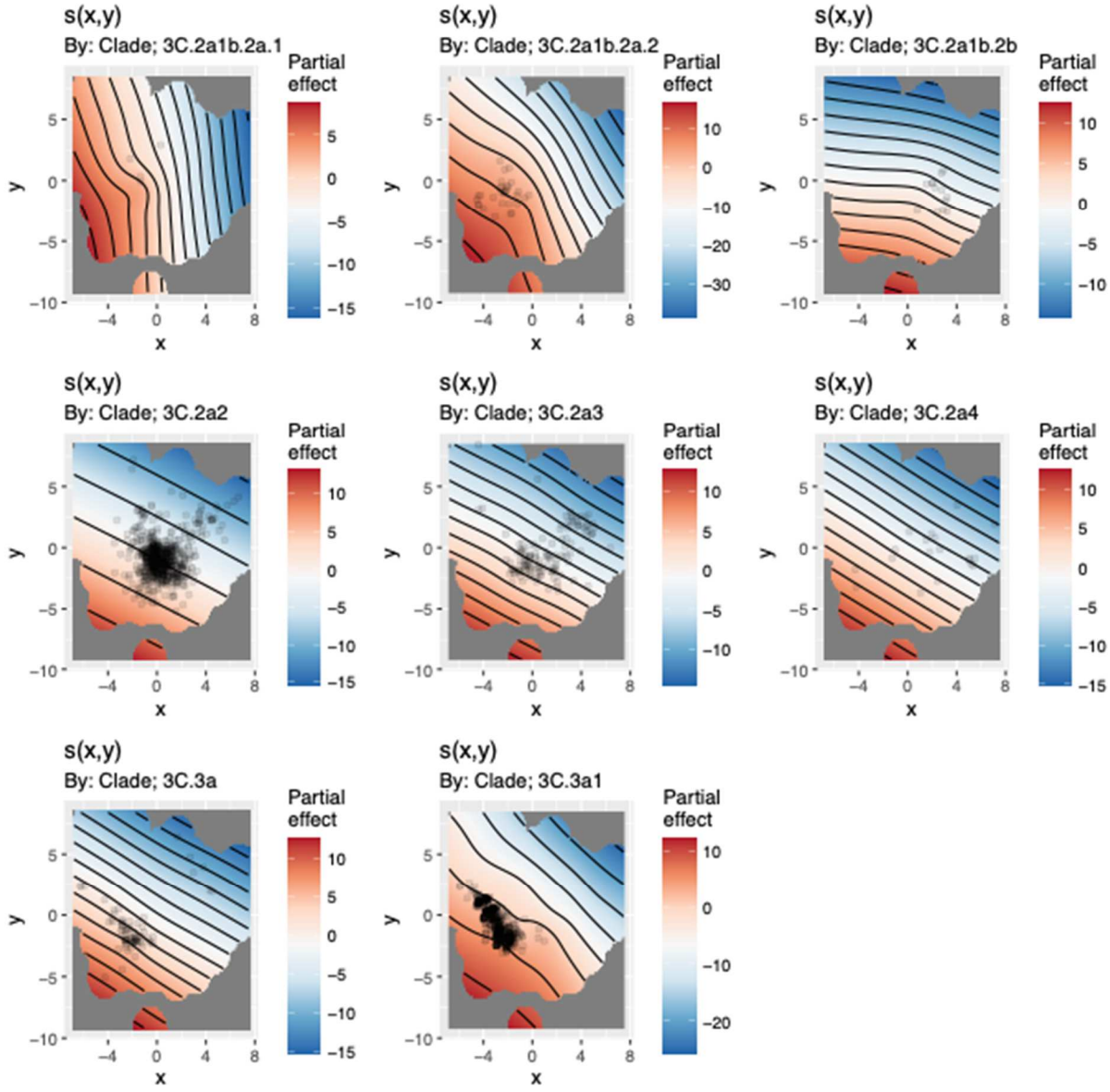


Figure S3.14. Visualization of GAM predictors for clade of H3N2 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.



H3N2

Figure S3.15. Visualization of GAM predictors for clade of H3N2 isolates (cont.). Predictors with statistically supported smooth terms and basis dimension check are visualized.

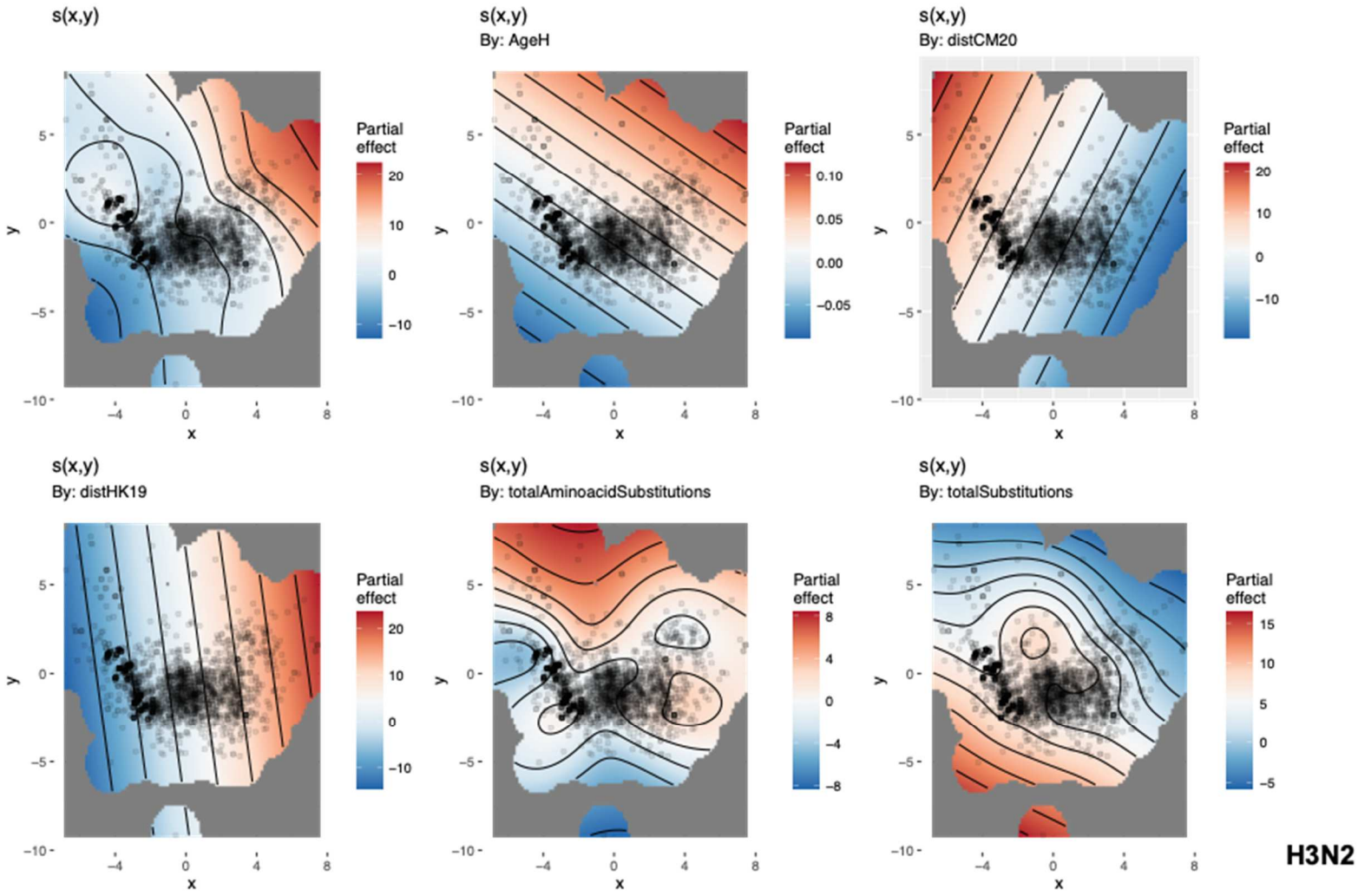


Figure S3.16. Visualization of GAM predictors for antigenic and genomic distance metrics for H3N2 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

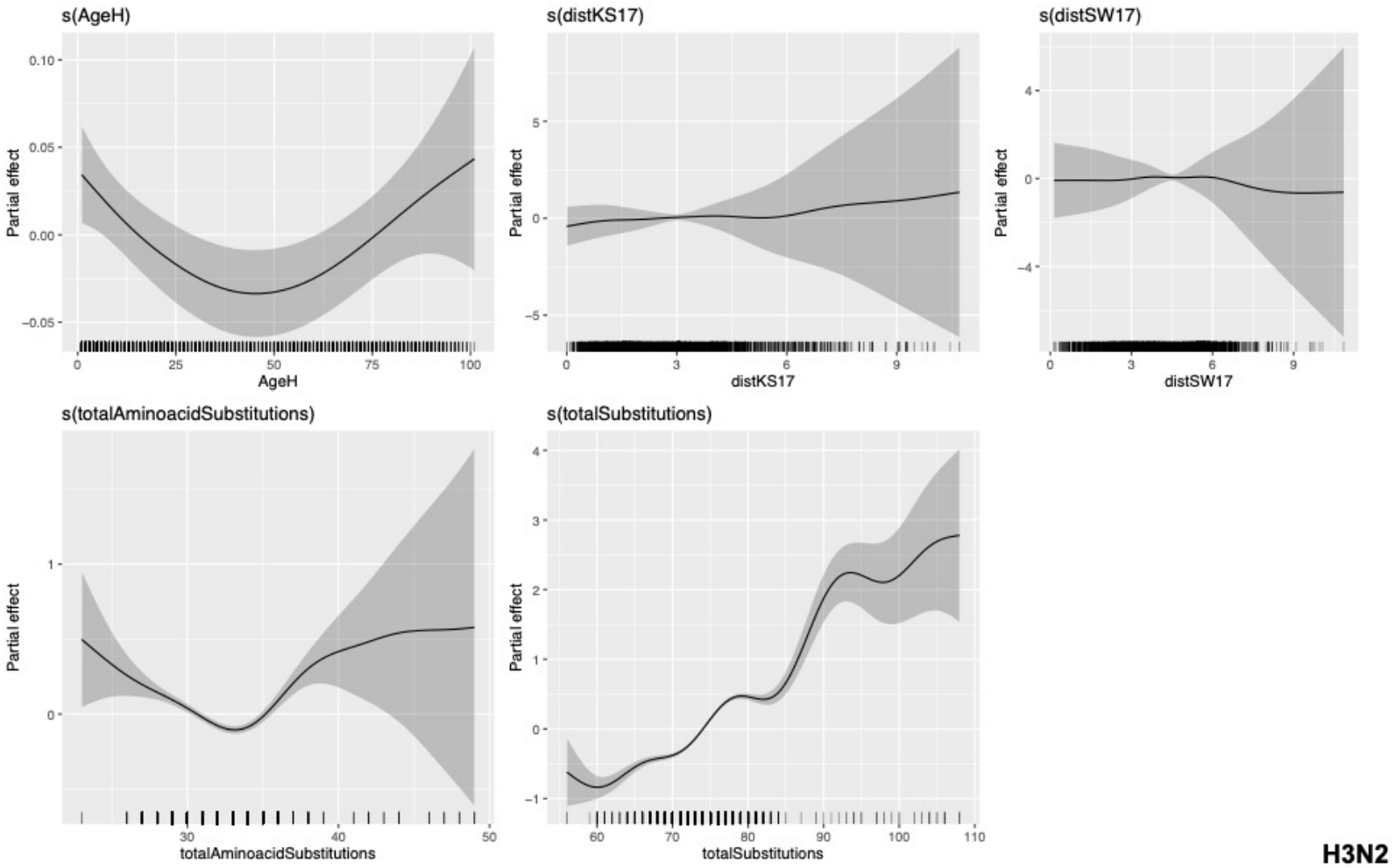


Figure S3.17. Visualization of GAM predictors of non-coordinate predictors for H3N2 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

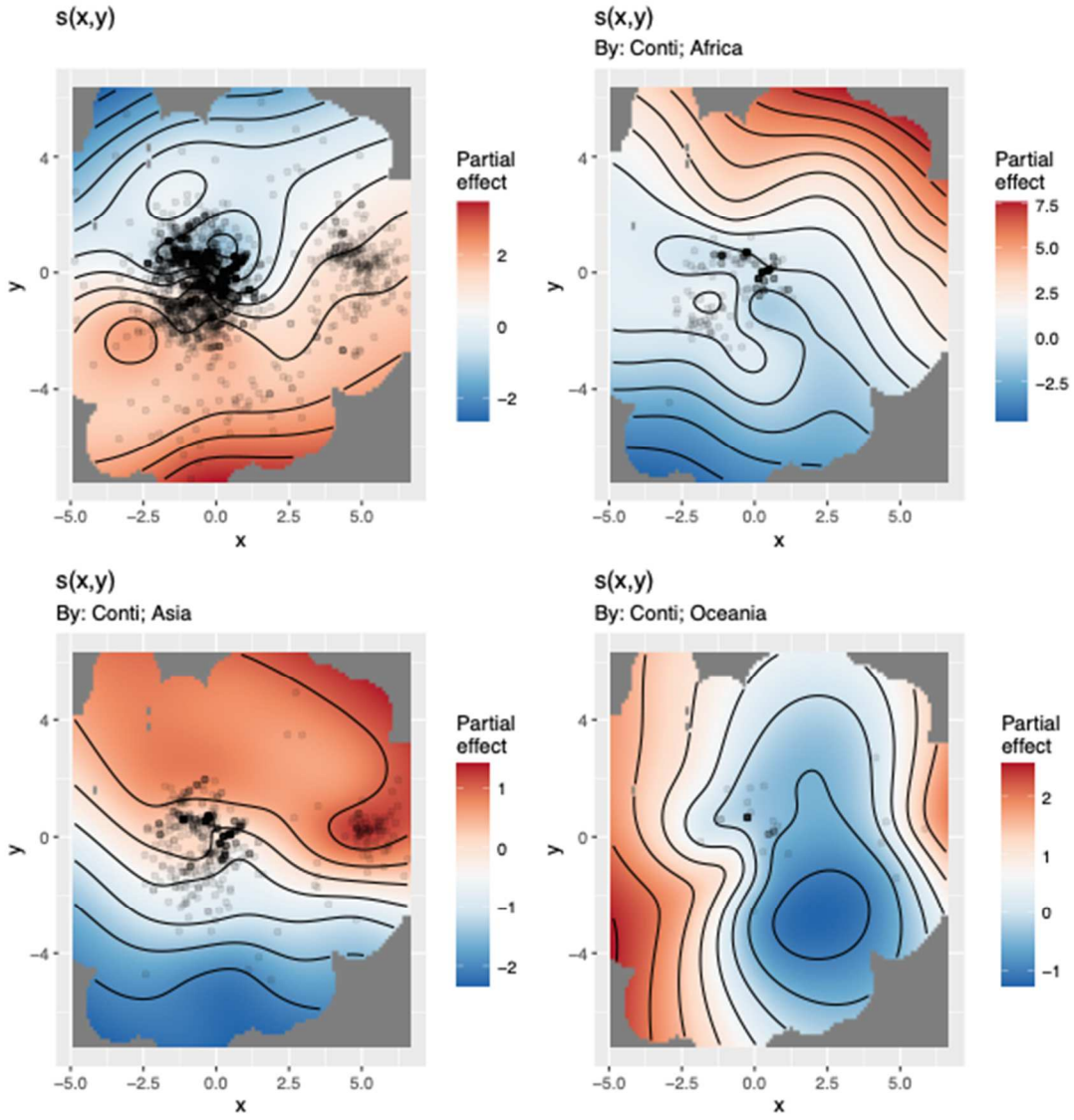


Figure S3.18. Visualization of GAM predictors for place of isolation for H1N1 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

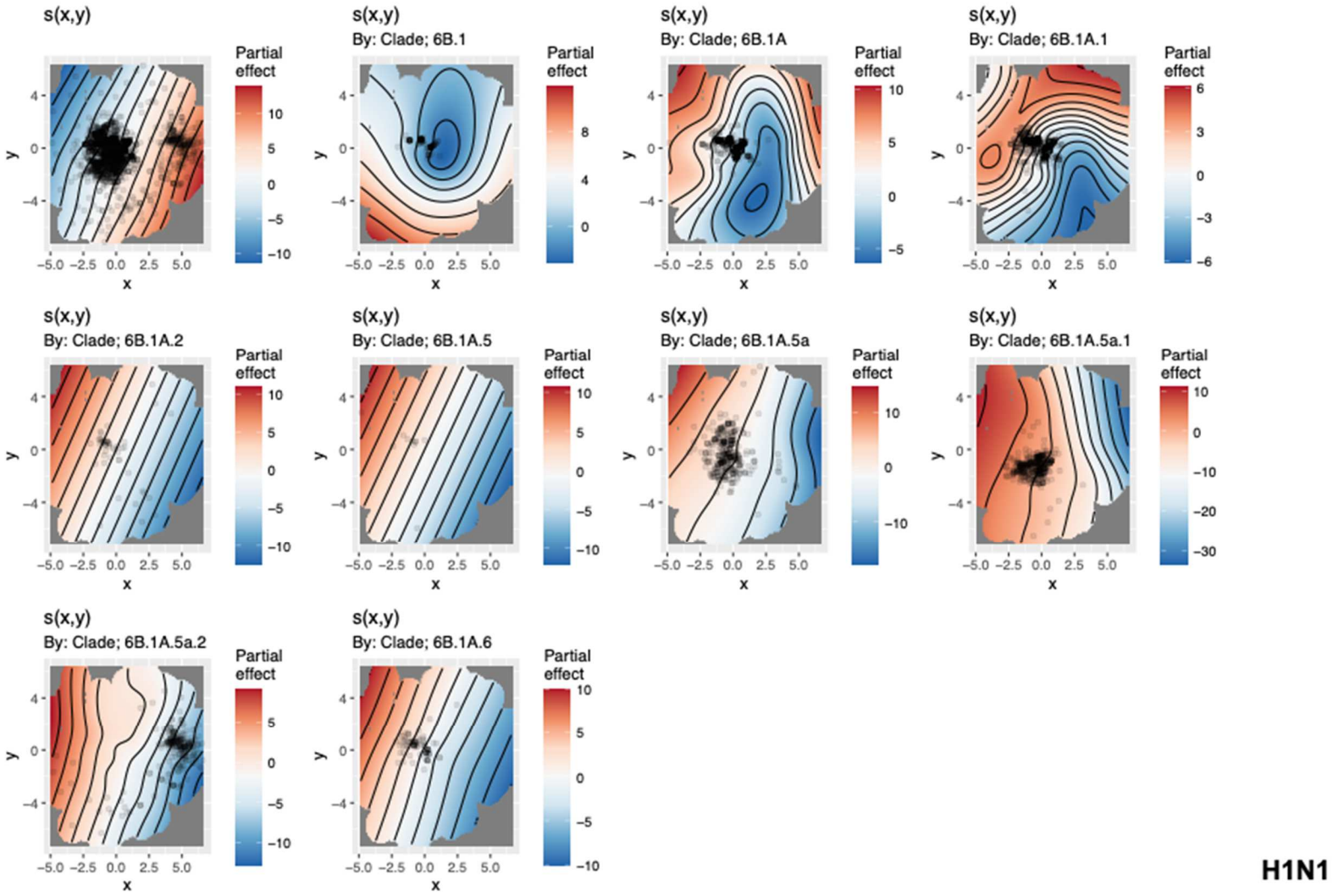


Figure S3.19. Visualization of GAM predictors for clade of H1N1 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

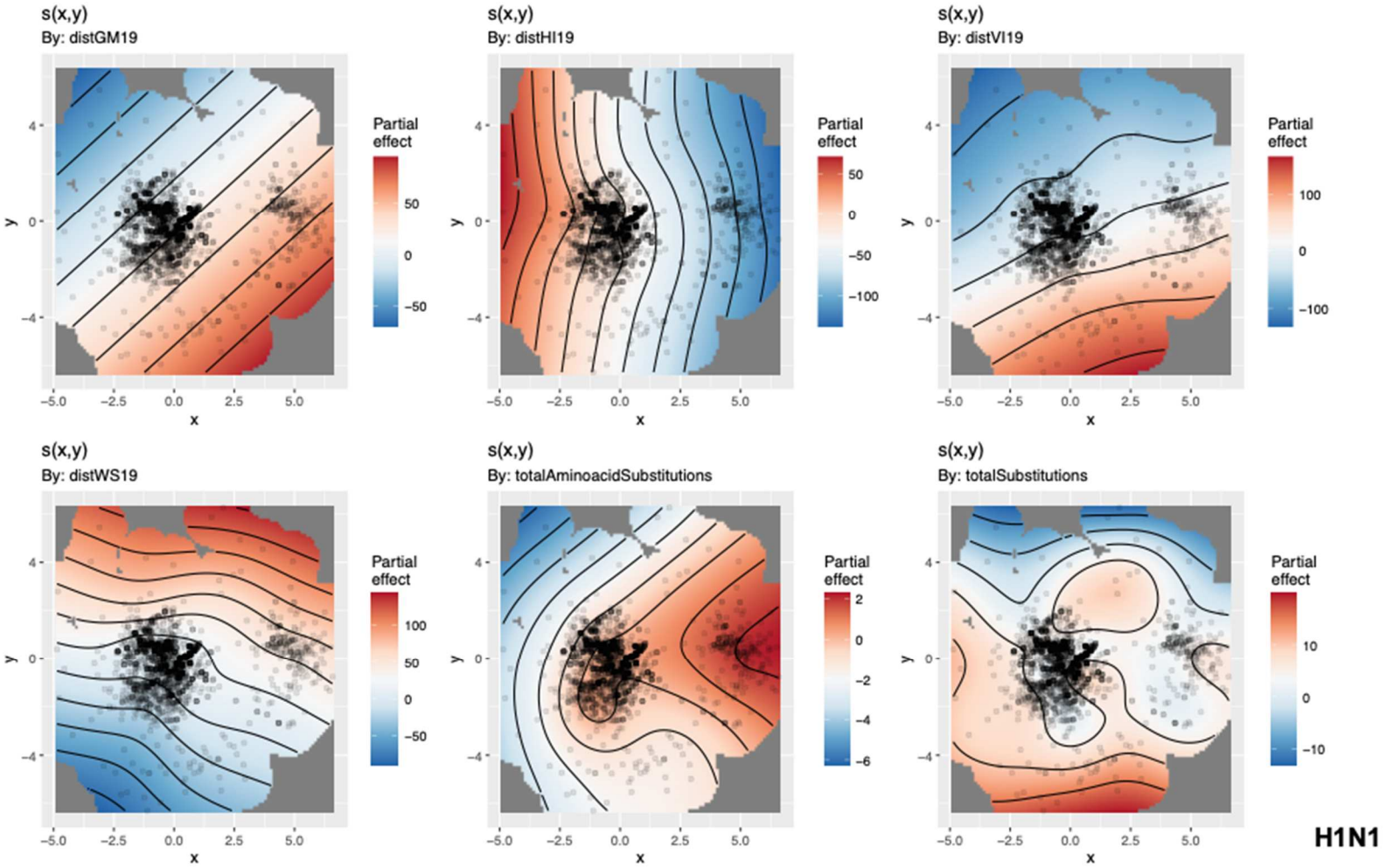


Figure S3.20. Visualization of GAM predictors for antigenic and genomic distance metrics for H1N1 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

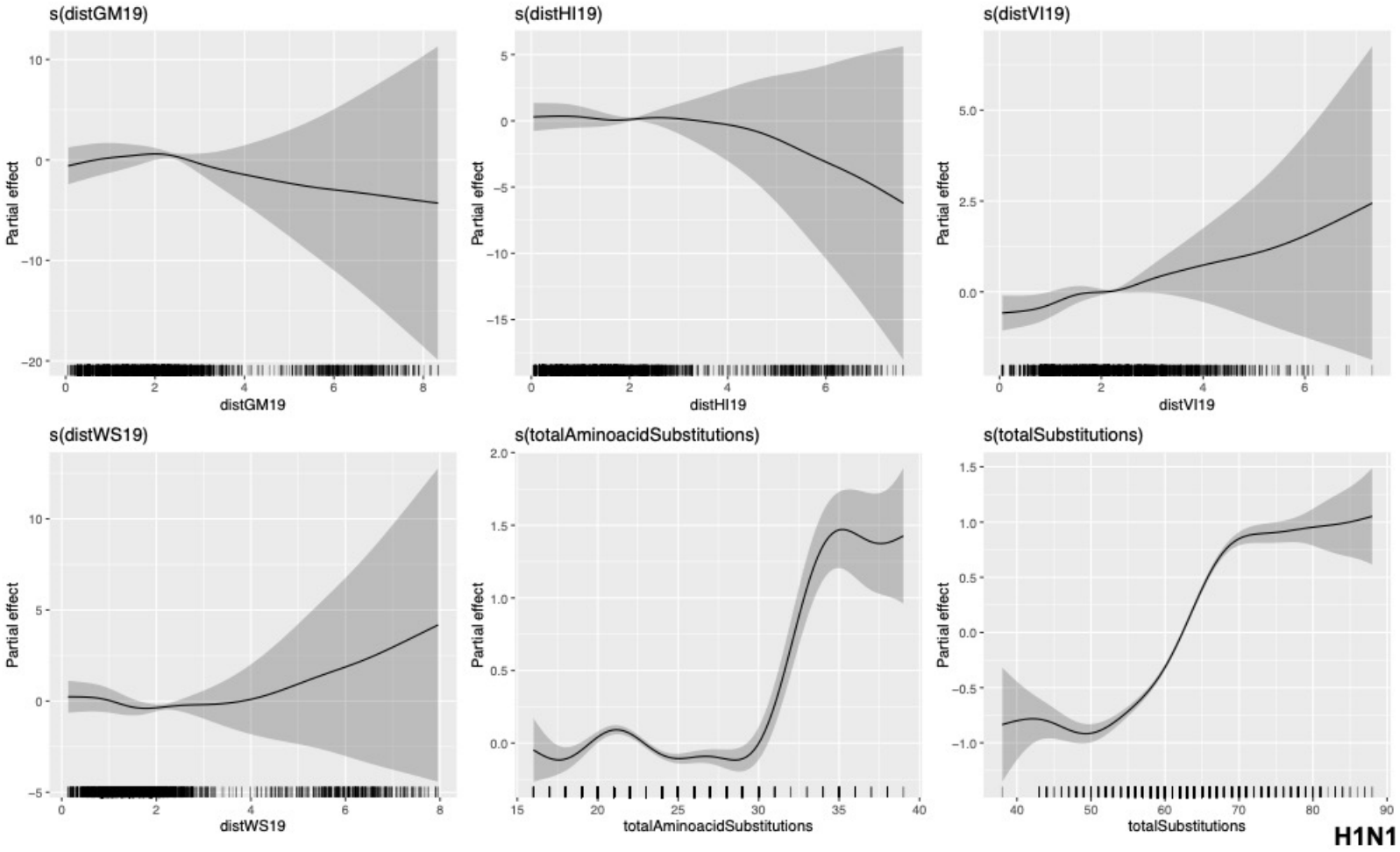


Figure S3.21. Visualization of GAM predictors of non-coordinate predictors for H1N1 isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

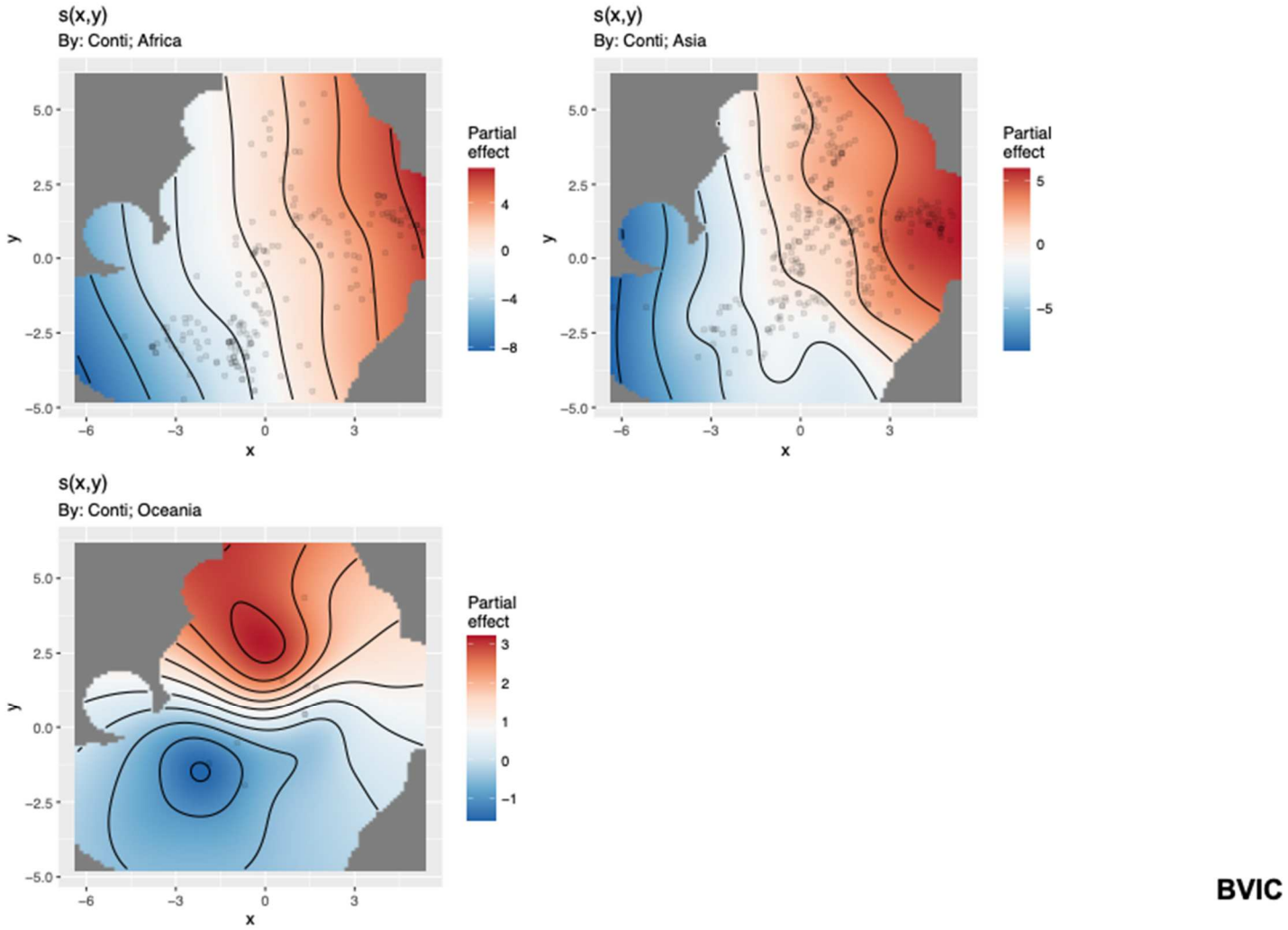


Figure S3.22. Visualization of GAM predictors for place of isolation for B-Vic isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

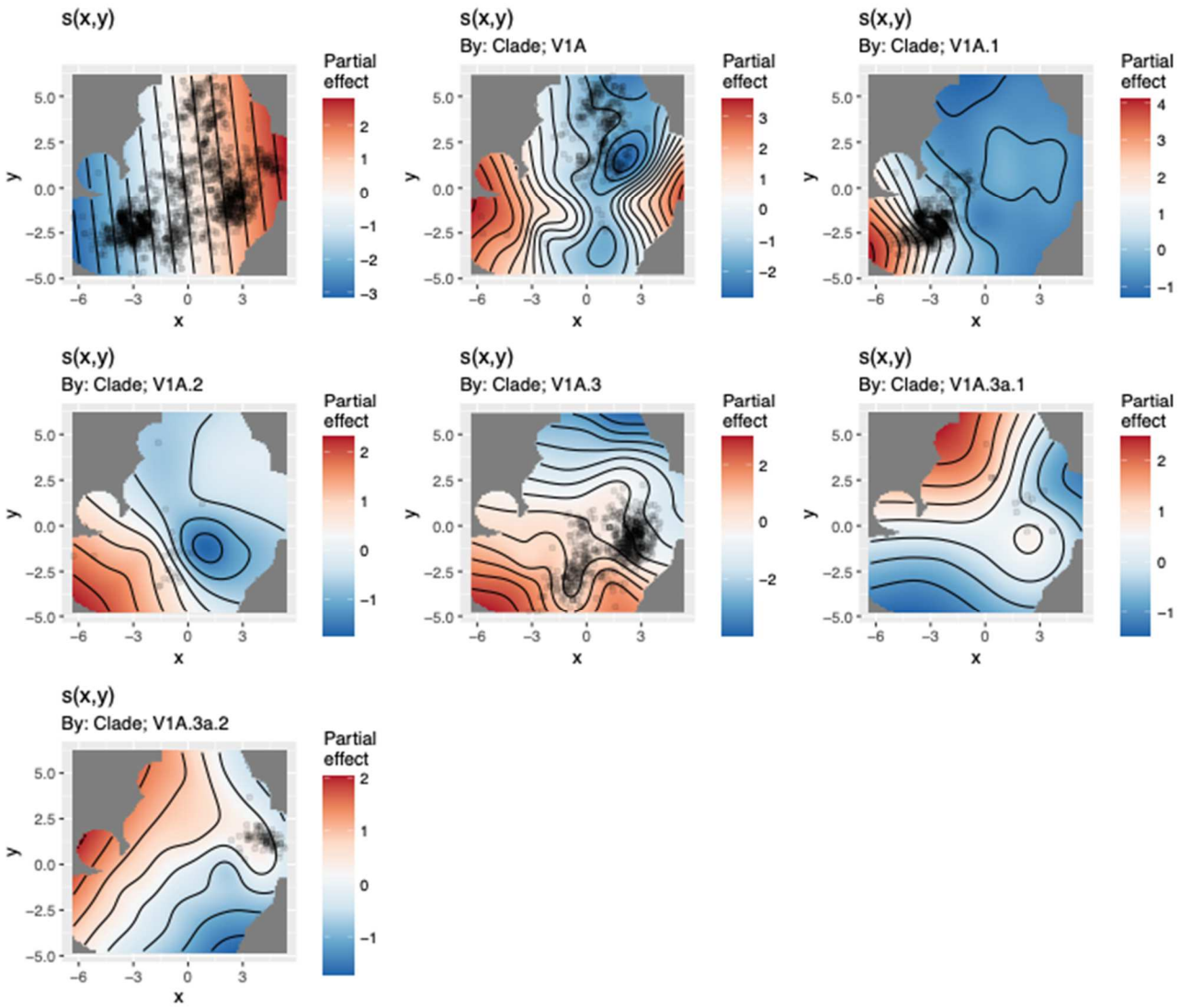


Figure S3.23. Visualization of GAM predictors for clade of B-Vic isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

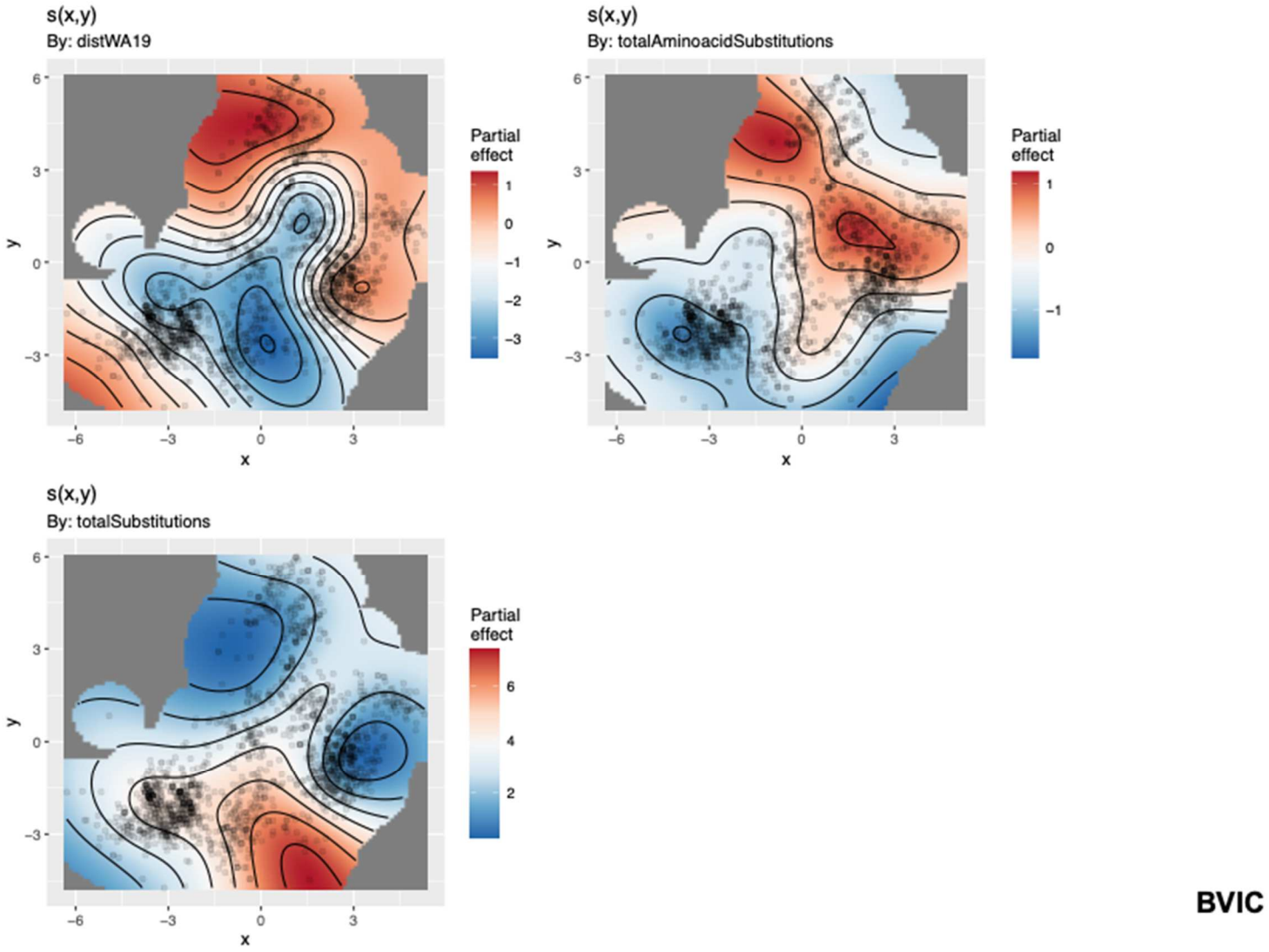


Figure S3.24. Visualization of GAM predictors for antigenic and genomic distance metrics for B-Vic isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

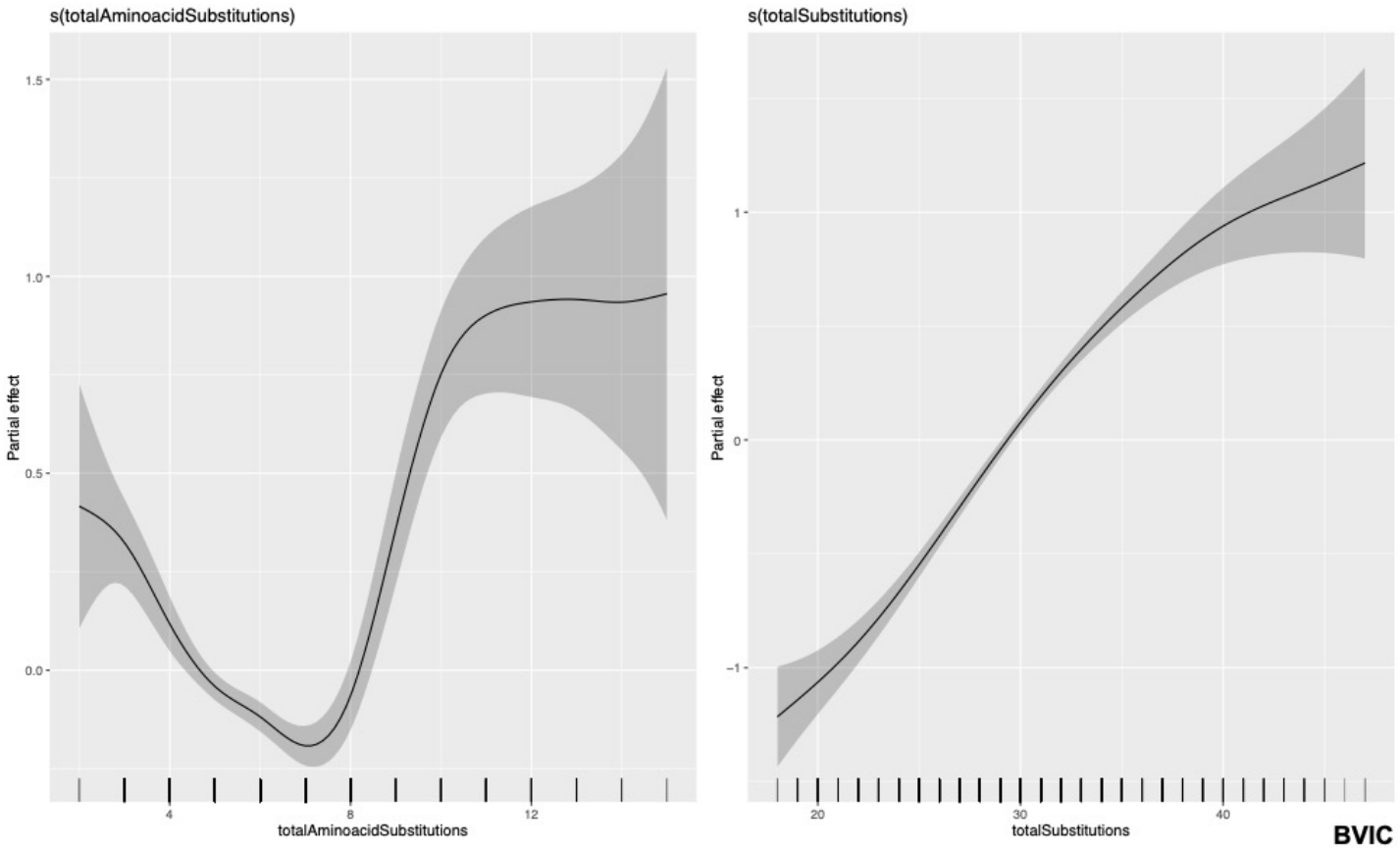


Figure S3.25. Visualization of GAM predictors of non-coordinate predictors for B-Vic isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

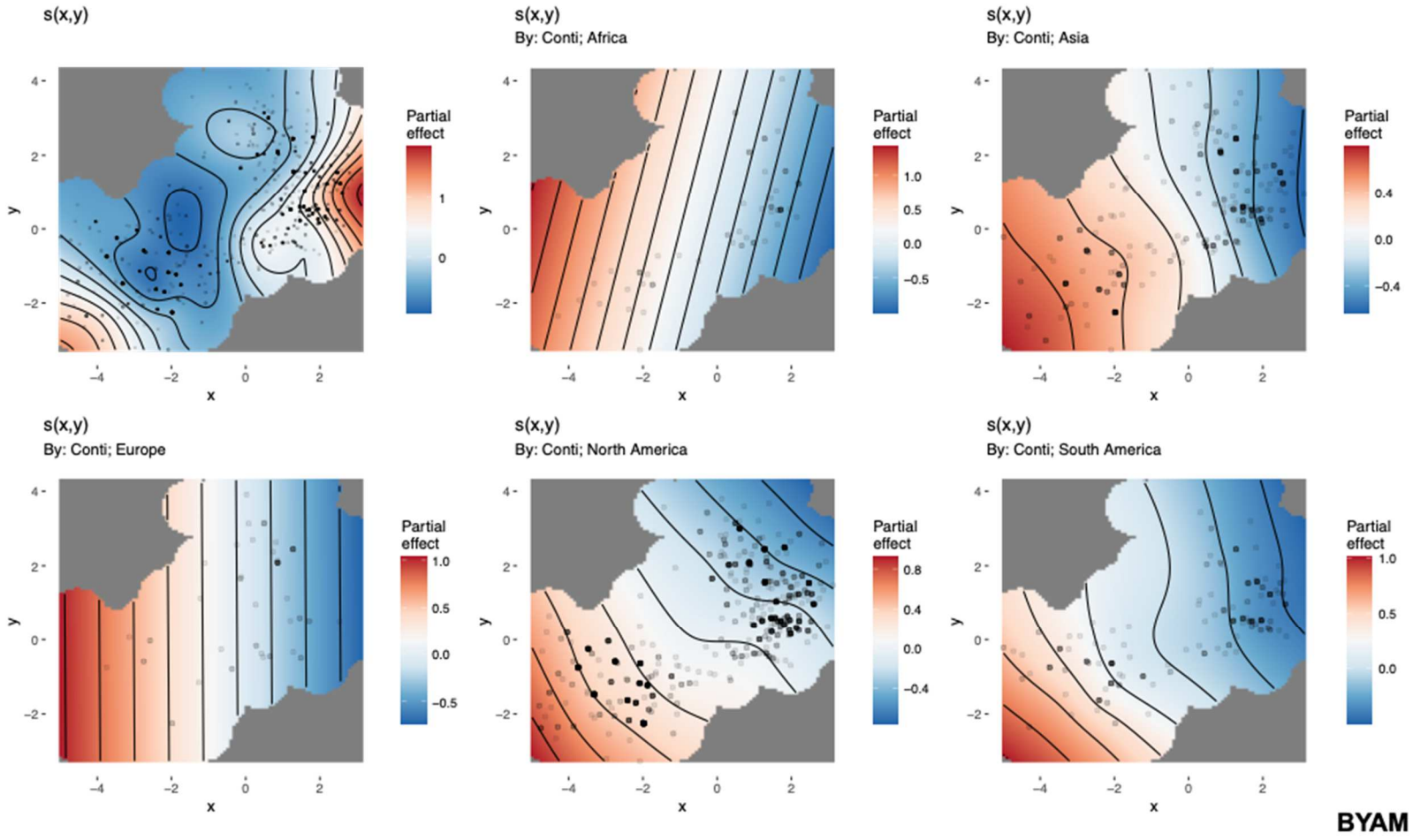


Figure S3.26. Visualization of GAM predictors for place of isolation for B-Yam isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

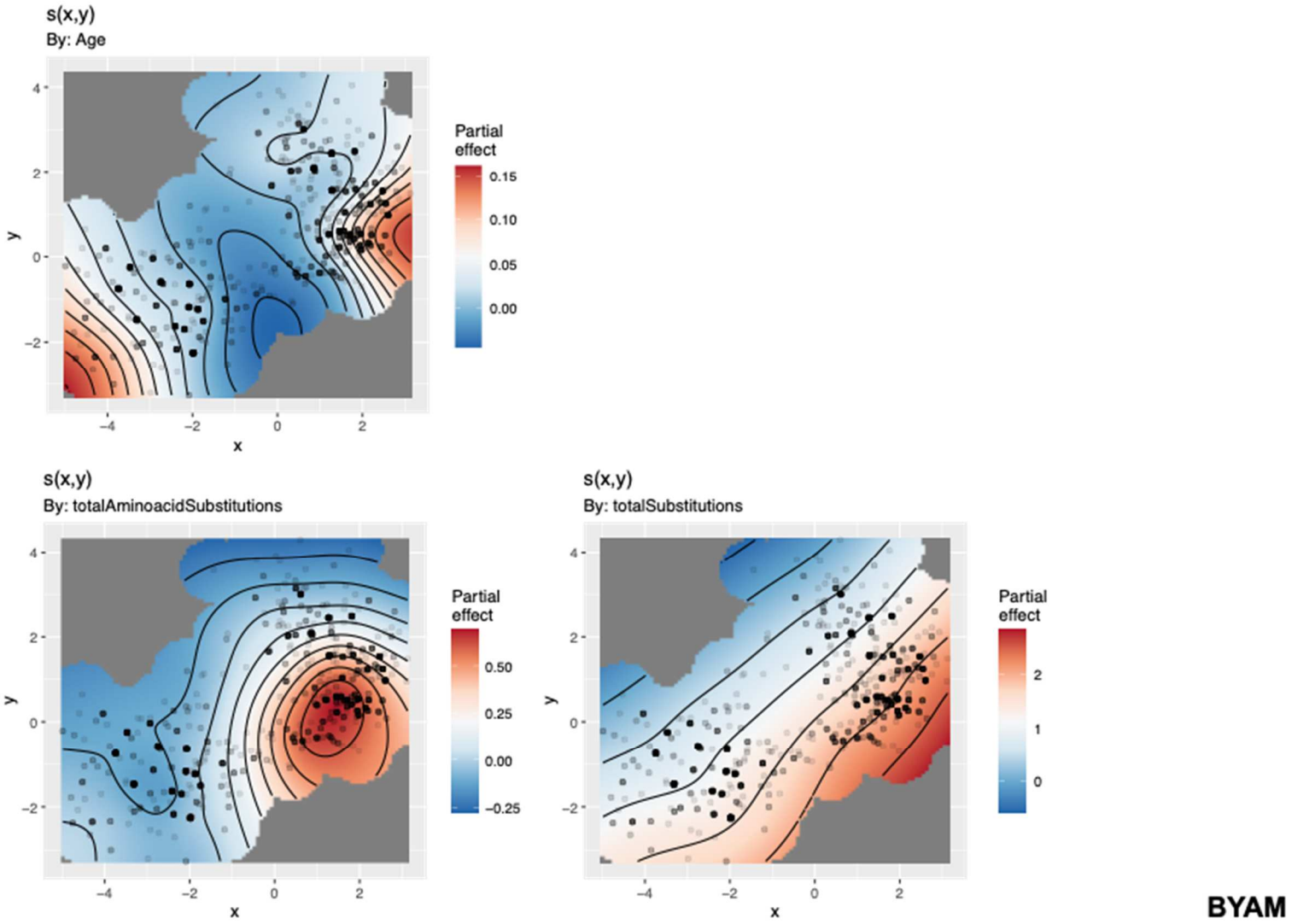


Figure S3.27. Visualization of GAM predictors for antigenic and genomic distance metrics for B-Yam isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

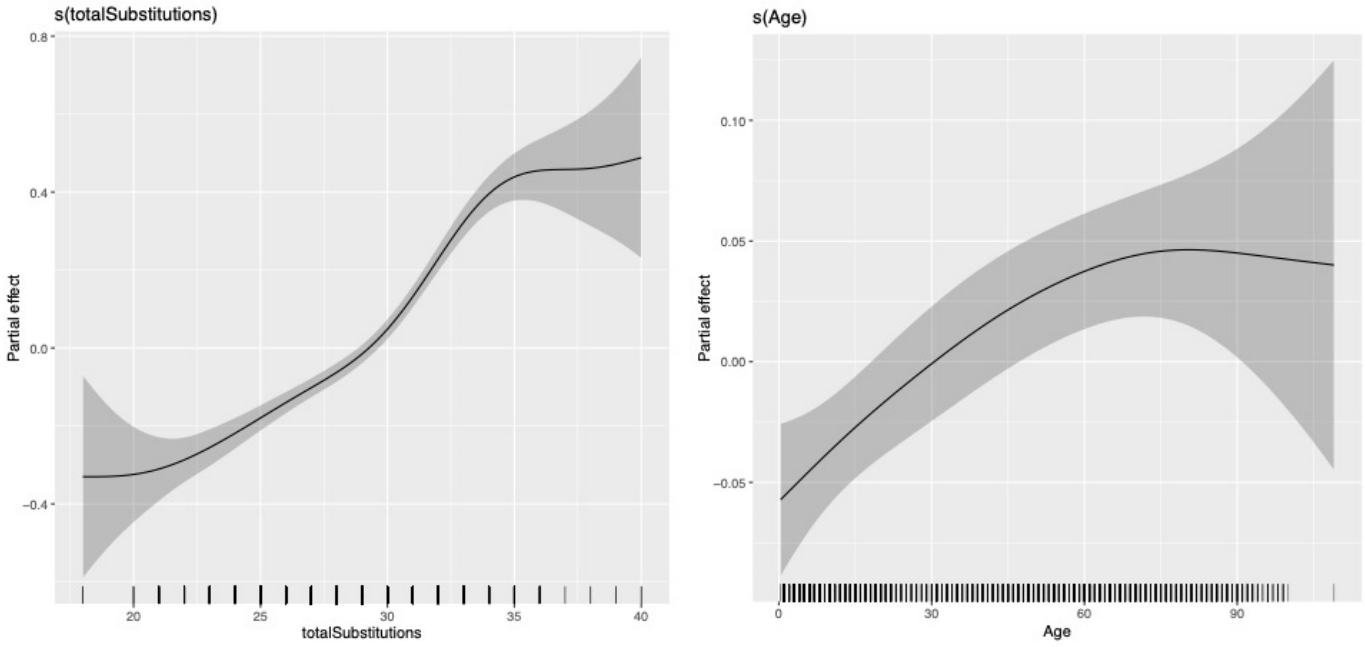


Figure S3.28. Visualization of GAM predictors of non-coordinate predictors for B-Yam isolates. Predictors with statistically supported smooth terms and basis dimension check are visualized.

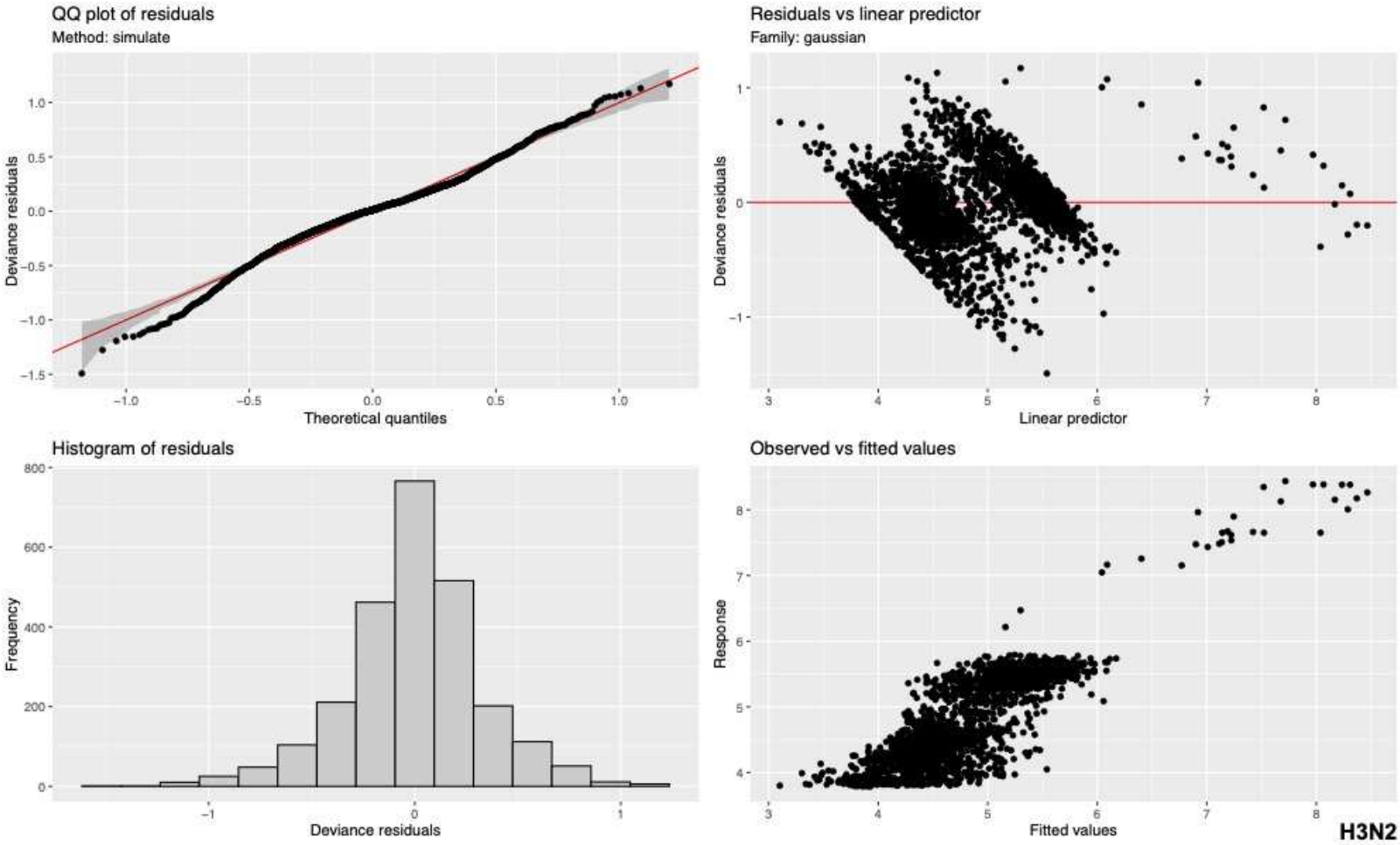


Figure S3.29. Diagnostic results for GAM of H3N2 isolates. QQ plot of residuals and residual frequencies, the residuals by the linear predictor and the observed vs fitted values are plotted.

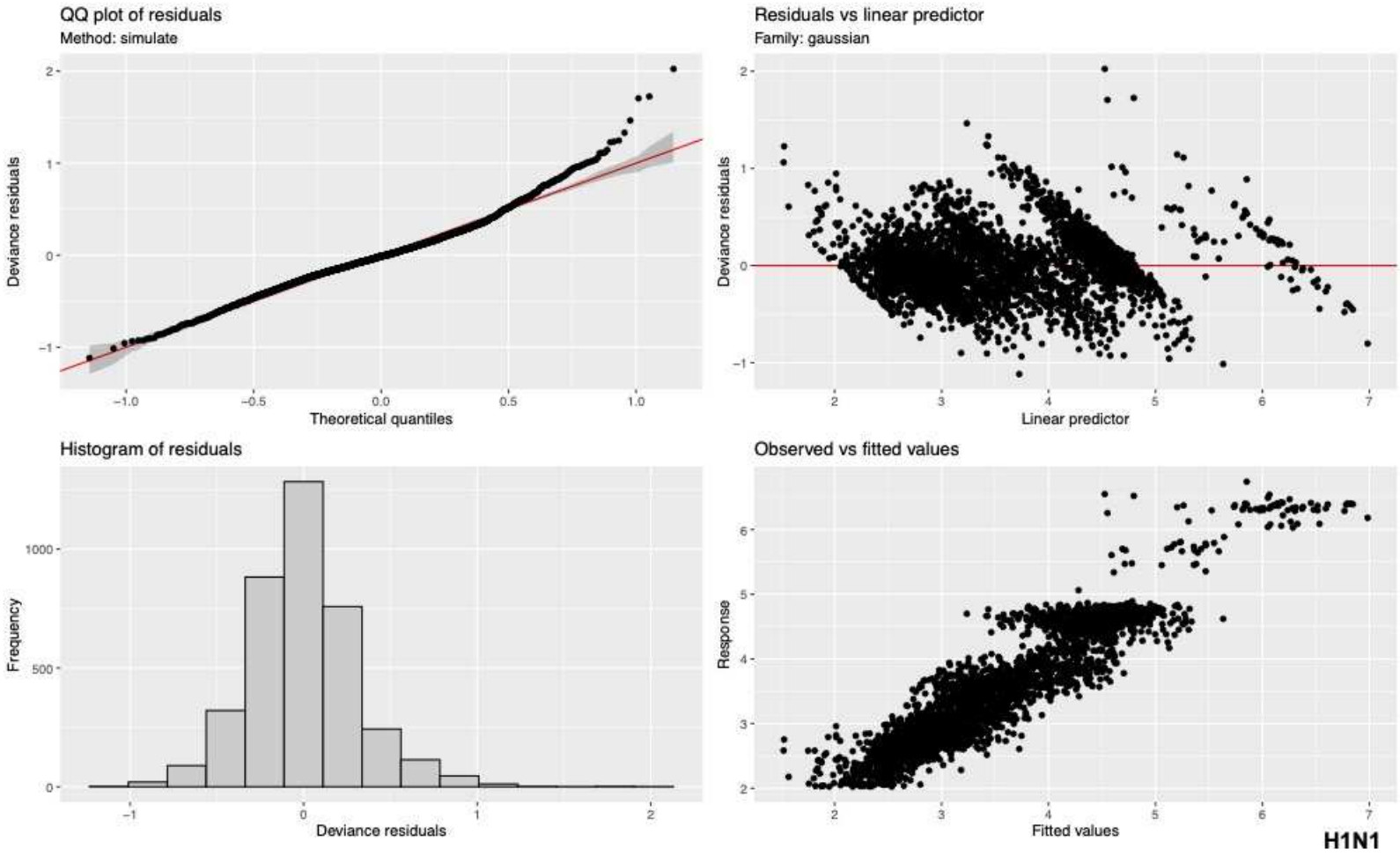


Figure S3.30. Diagnostic results for GAM of H1N1 isolates. QQ plot of residuals and residual frequencies, the residuals by the linear predictor and the observed vs fitted values are plotted.

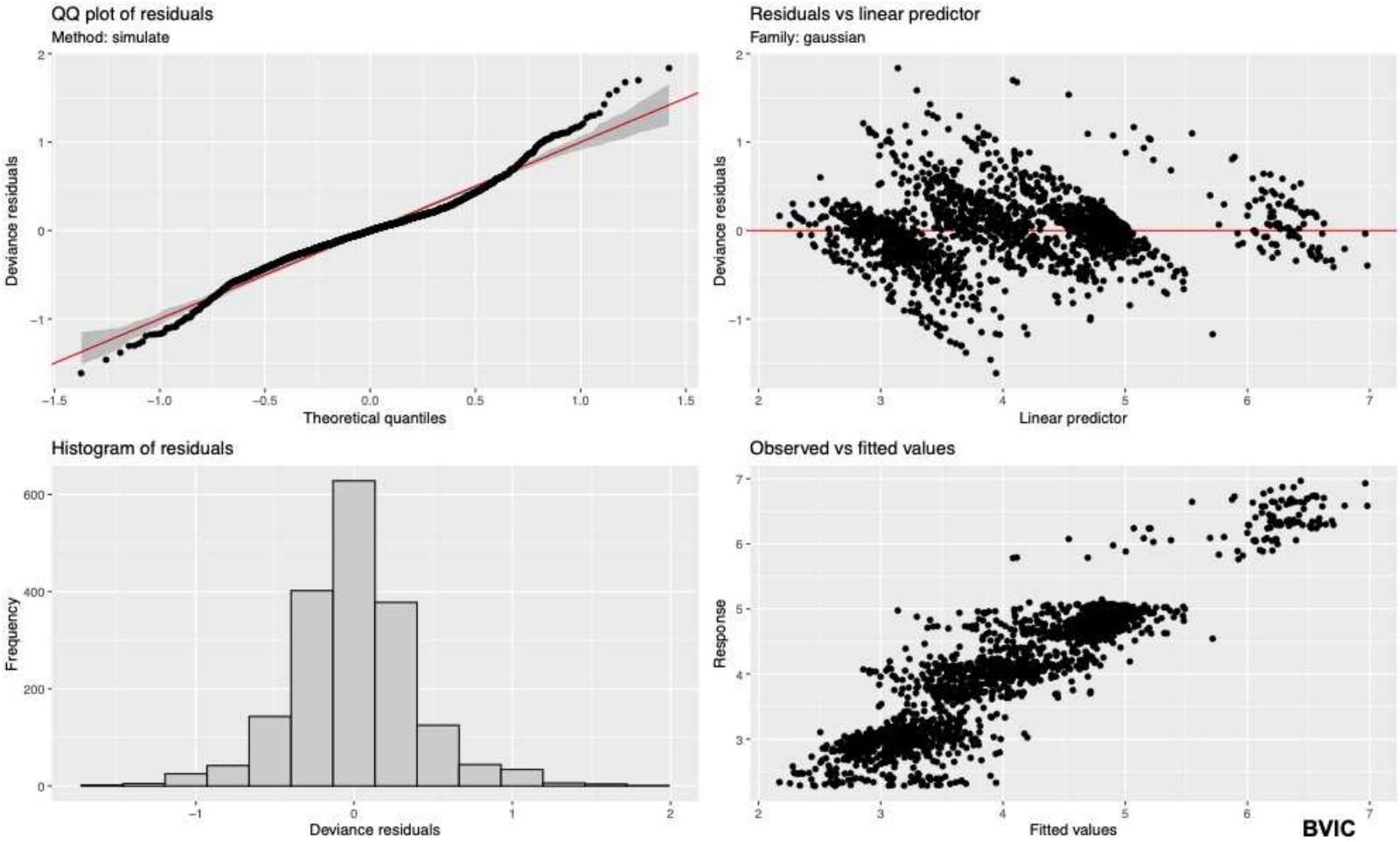


Figure S3.31. Diagnostic results for GAM of B-Vic isolates. QQ plot of residuals and residual frequencies, the residuals by the linear predictor and the observed vs fitted values are plotted.

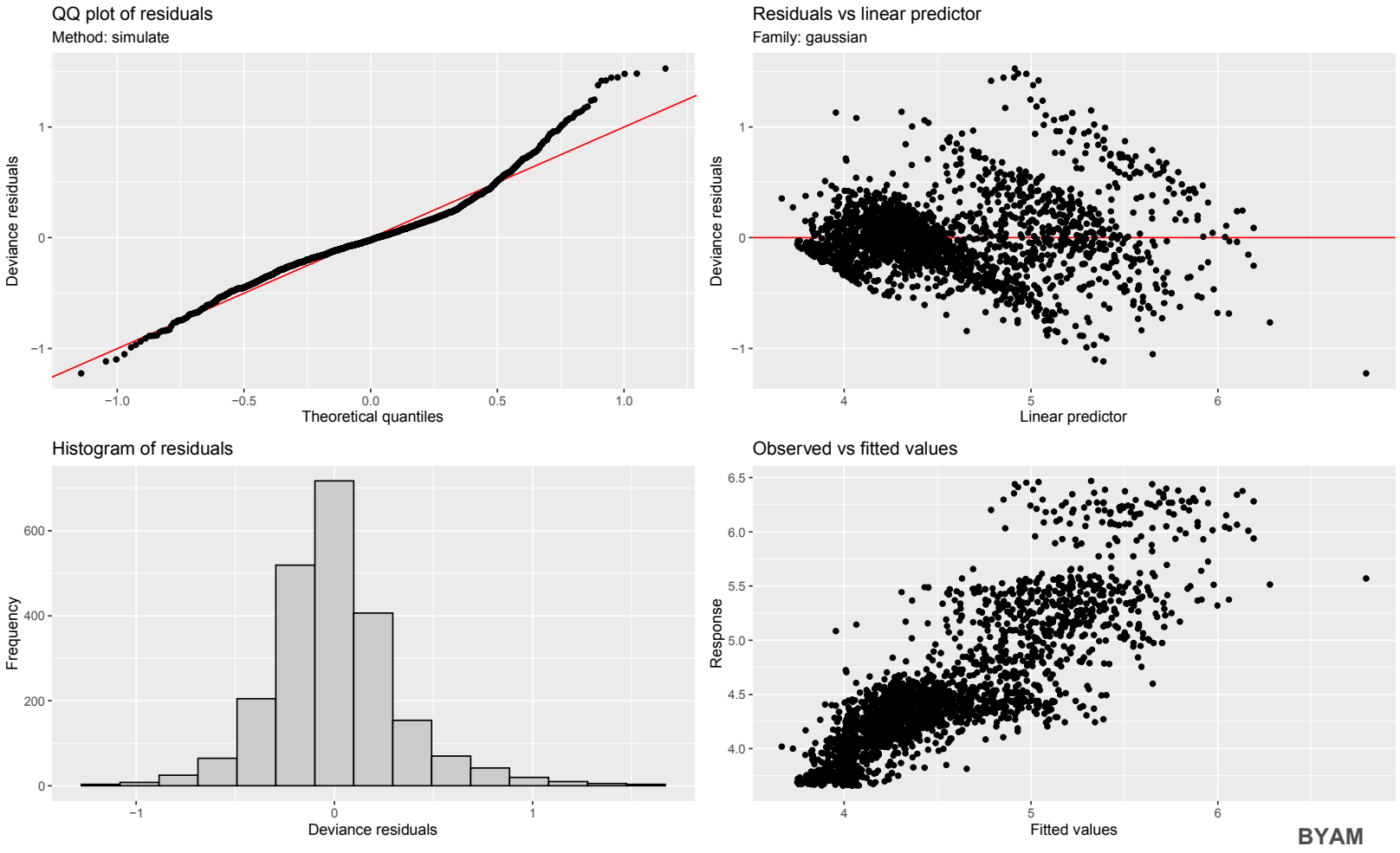


Figure S3.32. Diagnostic results for GAM of B-Yam isolates. QQ plot of residuals and residual frequencies, the residuals by the linear predictor and the observed vs fitted values are plotted.

Table S3.1. Significance of smoothing terms for H3N2 isolates. Predictors for the GAM implemented for H3N2 isolate root to tip distance as the response variable with corresponding statistical support.

H3N2 GAM results

Significance of smooth terms

Predictor	edf	Ref.df	F	p-value	Sig Code
s(x,y)	1.71E+01	2.03E+01	15.9	<2e-16	***
s(x,y):Clad3C.2a	2.00E+00	2.00E+00	67.22	<2e-16	***
s(x,y):Clad3C.2a1	6.93E+00	9.21E+00	25.32	<2e-16	***
s(x,y):Clad3C.2a1a	5.33E+00	7.13E+00	26.42	<2e-16	***
s(x,y):Clad3C.2a1b	4.22E+00	5.64E+00	34.39	<2e-16	***
s(x,y):Clad3C.2a1b.1	2.14E+01	2.51E+01	14.96	<2e-16	***
s(x,y):Clad3C.2a1b.1a	7.54E-05	1.51E-04	0	0.5	
s(x,y):Clad3C.2a1b.1b	6.44E+00	7.32E+00	13.27	<2e-16	***
s(x,y):Clad3C.2a1b.2	1.64E+01	1.92E+01	16.47	<2e-16	***
s(x,y):Clad3C.2a1b.2a.1	3.40E+00	3.67E+00	13.48	<2e-16	***
s(x,y):Clad3C.2a1b.2a.2	9.79E+00	1.14E+01	60.34	<2e-16	***
s(x,y):Clad3C.2a1b.2b	4.66E+00	5.58E+00	27.55	<2e-16	***
s(x,y):Clad3C.2a2	8.48E+00	1.09E+01	24.46	<2e-16	***
s(x,y):Clad3C.2a3	9.95E+00	1.30E+01	18.01	<2e-16	***
s(x,y):Clad3C.2a4	3.73E+00	4.75E+00	35.89	<2e-16	***
s(x,y):Clad3C.3a	8.29E+00	1.03E+01	22.29	<2e-16	***
s(x,y):Clad3C.3a1	1.55E+01	1.73E+01	23.87	<2e-16	***
s(x,y):OselS	9.022	10.693	0.301	0.986	
s(x,y):OselU	3.32	3.895	0.14	0.971	
s(x,y):ZanaS	11.022	12.693	1.047	0.377	
s(x,y):ZanaU	3.32	3.895	0.14	0.971	
s(x,y):ContAfrica	15.041	18.508	2.442	0.000508	***
s(x,y):ContAsia	16.019	20.158	1.681	0.025324	*
s(x,y):ContEurope	1.436	2.381	0.643	0.693068	
s(x,y):ContNorth_America	18.065	21.673	2.565	9.62E-05	***
s(x,y):ContOceania	11.711	14.423	4.102	9.92E-07	***
s(x,y):ContSouth_America	2.016	2.025	1.871	0.153897	
s(x,y):AgeH	3	3.001	2.625	0.04887	*
s(x,y):totalSubstitutions	17.322	20.824	52.187	<2e-16	***
s(x,y):totalAminoacidSubstitutions	21.551	24.795	4.42	<2e-16	***
s(x,y):distSW17	3.001	3.001	0.499	0.68338	
s(x,y):distKS17	3.001	3.001	2.312	0.07421	.
s(x,y):distSA19	5.264	6.642	0.323	0.93628	
s(x,y):distHK19	3.01	3.017	2.934	0.03198	*
s(x,y):distCM20	3	3.001	4.302	0.00491	**
s(AgeH)	2.704	3.358	3.966	0.00613	**
s(totalSubstitutions)	8.469	8.869	119.994	<2e-16	***
s(totalAminoacidSubstitutions)	5.439	6.528	18.512	<2e-16	***
s(distSW17)	6.189	7.152	1.911	0.0495	*
s(distKS17)	6.585	7.397	2.691	0.00671	**
s(distSA19)	2.996	3.802	0.32	0.83115	
s(distHK19)	1.074	1.114	0.152	0.76645	
s(distCM20)	5.782	6.79	1.834	0.0561	.

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.2. Basis dimension checking for H3N2 isolates. Basis dimension indices for predictors for the GAM implemented for H3N2 isolate root to tip distance as the response variable with corresponding statistical support.

H3N2 GAM results

Basis dimension (k) checking

Predictor	k'	edf	k-index	p-value	Sig Code
s(x,y)	2.90E+01	1.71E+01	0.84	<2e-16	***
s(x,y):Clad3C.2a	2.90E+01	2.00E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1	2.90E+01	6.93E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1a	2.90E+01	5.33E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1b	2.90E+01	4.22E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.1	2.90E+01	2.14E+01	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.1a	2.90E+01	7.54E-05	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.1b	2.90E+01	6.44E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.2	2.90E+01	1.64E+01	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.2a.1	2.90E+01	3.40E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.2a.2	2.90E+01	9.79E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a1b.2b	2.90E+01	4.66E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a2	2.90E+01	8.48E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a3	2.90E+01	9.95E+00	0.84	<2e-16	***
s(x,y):Clad3C.2a4	2.90E+01	3.73E+00	0.84	<2e-16	***
s(x,y):Clad3C.3a	2.90E+01	8.29E+00	0.84	<2e-16	***
s(x,y):Clad3C.3a1	2.90E+01	1.55E+01	0.84	<2e-16	***
s(x,y):OselS	29	9.02	0.85	<2e-16	***
s(x,y):OselU	29	3.32	0.85	<2e-16	***
s(x,y):ZanaS	29	11.02	0.85	<2e-16	***
s(x,y):ZanaU	29	3.32	0.85	<2e-16	***
s(x,y):ContAfrica	29	15.04	0.85	<2e-16	***
s(x,y):ContAsia	29	16.02	0.85	<2e-16	***
s(x,y):ContEurope	29	1.44	0.85	<2e-16	***
s(x,y):ContNorth_America	29	18.06	0.85	<2e-16	***
s(x,y):ContOceania	29	11.71	0.85	<2e-16	***
s(x,y):ContSouth_America	29	2.02	0.85	<2e-16	***
s(x,y):AgeH	30	3	0.9	<2e-16	***
s(x,y):totalSubstitutions	30	17.32	0.9	<2e-16	***
s(x,y):totalAminoacidSubstitutions	30	21.55	0.9	<2e-16	***
s(x,y):distSW17	30	3	0.9	<2e-16	***
s(x,y):distKS17	30	3	0.9	<2e-16	***
s(x,y):distSA19	30	5.26	0.9	<2e-16	***
s(x,y):distHK19	30	3.01	0.9	<2e-16	***
s(x,y):distCM20	30	3	0.9	<2e-16	***
s(AgeH)	9	2.7	0.74	<2e-16	***
s(totalSubstitutions)	9	8.47	0.38	<2e-16	***
s(totalAminoacidSubstitutions)	9	5.44	0.57	<2e-16	***
s(distSW17)	9	6.19	0.91	<2e-16	***
s(distKS17)	9	6.59	0.97	0.05	*
s(distSA19)	9	3	0.92	<2e-16	***
s(distHK19)	9	1.07	0.9	<2e-16	***
s(distCM20)	9	5.78	0.91	<2e-16	***

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.3. Significance of smoothing terms for H1N1 isolates. Predictors for the GAM implemented for H1N1 isolate root to tip distance as the response variable with corresponding statistical support.

H1N1 GAM results

Significance of smooth terms

Predictor	edf	Ref.df	F	p-value	Sig code
s(x,y)	28.011	28.87	115.339	<2e-16	***
s(x,y):Clade6B	3.031	3.713	3.22	0.0693	.
s(x,y):Clade6B.1	7.568	8.678	22.958	<2e-16	***
s(x,y):Clade6B.1A	19.292	20.923	19.98	<2e-16	***
s(x,y):Clade6B.1A.1	18.853	20.915	9.753	<2e-16	***
s(x,y):Clade6B.1A.2	2.002	2.004	1.364	0.25597	
s(x,y):Clade6B.1A.3	13.522	15.605	7.477	<2e-16	***
s(x,y):Clade6B.1A.4	2	2	1.055	0.34835	
s(x,y):Clade6B.1A.5	2	2	1.266	0.28199	
s(x,y):Clade6B.1A.5a	21.282	23.662	4.467	<2e-16	***
s(x,y):Clade6B.1A.5a.1	19.619	21.519	11.115	<2e-16	***
s(x,y):Clade6B.1A.5a.2	17.246	19.255	4.006	<2e-16	***
s(x,y):Clade6B.1A.5b	20.538	23.043	4.881	<2e-16	***
s(x,y):Clade6B.1A.6	6.841	8.707	1.858	0.05447	.
s(x,y):Clade6B.1A.7	10.202	12.532	2.596	0.00169	**
s(x,y):OselR	5.05E-04	0.001001	0.001	0.999	
s(x,y):OselS	4.58E+00	5.541369	0.204	0.976	
s(x,y):OselU	4.14E+00	5.395094	0.203	0.976	
s(x,y):ZanaS	3.95E+00	4.660211	0.083	0.994	
s(x,y):ZanaU	4.14E+00	5.395094	0.203	0.976	
s(x,y):ContiAfrica	15.351	17.346	13.813	<2e-16	***
s(x,y):ContiAsia	19.113	23.179	5.311	<2e-16	***
s(x,y):ContiEurope	7.914	10.582	1.351	0.19019	
s(x,y):ContiNorth_America	2.033	2.047	0.168	0.8512	
s(x,y):ContiOceania	7.198	9.21	2.786	0.00235	**
s(x,y):ContiSouth_America	7.425	10.039	1.189	0.29331	
s(x,y):Age	9.769	13.044	1.399	0.14529	
s(x,y):totalSubstitutions	26.836	28.691	42.053	<2e-16	***
s(x,y):totalAminoacidSubstitutions	11.12	14.457	2.551	0.00122	**
s(x,y):distBR18	16.544	18.612	1.172	0.27107	
s(x,y):distGM19	3.002	3.003	5.227	0.00135	**
s(x,y):distHI19	17.034	18.386	1.697	0.03715	*
s(x,y):distVI19	23.224	24.903	6.987	<2e-16	***
s(x,y):distWS19	20.032	21.585	1.888	0.00778	**
s(Age)	1.001	1.002	0.007	0.937	
s(totalSubstitutions)	7.697	8.506	189.611	<2e-16	***
s(totalAminoacidSubstitutions)	8.362	8.823	29.577	<2e-16	***
s(distBR18)	7.648	8.423	14.352	<2e-16	***
s(distGM19)	7.07	7.831	13.221	<2e-16	***
s(distHI19)	7.17	7.876	6.953	<2e-16	***
s(distVI19)	7.124	8.111	8.124	<2e-16	***
s(distWS19)	7.49	8.231	14.349	<2e-16	***

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.4. Basis dimension checking for H1N1 isolates. Basis dimension indices for predictors for the GAM implemented for H1N1 isolate root to tip distance as the response variable with corresponding statistical support.

H1N1 GAM results

Basis dimension (k) checking

Predictor	k'	edf	k-index	p-value	Sig code
s(x,y)	29	17.41	0.69	<2e-16	***
s(x,y):Clade6B	29	3.03	0.69	<2e-16	***
s(x,y):Clade6B.1	29	7.57	0.69	<2e-16	***
s(x,y):Clade6B.1A	29	19.29	0.69	<2e-16	***
s(x,y):Clade6B.1A.1	29	18.85	0.69	<2e-16	***
s(x,y):Clade6B.1A.2	29	2	0.69	<2e-16	***
s(x,y):Clade6B.1A.3	29	13.52	0.69	<2e-16	***
s(x,y):Clade6B.1A.4	29	2	0.69	<2e-16	***
s(x,y):Clade6B.1A.5	29	2	0.69	<2e-16	***
s(x,y):Clade6B.1A.5a	29	21.28	0.69	<2e-16	***
s(x,y):Clade6B.1A.5a.1	29	19.62	0.69	<2e-16	***
s(x,y):Clade6B.1A.5a.2	29	17.25	0.69	<2e-16	***
s(x,y):Clade6B.1A.5b	29	20.54	0.69	<2e-16	***
s(x,y):Clade6B.1A.6	29	6.84	0.69	<2e-16	***
s(x,y):Clade6B.1A.7	29	10.2	0.69	<2e-16	***
s(x,y):OselR	2.90E+01	5.05E-04	0.41	<2e-16	***
s(x,y):OselS	2.90E+01	4.58E+00	0.41	<2e-16	***
s(x,y):OselU	2.90E+01	4.14E+00	0.41	<2e-16	***
s(x,y):ZanaS	2.90E+01	3.95E+00	0.41	<2e-16	***
s(x,y):ZanaU	2.90E+01	4.14E+00	0.41	<2e-16	***
s(x,y):ContiAfrica	29	15.35	0.41	<2e-16	***
s(x,y):ContiAsia	29	19.11	0.41	<2e-16	***
s(x,y):ContiEurope	29	7.91	0.41	<2e-16	***
s(x,y):ContiNorth_America	29	2.03	0.41	<2e-16	***
s(x,y):ContiOceania	29	7.2	0.41	<2e-16	***
s(x,y):ContiSouth_America	29	7.42	0.41	<2e-16	***
s(x,y):Age	30	9.77	0.72	<2e-16	***
s(x,y):totalSubstitutions	30	26.84	0.72	<2e-16	***
s(x,y):totalAminoacidSubstitutions	30	11.12	0.72	<2e-16	***
s(x,y):distBR18	30	16.54	0.72	<2e-16	***
s(x,y):distGM19	30	3	0.72	<2e-16	***
s(x,y):distHI19	30	17.03	0.72	<2e-16	***
s(x,y):distVI19	30	23.22	0.72	<2e-16	***
s(x,y):distWS19	30	20.03	0.72	<2e-16	***
s(Age)	9	1	0.98	0.11	
s(totalSubstitutions)	9	7.7	0.81	<2e-16	***
s(totalAminoacidSubstitutions)	9	8.36	0.76	<2e-16	***
s(distBR18)	9	7.65	0.73	<2e-16	***
s(distGM19)	9	7.07	0.73	<2e-16	***
s(distHI19)	9	7.17	0.73	<2e-16	***
s(distVI19)	9	7.12	0.7	<2e-16	***
s(distWS19)	9	7.49	0.71	<2e-16	***

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.5. Significance of smoothing terms for B-Vic isolates. Predictors for the GAM implemented for B-Vic isolate root to tip distance as the response variable with corresponding statistical support.

B-Vic GAM results

Significance of smooth terms

Position	edf	Ref.df	F	p-value	Sig code
s(x,y)	27.22052	28.675	37.661	<2e-16	***
s(x,y):CladeV1A	19.215	21.514	13.385	<2e-16	***
s(x,y):CladeV1A.1	13.054	14.828	23.964	<2e-16	***
s(x,y):CladeV1A.2	7.697	9.157	4.022	3.69E-05	***
s(x,y):CladeV1A.3	18.485	20.982	13.04	<2e-16	***
s(x,y):CladeV1A.3a	1.123	1.602	0.416	0.50576	
s(x,y):CladeV1A.3a.1	5.365	6.333	3.432	0.00179	**
s(x,y):CladeV1A.3a.2	5.575	6.565	2.403	0.01525	*
s(x,y):OselS	2.006644	2.012	1.432	0.24	
s(x,y):OselU	0.698478	1.202	0.031	0.937	
s(x,y):ZanaS	0.006644	0.012	0.043	0.982	
s(x,y):ZanaU	0.698478	1.202	0.031	0.937	
s(x,y):ContiAfrica	13.327	17.102	1.816	0.0212	*
s(x,y):ContiAsia	22.974	26.043	4.012	<2e-16	***
s(x,y):ContiEurope	2.004	2.006	1.44	0.2368	
s(x,y):ContiNorth_America	10.915	13.573	0.491	0.9552	
s(x,y):ContiOceania	4.779	5.59	3.303	0.0495	*
s(x,y):ContiSouth_America	8.346	11.232	1.379	0.1577	
s(x,y):Age	10.592	14.05	0.763	0.7082	
s(x,y):totalSubstitutions	20.436	23.3	27.095	<2e-16	***
s(x,y):totalAminoacidSubstitutions	19.423	23.41	7.47	<2e-16	***
s(x,y):distWA19	19.547	22.26	1.6	0.0364	*
s(Age)	1.783	2.211	0.836	0.428	
s(totalSubstitutions)	4.305	5.356	111.089	<2e-16	***
s(totalAminoacidSubstitutions)	6.909	7.906	23.172	<2e-16	***
s(distWA19)	3.884	4.961	0.66	0.699	

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.6. Basis dimension checking for B-Vic isolates. Basis dimension indices for predictors for the GAM implemented for B-Vic isolate root to tip distance as the response variable with corresponding statistical support.

B-Vic GAM results

Basis dimension (k) checking

Position	k'	edf	k-index	p-value	Sig code
s(x,y)	29	2.01	0.85	<2e-16	***
s(x,y):CladeV1A	29	19.22	0.85	<2e-16	***
s(x,y):CladeV1A.1	29	13.05	0.85	<2e-16	***
s(x,y):CladeV1A.2	29	7.7	0.85	<2e-16	***
s(x,y):CladeV1A.3	29	18.49	0.85	<2e-16	***
s(x,y):CladeV1A.3a	29	1.12	0.85	<2e-16	***
s(x,y):CladeV1A.3a.1	29	5.36	0.85	<2e-16	***
s(x,y):CladeV1A.3a.2	29	5.57	0.85	<2e-16	***
s(x,y):OselS	29	2.00664	0.84	<2e-16	***
s(x,y):OselU	29	0.69848	0.84	<2e-16	***
s(x,y):ZanaS	29	0.00664	0.84	<2e-16	***
s(x,y):ZanaU	29	0.69848	0.84	<2e-16	***
s(x,y):ContiAfrica	29	13.33	0.87	<2e-16	***
s(x,y):ContiAsia	29	22.97	0.87	<2e-16	***
s(x,y):ContiEurope	29	2	0.87	<2e-16	***
s(x,y):ContiNorth_America	29	10.92	0.87	<2e-16	***
s(x,y):ContiOceania	29	4.78	0.87	<2e-16	***
s(x,y):ContiSouth_America	29	8.35	0.87	<2e-16	***
s(x,y):Age	30	10.6	0.94	<2e-16	***
s(x,y):totalSubstitutions	30	20.4	0.94	0.005	**
s(x,y):totalAminoacidSubstitutions	30	19.4	0.94	<2e-16	***
s(x,y):distWA19	30	19.6	0.94	0.005	**
s(Age)	9	1.78	0.96	0.015	*
s(totalSubstitutions)	9	4.31	0.81	<2e-16	***
s(totalAminoacidSubstitutions)	9	6.91	0.8	<2e-16	***
s(distWA19)	9	3.88	0.91	<2e-16	***

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Table S3.7. Significance of smoothing terms for B-Yam isolates. Predictors for the GAM implemented for B-Yam isolate root to tip distance as the response variable with corresponding statistical support.

B-Yam GAM results

Significance of smooth terms

Predictor	edf	Ref.df	F	p-value	Sig code
s(x,y)	24.170192	26.452804	20.744	<2e-16	***
s(x,y):OselS	0.004338	0.006541	0.003	0.996	
s(x,y):OselU	5.758717	7.118433	0.919	0.537	
s(x,y):ZanaS	0.004338	0.006541	0.003	0.996	
s(x,y):ZanaU	3.758717	5.118433	0.141	0.986	
s(x,y):ContiAfrica	2.001276	2.00243	22.883	<2e-16	***
s(x,y):ContiAsia	7.077669	9.65122	7.004	<2e-16	***
s(x,y):ContiEurope	2.000328	2.00062	12.931	3.02E-06	***
s(x,y):ContiNorth_America	7.861852	10.56783	3.191	0.000409	***
s(x,y):ContiOceania	0.007441	0.01411	0	0.5	
s(x,y):ContiSouth_America	6.268093	8.50093	3.711	0.000222	***
s(x,y):Age	11.039	14.55	2.646	0.000707	***
s(x,y):totalSubstitutions	9.162	10.6	31.174	<2e-16	***
s(x,y):totalAminoacidSubstitutions	11.896	14.03	4.203	<2e-16	***
s(Age)	2.393	2.983	4.822	0.00216	**
s(totalSubstitutions)	5.813	6.881	67.804	<2e-16	***

(Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ’ ’ 1)

Table S3.8. Basis dimension checking for B-Yam isolates. Basis dimension indices for predictors for the GAM implemented for B-Yam isolate root to tip distance as the response variable with corresponding statistical support.

B-Yam GAM results

Basis dimension (k) checking

Predictor	k'	edf	k-index	p-value	Sig code
s(x,y)	29	24.17019	0.44	<2e-16	***
s(x,y):OselS	29	0.00434	0.44	<2e-16	***
s(x,y):OselU	29	5.75872	0.44	<2e-16	***
s(x,y):ZanaS	29	0.00434	0.44	<2e-16	***
s(x,y):ZanaU	29	3.75872	0.44	<2e-16	***
s(x,y):ContiAfrica	29	2.00128	0.43	<2e-16	***
s(x,y):ContiAsia	29	7.07767	0.43	<2e-16	***
s(x,y):ContiEurope	29	2.00033	0.43	<2e-16	***
s(x,y):ContiNorth_America	29	7.86185	0.43	<2e-16	***
s(x,y):ContiOceania	29	0.00744	0.43	<2e-16	***
s(x,y):ContiSouth_America	29	6.26809	0.43	<2e-16	***
s(x,y):Age	30	11.04	0.52	<2e-16	***
s(x,y):totalSubstitutions	30	9.16	0.52	<2e-16	***
s(x,y):totalAminoacidSubstitutions	30	11.9	0.52	<2e-16	***
s(Age)	9	2.39	0.97	0.095	.
s(totalSubstitutions)	9	5.81	0.84	<2e-16	***

(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

Code availability. All BEAST XMLs, and code used for analyses and visualization of results are provided in the following GitHub repository: https://github.com/ldamodaran/Influenza-antigenic-2017-2022

Supplemental Figures Chapter 4: Comparison of estimated antigenic space between different laboratory assays for H3N2 influenza viruses

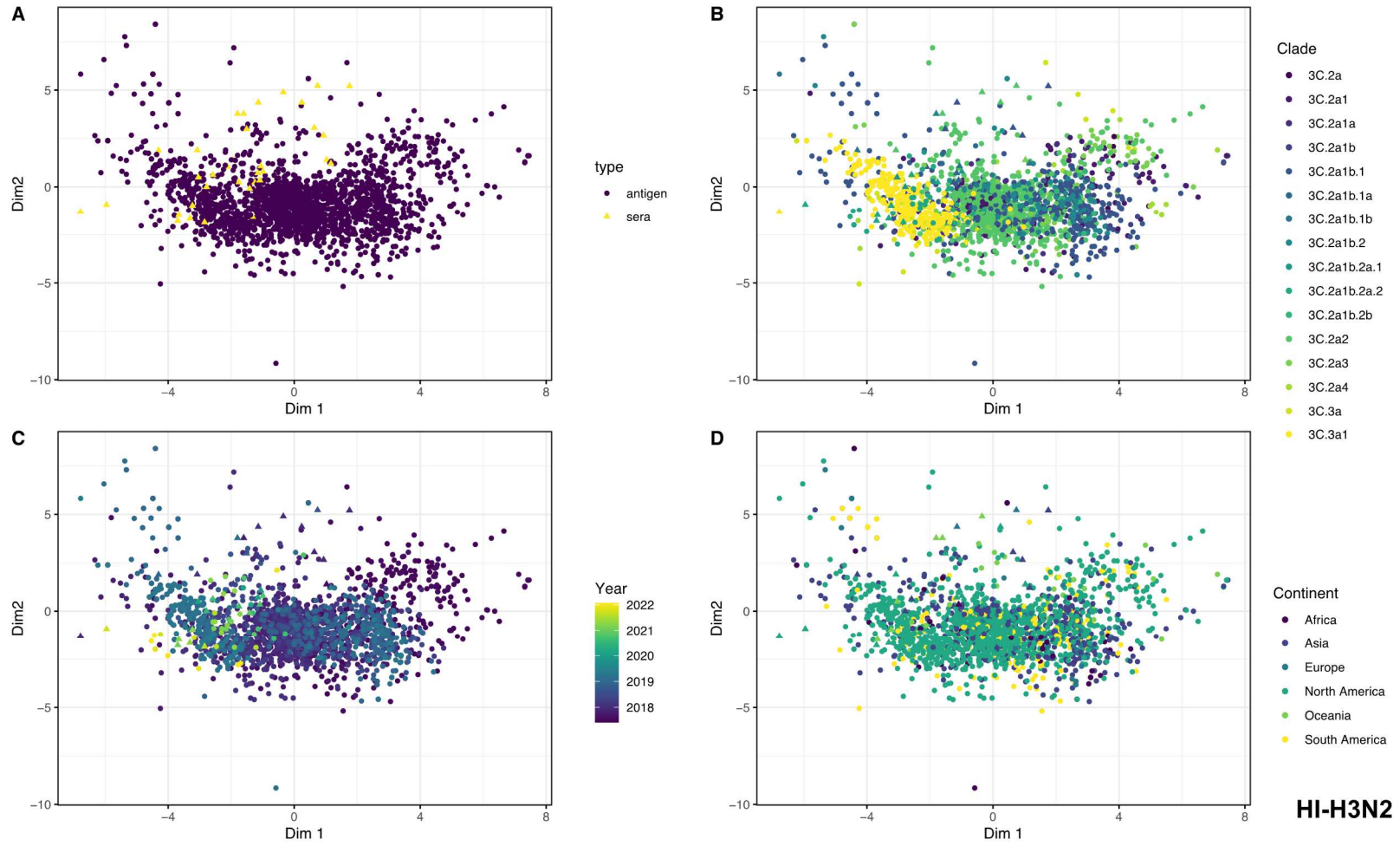


Figure S4.1. Antigenic cartography of HI assay data collected between 2017 and 2022 colored by discrete associated metadata. A) Colored by antigen and sera. B) Colored by lineage as determined by NextClade. C) Colored by year of isolation. D) Colored by continent of isolation.

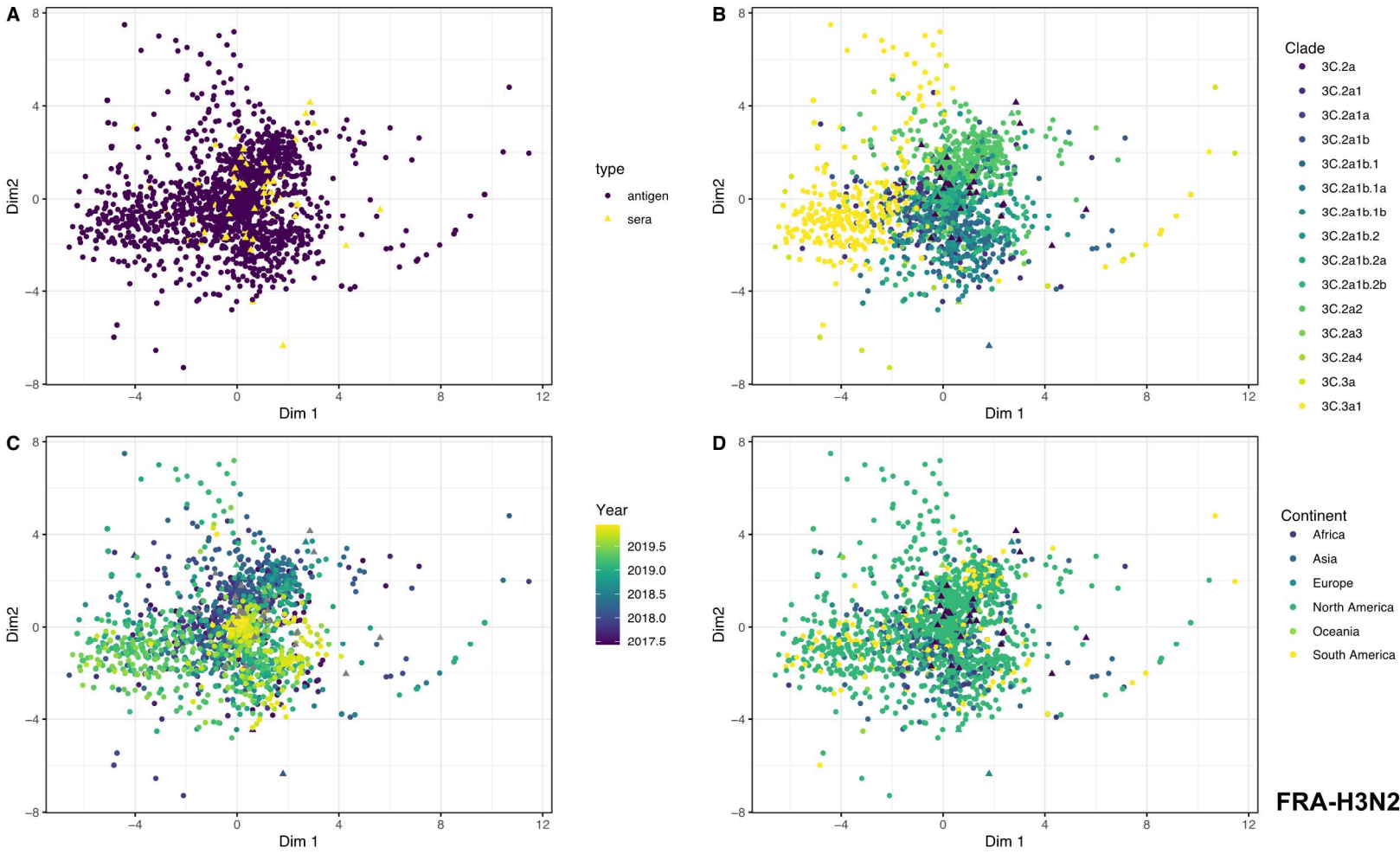


Figure S4.2. Antigenic cartography of FRA assay data collected between 2017 and 2022 colored by discrete associated metadata. A) Colored by antigen and sera. B) Colored by lineage as determined by NextClade. C) Colored by year of isolation. D) Colored by continent of isolation.

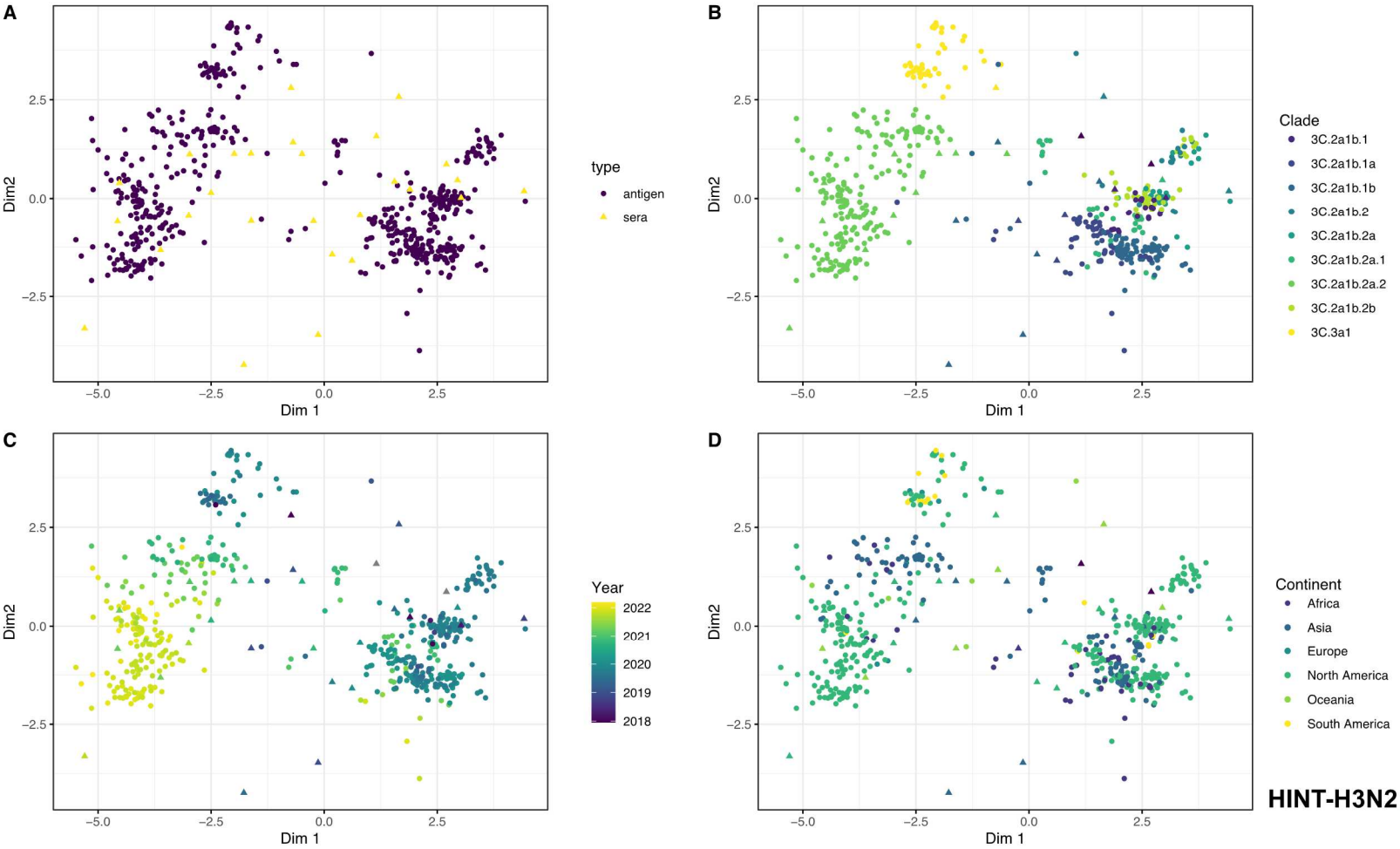


Figure S4.3. Antigenic cartography of HINT assay data collected between 2017 and 2022 colored by discrete associated metadata. A) Colored by antigen and sera. B) Colored by lineage as determined by NextClade. C) Colored by year of isolation. D) Colored by continent of isolation.

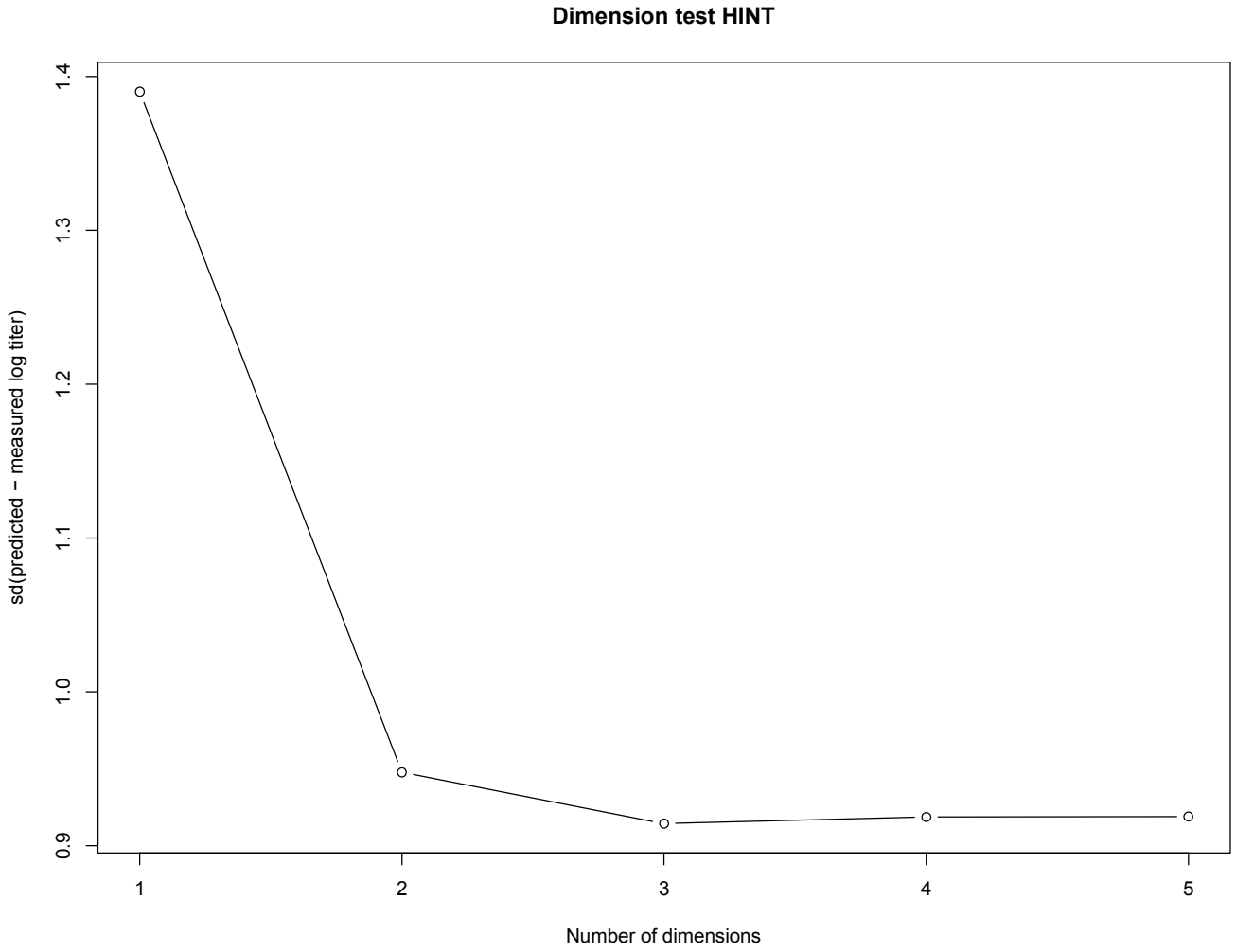


Figure S4.4. Scree plot for the dimension test of HI assay data. Dimension test was carried out for 5 dimensions with 100 optimizations and 100 replicates for each dimension of data.

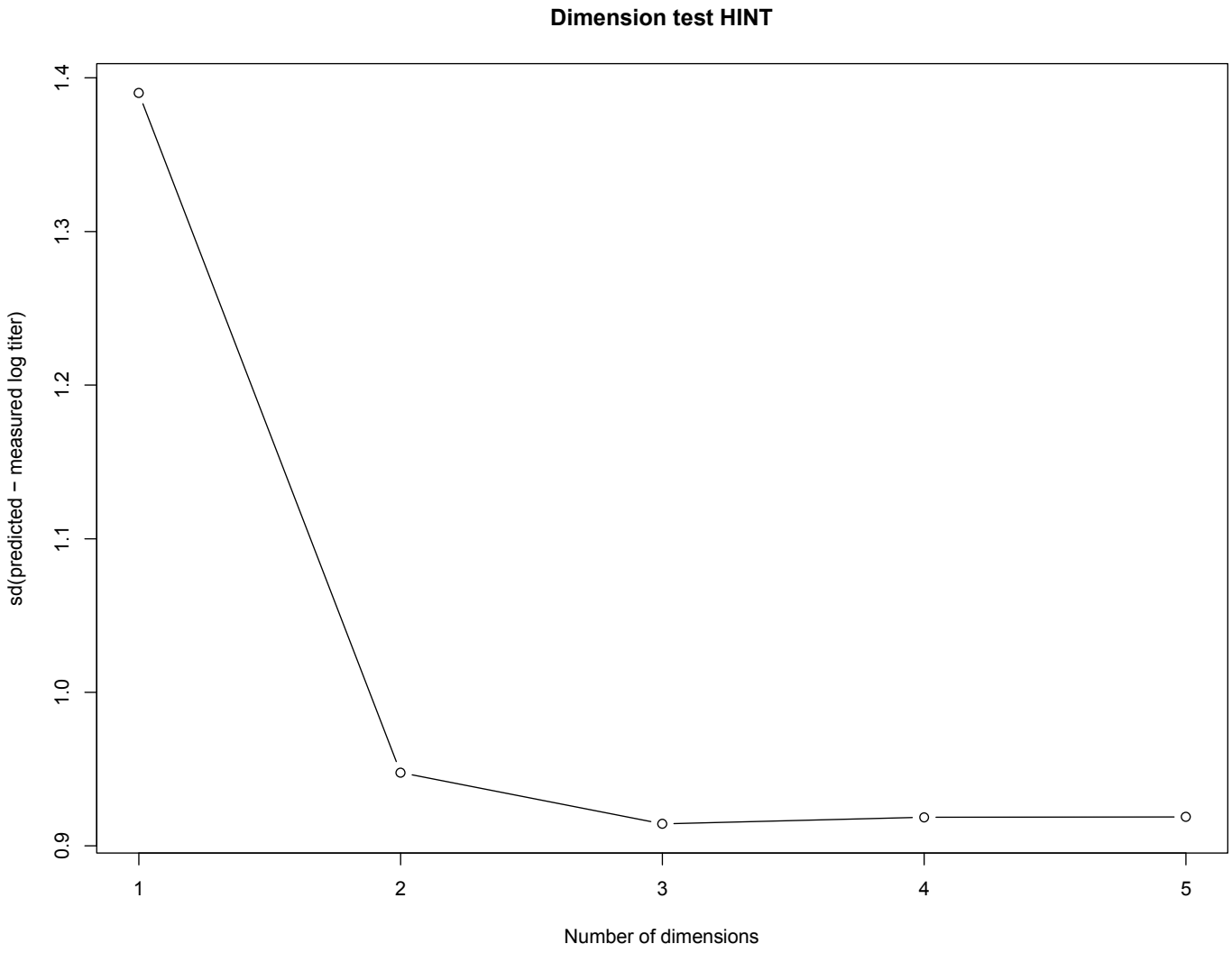


Figure S4.5. Scree plot for the dimension test of FRA assay data. Dimension test was carried out for 5 dimensions with 100 optimizations and 100 replicates for each dimension of data.

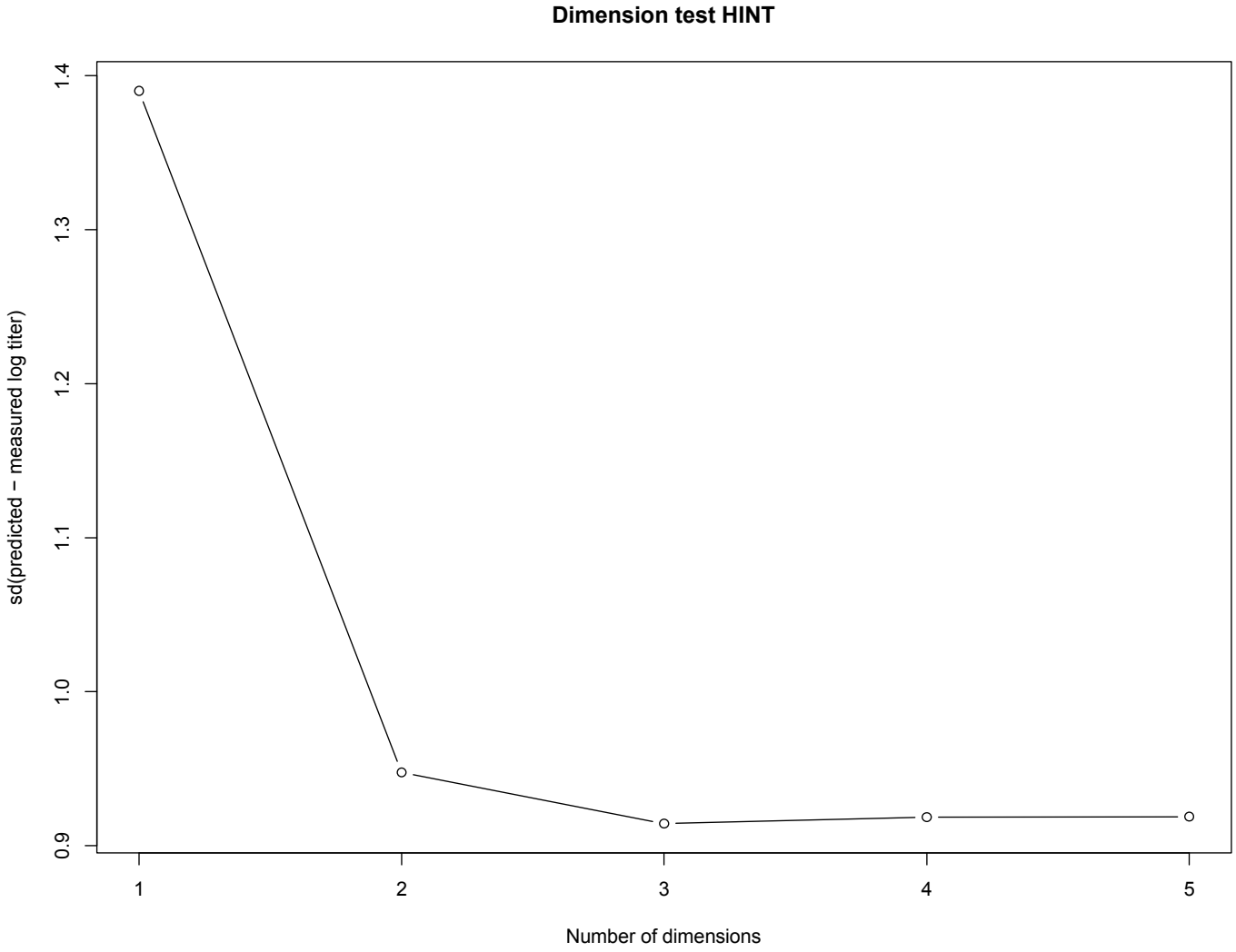


Figure S4.6. Scree plot for the dimension test of HINT assay data. Dimension test was carried out for 5 dimensions with 100 optimizations and 100 replicates for each dimension of data.

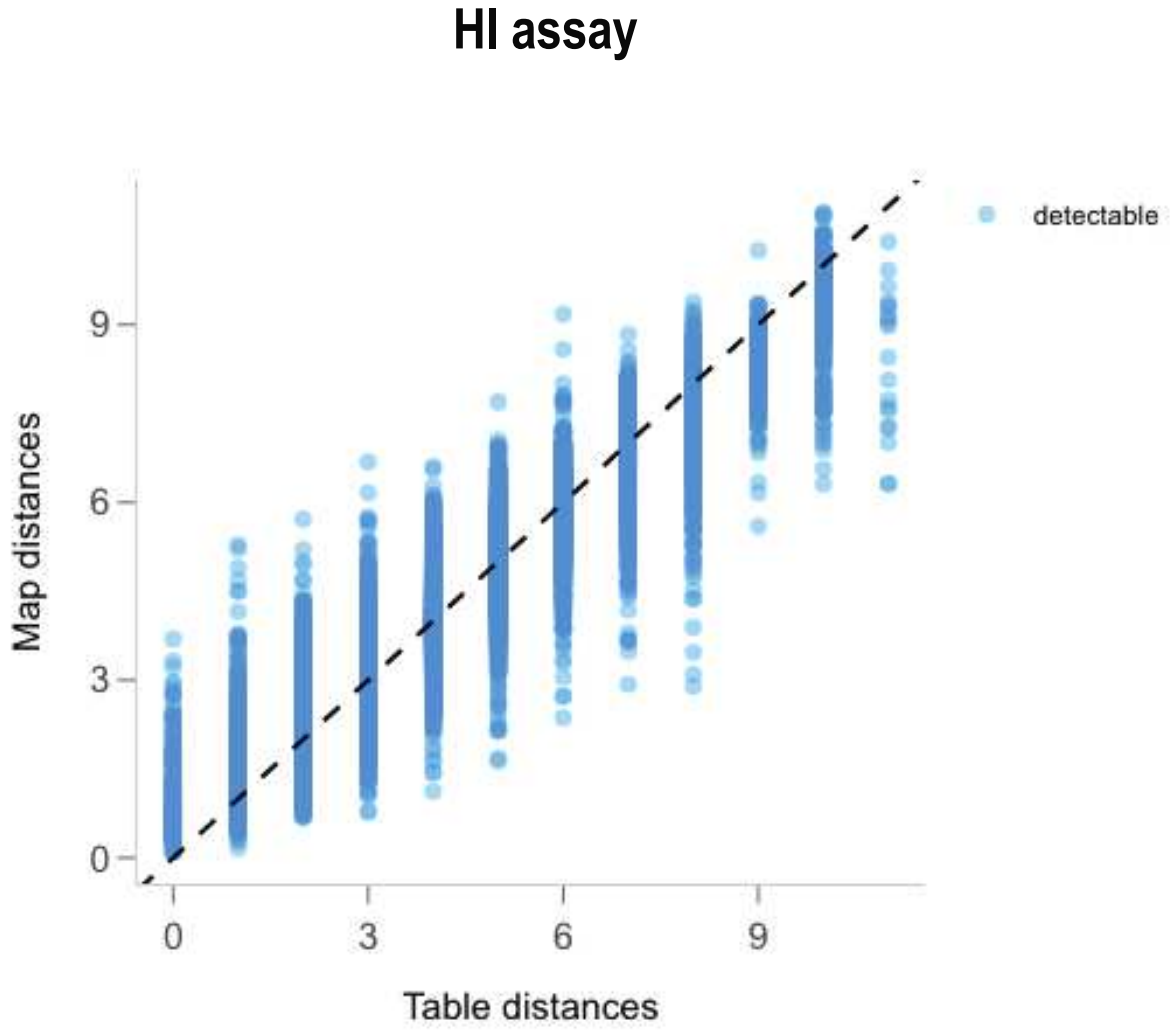


Figure S4.7. Shepard stress plot for table distances vs map distances for HI assay data. Dotted line represents the linear regression of the datapoints.

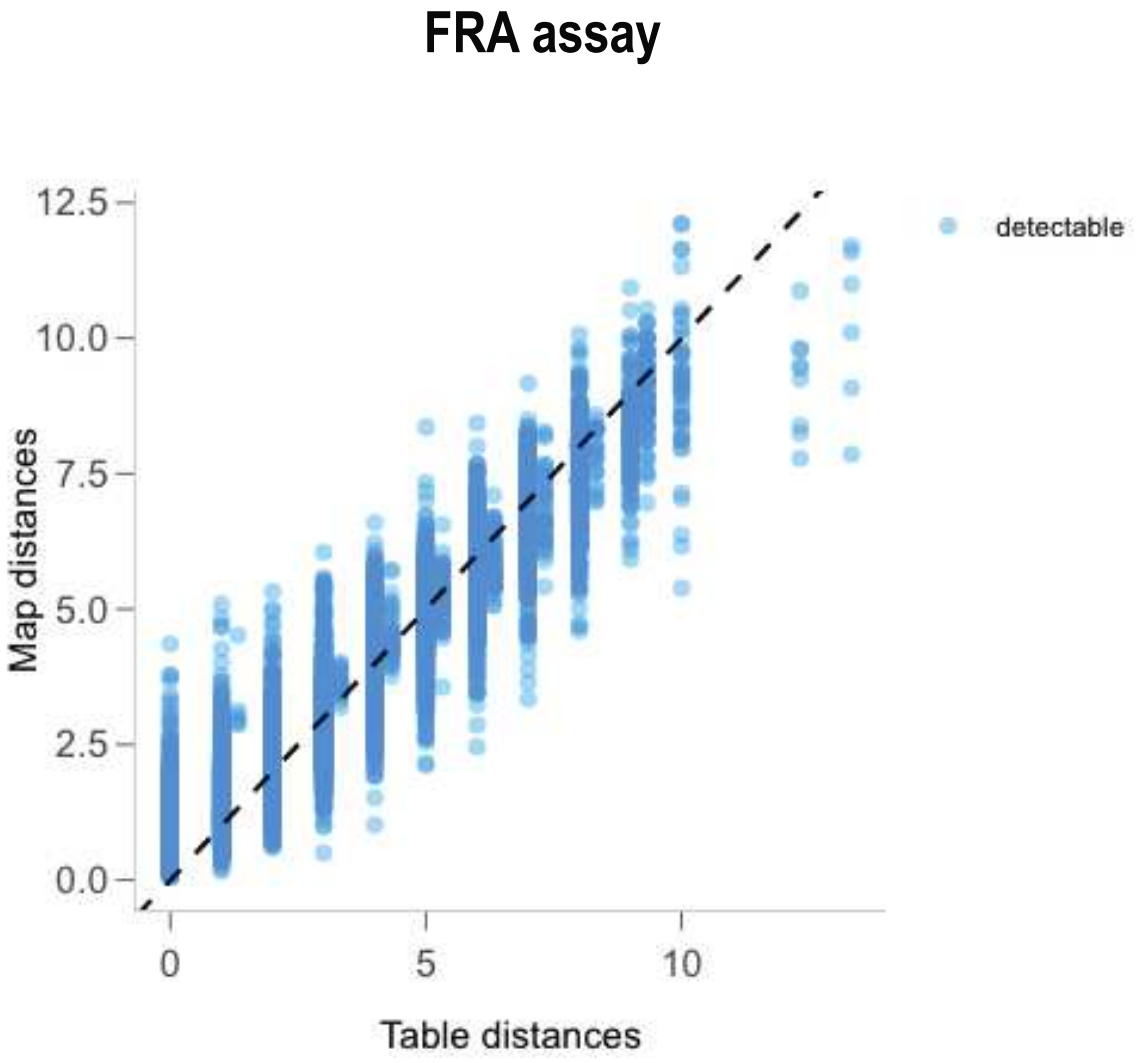


Figure S4.8. Shepard stress plot for table distances vs map distances for FRA assay data. Dotted line represents the linear regression of the datapoints.

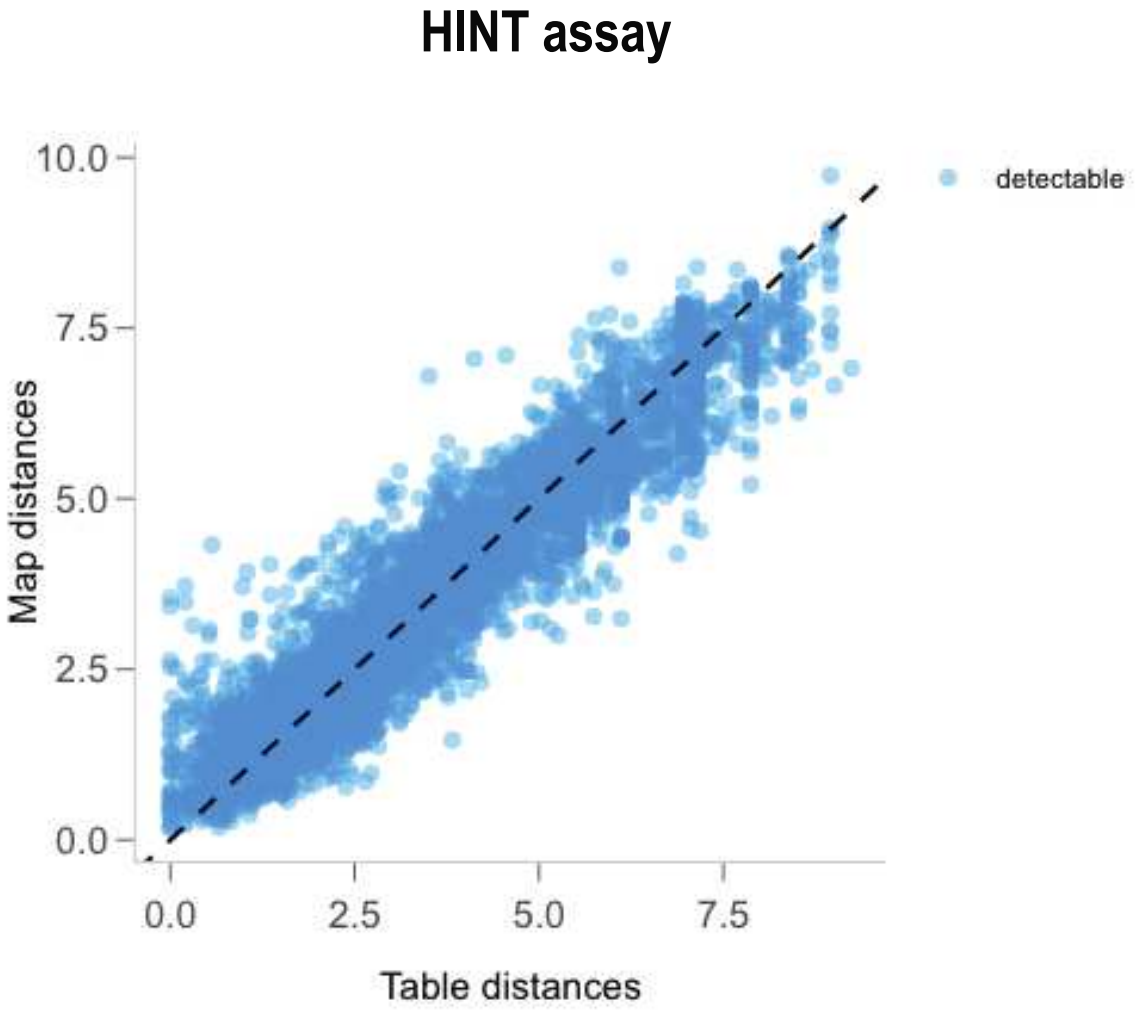


Figure S4.8. Shepard stress plot for table distances vs map distances for HINT assay data. Dotted line represents the linear regression of the datapoints.

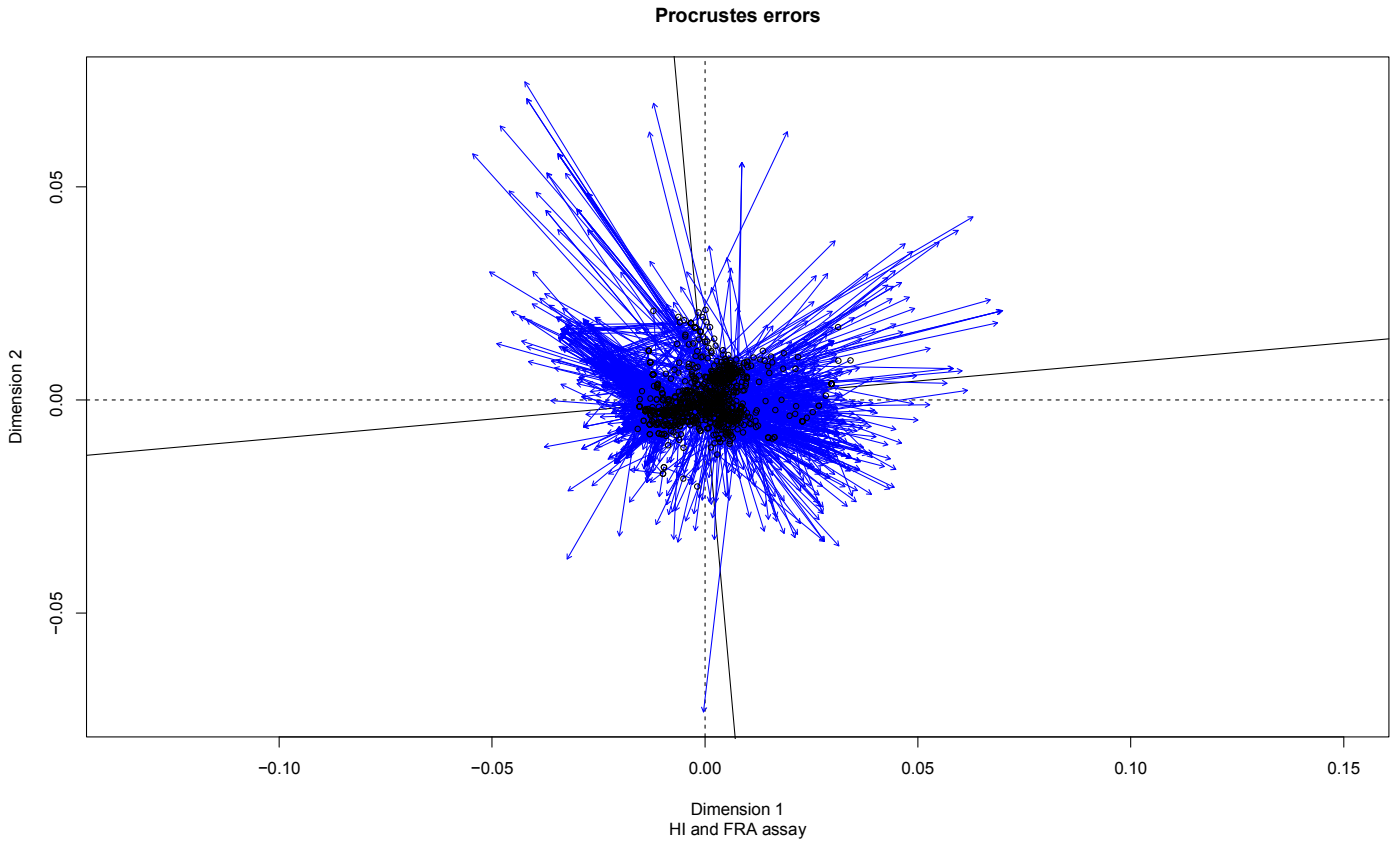


Figure S4.10. Procrustes errors for the comparison of HI and FRA assays. Blue arrows represent the position in the second assay for a given isolate in the first assay being compared.

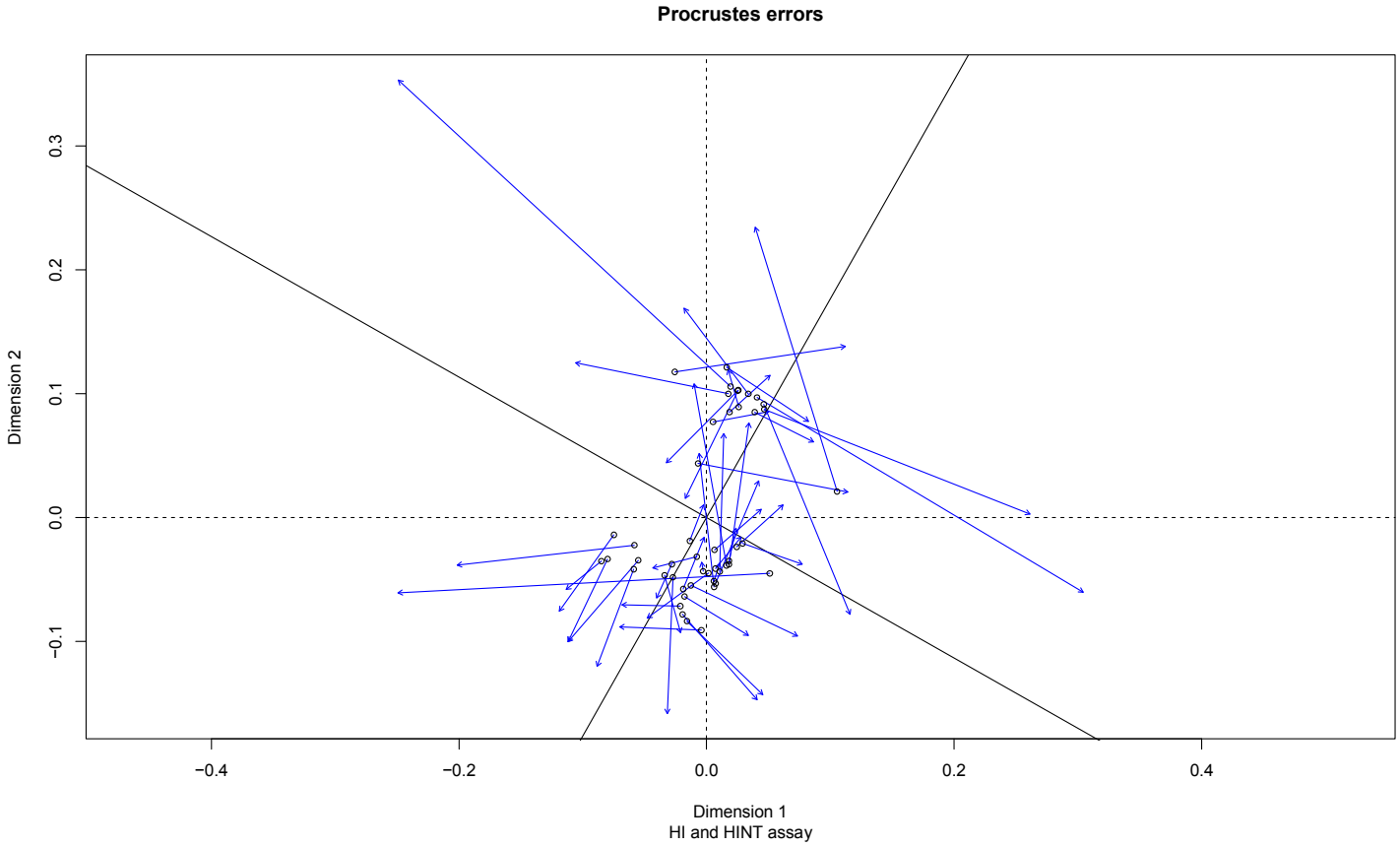


Figure S4.11. Procrustes errors for the comparison of HI and HINT assays. Blue arrows represent the position in the second assay for a given isolate in the first assay being compared.

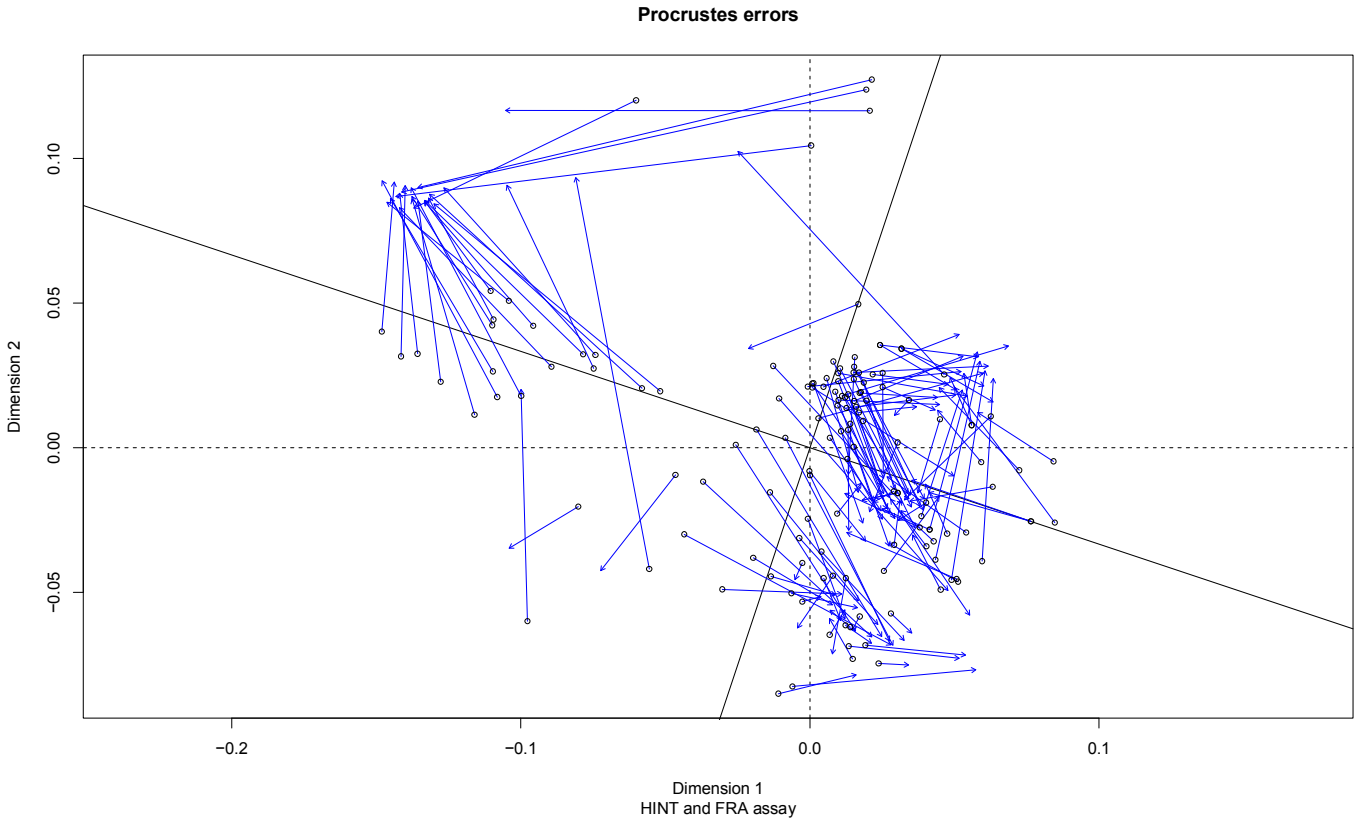


Figure S4.12. Procrustes errors for the comparison of HINT and FRA assays. Blue arrows represent the position in the second assay for a given isolate in the first assay being compared.

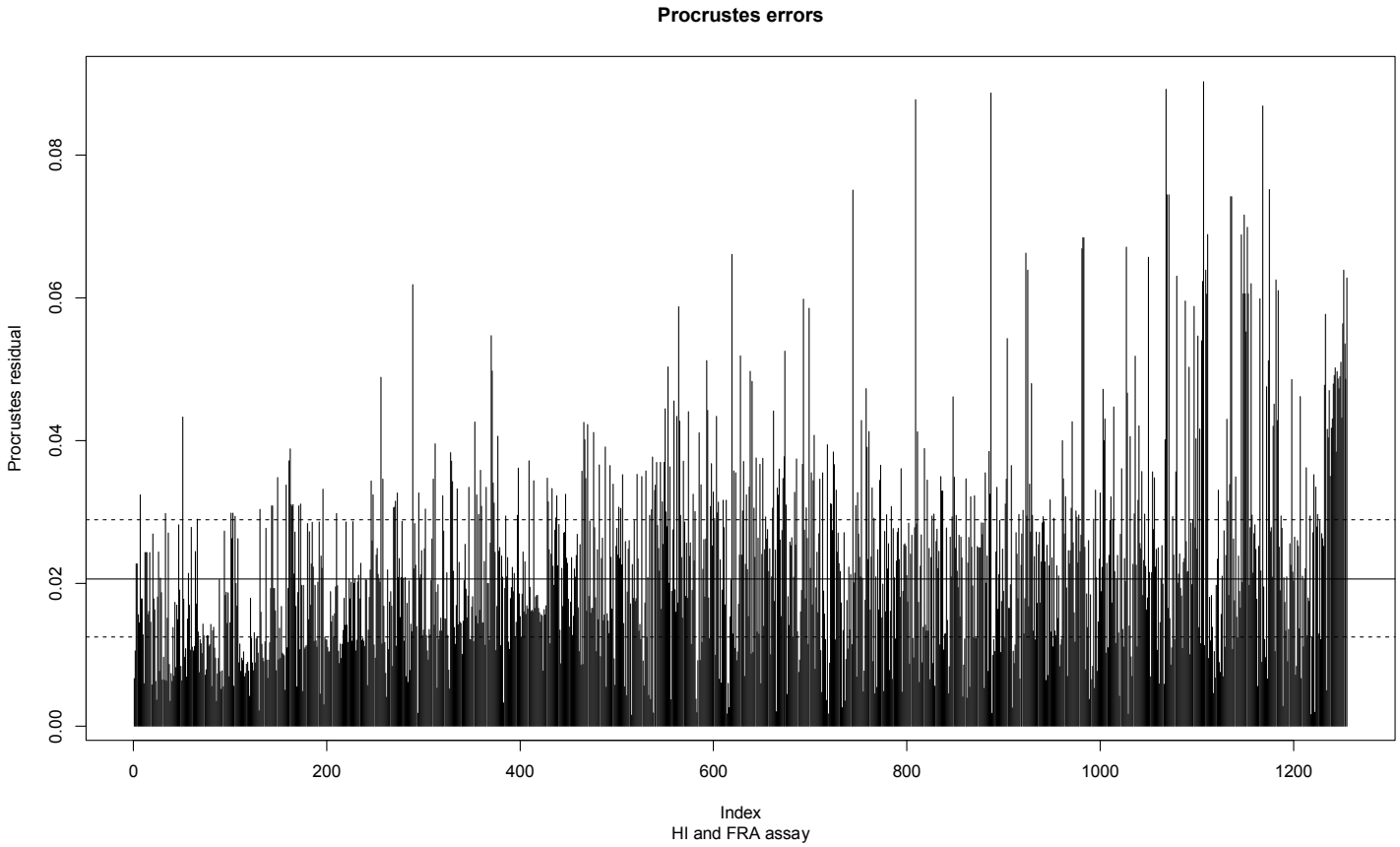


Figure S4.13. Procrustes residuals for individual isolates between the HI and FRA assays. The index represents a given isolate and the horizontal line represents the mean Procrustes residual value and 95% confidence interval.

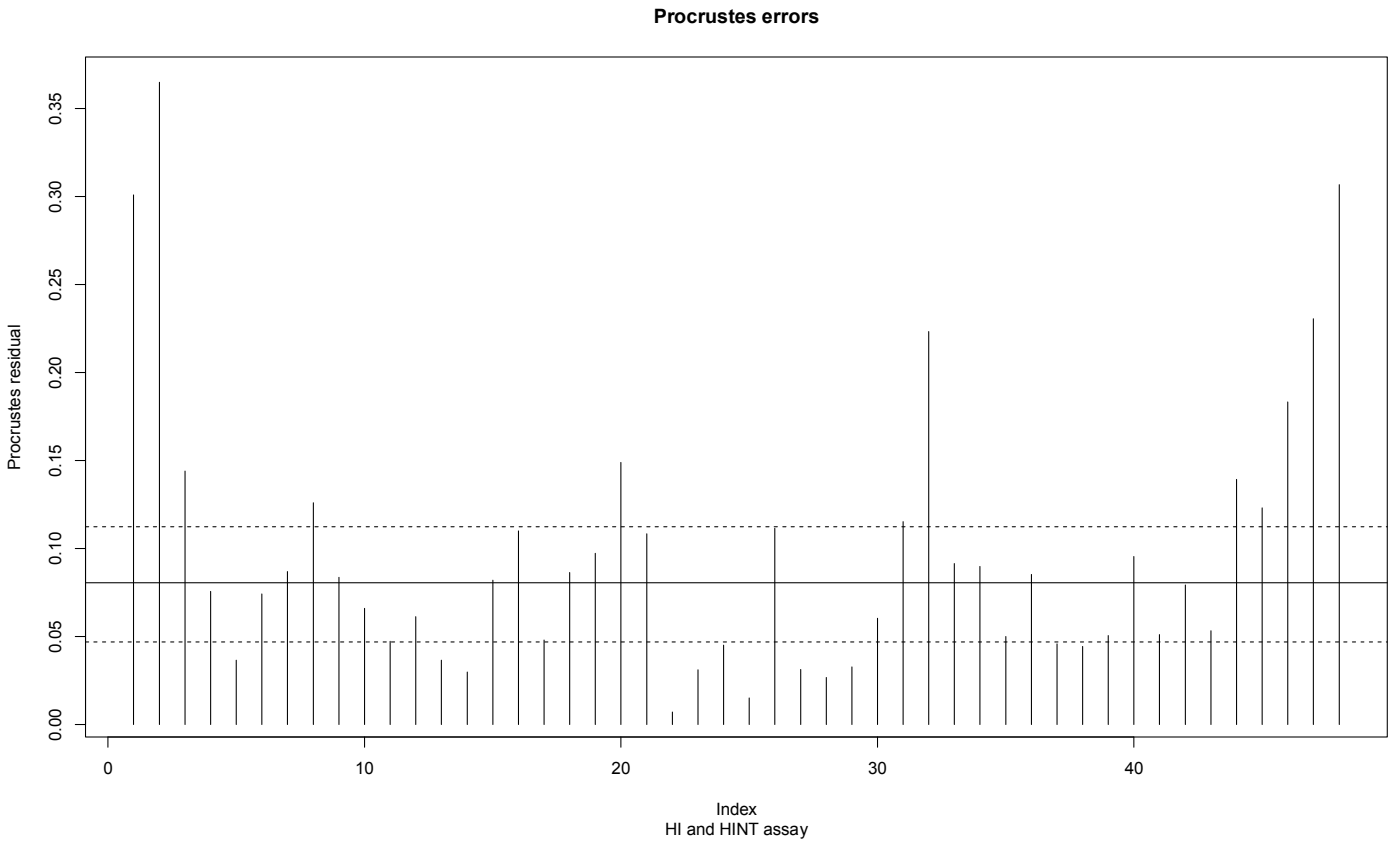


Figure S4.14. Procrustes residuals for individual isolates between the HI and HINT assays. The index represents a given isolate and the horizontal line represents the mean Procrustes residual value and 95% confidence interval.

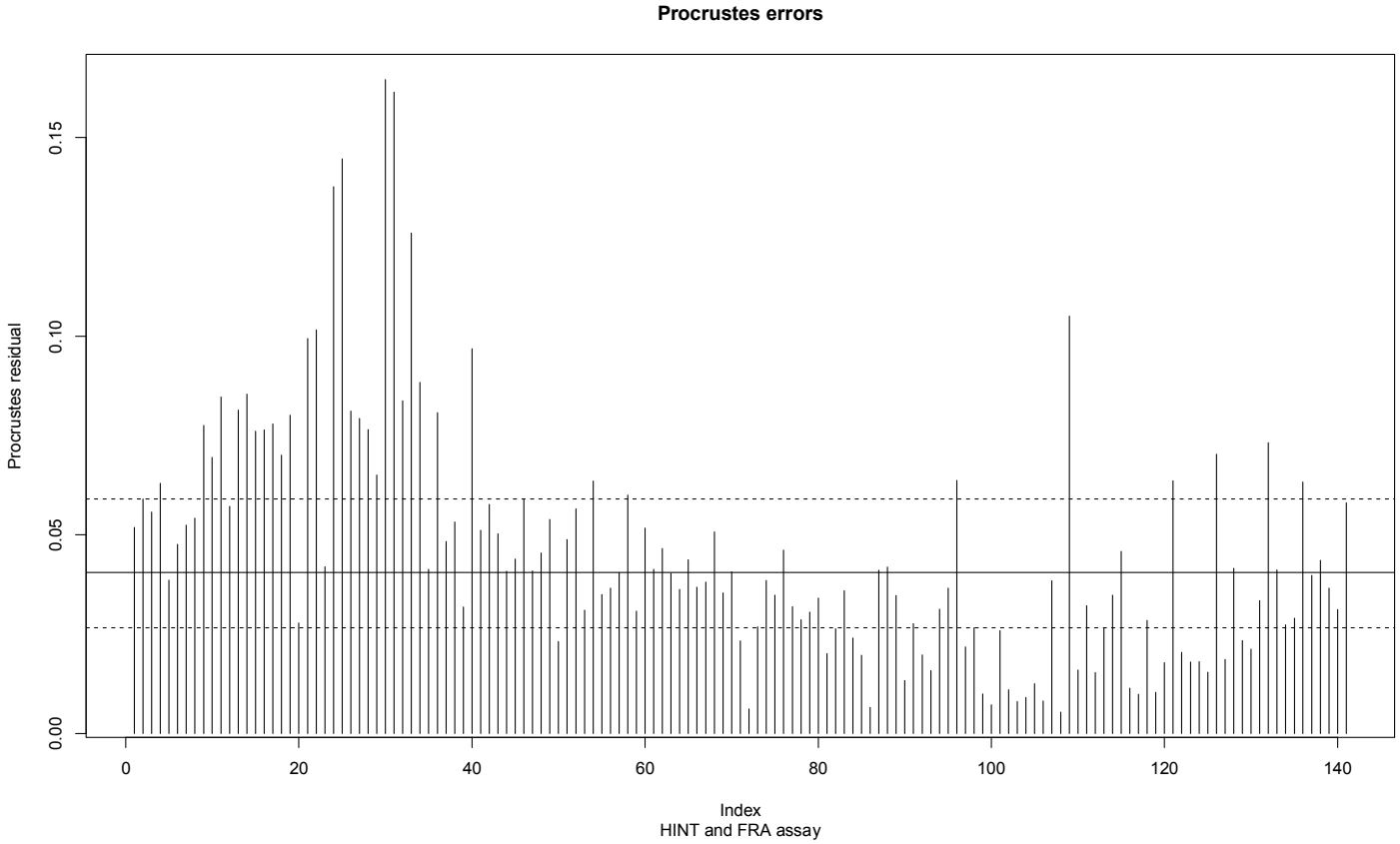


Figure S4.15. Procrustes residuals for individual isolates between the HINT and FRA assays. The index represents a given isolate and the horizontal line represents the mean Procrustes residual value and 95% confidence interval.

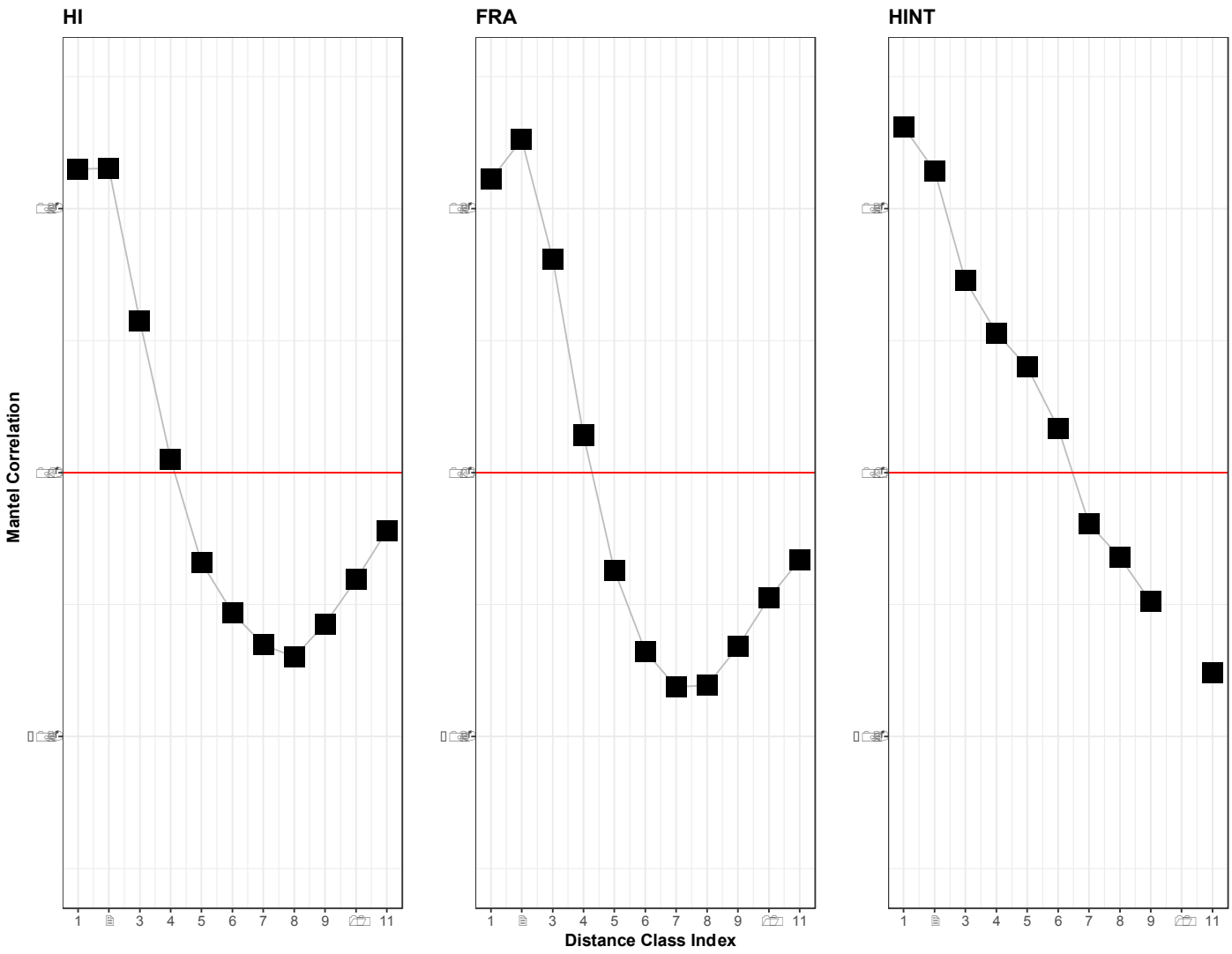


Figure S4.16. Mantel correlogram for antigenic assays tested against corresponding phylogenies for HA1 genomic region of the HA protein. The non-significant correlation values were not plotted for a given distance class index.