APPLIED PHYLODYNAMIC MODELING OF INFLUENZA VIRUSES TO ENHANCE THE UNDERSTANDING OF VIRAL EVOLUTION AND DIFFUSION DYNAMICS

by

XUETING QIU

(Under the Direction of Justin Bahl)

ABSTRACT

Influenza continues to pose global public health threats, including several pandemics in history, heavy disease burdens in humans with annual seasonal outbreaks, and high economic loss with highly pathogenic virus causing large outbreaks in poultry. In addition, seasonal influenza requires tremendous efforts to update and distribute vaccines annually, while zoonosis of potential pandemic strains poses another threat of no stockpile of effective vaccines. To overcome the predicament of influenza vaccine and disease prevention, phylogenetic modeling with currently advanced mathematical models, global genomic data sharing and increased computing capability becomes a powerful tool to understand rapidly evolving pathogens and ultimately achieve the goal of effective prevention. In this dissertation, I aimed to develop novel models and apply advanced models to study the evolutionary and epidemiological dynamics of influenza virus across ecological scales in order to improve disease control. Specifically, aim 1 focused on developing a model to incorporating hemagglutinin (HA) protein structure to better understand the evolution of vaccine targets. The new phylogenetic model that accounted for rate variations across protein structural domains and across codon positions significantly improved the reconstruction of influenza viruses. It revealed valuable biological insights on protein

structural domain-specific evolutionary characteristics and approximate selection pressure on these domains, which can provide new approach for broadly-reactive vaccine design. Aim 2 explored viral diffusion patterns and ecological factors that potentially affect the diffusion in the U.S. via phylodynamic modeling. It identified regions with busiest airports played as a primary hub for viral diffusions in the U.S. Higher proportion of high-risk populations including the youth and the elderly and more flight connections may significantly increase viral migration rates. Aim 3 explored the impacts of H3N2 live attenuated influenza vaccine (LAIV) on viral genetic diversity and diffusion dynamics in Central Texas via discrete trait analysis and structured coalescent model. The vaccinated population needed more external introductions to sustain the epidemics and disseminated less to external regions, which provided the phylogenetic evidence of vaccination benefits. Taken together, findings from these studies provided instructive insights/recommendations on vaccine design, administration strategy and effective prevention measures of influenza viruses.

INDEX WORDS: Phylodynamic modeling, Influenza viruses, Viral evolution, Diffusion dynamics, Vaccination

APPLIED PHYLODYNAMIC MODELING OF INFLUENZA VIRUSES TO ENHANCE THE UNDERSTANDING OF VIRAL EVOLUTION AND DIFFUSION DYNAMICS

by

XUETING QIU

BM, Shanghai Jiao Tong University, China, 2013

MS, The University of Texas Health Science Center at Houston, 2015

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

2019

APPLIED PHYLODYNAMIC MODELING OF INFLUENZA VIRUSES TO ENHANCE THE UNDERSTANDING OF VIRAL EVOLUTION AND DIFFUSION DYNAMICS

by

XUETING QIU

Major Professor:	Justin Bahl
Committee:	Pedro A. Piedra
	Pejman Rohani
	Liliana Salvador
	Stephen M. Tompkins

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
December 2019

DEDICATION

To my dearest sister, XueRong Qiu. For fully supporting me to pursue my dreams and for being the one whom I can always turn to throughout the years. She has sustained me in ways that I never knew I needed.

ACKNOWLEDGEMENTS

This dissertation is a long journey from China to the US, from Houston to Athens, with so many great people and opportunities preparing me, which I cannot take all the credits. The highlighted memories are the organic complement of my science life to make it a whole journey. I will first say, since I have been completely SERIOUS about the scientific part, I am writing some part of the acknowledgement with a humor sense, just in case you cannot get it (don't blame yourself – I have pretty clueless jokes).

Firstly, a special BIG thank-you goes to Joe Hicks. Such a fortune to have a labmate and best friend at the same time! Joe is one of the most patient, kind and talented people I have met in my life. And he is the one who can get all my jokes (even I did not intend to say one) and thinks I am funnier than I ever thought I could be. This brought so many good laughs together that benefited our cardiac health. From courses to lab work, he is a role model that inspires me to pursue high standards in our work. I guess not so many friends can speak so many types of languages together: technical languages, daily English, culture and history, books and movies, Ghibli anime, nerdy medical/statistical jokes, sometimes Chinese in countable but fun enough phrases/sentences. With three years in Houston and one year in Athens, all the exploration and adventures are just unbeatable. While he saw all my clumsiness and awkward teary moments, it turns out we are still best friends! Ph.D. life is challenging but turns out to be fun with this guy. After we have separately headed to our next stage of adventures, I miss his snowballs the most (notes – snowball: one of the best sweets Joe makes).

I fortunately have met Avette and Rolando Farias while I was in Houston. And I always think of them as my US parents, whose love and encouragement have always lifted my spirit to handle challenges. Like most parents, they help in any ways that they can to make me worry-free: they have taken my old car to a full inspection and fixed everything before I drove to Athens from Houston; they unceasingly confirmed that I kept warm by mailing me hot teas, tea mugs and snuggling blanket during the winter in Athens (for a previous Houstonian, it is real winter); and they have tolerated my one-hour dissertation presentation without knowing much background of my work. The most special thing they have given me is the feeling that I am at home when I am with them.

My Ph.D. story started with Jessie Wang, my first roommate (later best friend) when I came to the US. She patiently helped me with so many big or small things while my English was not that functional. No matter I admit it or not, she is the one who first paid the adoption fee of my cat (which I paid back later to make sure there was no conflict of interests) while I was hesitating to become a cat person (I thought I was more a dog person). It tells the truth that sometimes our friends know us better. It is a great thing if you have a knowing-you-better friend!

Amazingly I have more. Boshi Yang, a high-tech (=strange) friend who learned Chinese via the similar ways as learning mathematics which I cannot figure out how that worked, patiently guides me how to tell a good story. He has been the first person to emphasize that telling good stories is an essential skill that I should master. He has been so strictly and unstoppingly taking the chance to improve my critical thinking and presentation skills by almost every conversation we had. If possible, please allow me to show my hidden googly eyes here.

Many Houston friends beyond what I can list here have supported me along the years. One representative would be Linda Greg and her adorable daughter Eliana, who have given me

so many pure joy moments. Linda always thinks that my accent is cute, which encouraged me to overcome my fear to talk in front of people during the early years. She just has the magic power to see people's strengths and always be a cheering-up friend. Another similar friend is William Dunn, who has an extra magic skill – artistically packing/unpacking and organization, which is critical for relocation. He kindly drove with both Joe and me last year from Houston to Athens and made the relocation and settle-down very smooth, where the last year of our Ph.D. started.

The opportunity to meet two great woman scientists, Liliana Salvador and Ana Bento, in Athens is a great treasure for me. Their passion and perseverance in science set up role models for me. And they have generously spared their time for me during my dissertation development. Ana has provided so many valuable suggestions on my projects and career path. It is well deserved that Ana has been appointed as a faculty in Indiana University Bloomington right before I finished up my work. Liliana serves on my committee, and more than a mentor, she is a great friend that always keeps her door open for me. She has spent so much time on my job interviews, dissertation work, final defense and beyond. I cannot forget how she spent hours to talk with me about academic career and how she sacrificed a beautiful Saturday to guide me to revise my defense slides. She is the one that you can always reach to when you need help. If I fortunately get to the point in the future where I will have my own students, I want to be like her.

Since my Master's, I have been under the mentorship of Dr. Justin Bahl. I am so grateful that he not only took efforts to transfer me to Athens to continue these exciting projects but also provided me an office with windows, for which I had longed for years! He meanwhile supplies me great relaxing time with his wife Alex, his twin babies Wolfgang and Vedan, and his two dogs Bernie and Pippa. Other than that, he is an absolutely great mentor to provide me so many learning opportunities and stand for me when I need help. Whenever I had some thoughts, he

would be the great one to talk with. His guidance on the direction of my research is critical for me to not get lost or sway away too far from the main road.

I really appreciate all the dedicated committee members to help build my work. This dissertation could not have been done without the excellent data and guidance from Dr. Pedro Piedra, who took efforts to remotely join every meeting. The work has been much improved by the constructive suggestions from Drs. Pejman Rohani and Mark Tompkins, who both contributed great thoughts with their expertise. Without the growing-me-to-the-best committee, I would have not been able to learn so much and accomplish this work.

Several training grants have provided me valuable opportunities to acquire critical skills to develop my projects. The NIH Centers of Excellence for Influenza Research and Surveillance (CEIRS) Cross-Network Graduate Student Training Award (HHSN272201400008C) in 2016 supported me to be trained in the Krammer Lab at Icahn School of Medicine at Mount Sinai, where I learned the Hemagglutinin HA protein structure and related experiments used in Aim 1 of the dissertation. The R. Palmer Beasley Travel Award in International Research in 2017 has funded me to be trained in the Cowling Lab at the University of Hong Kong to work on influenza household transmission model with Dr. Benjamin Cowling, which established my fundamental skillset of phylogeographic modeling used in Aims 2 and 3. Lastly, grant supported by the NIH National Institute of General Medical Sciences (NIGMS) under award number 5R25GM089694 allowed me to attend the Summer Institute in Statistics and Modeling in Infectious Diseases (SISMID) at the University of Washington. Modules in this training have enriched and broadened my skillset in infectious disease modeling, network epidemiology, and pathogen evolution, selection and immunity.

I would like to finish with two fluffy friends. My cat Violin, she has tolerated twice our long-distance relocations for keeping me accompanied along the years. She kindly allowed me to be absent from home more than 14 hours per day during the last year. She did not complain but slept tight while I worked late with all lights on. She amazed me by fitting in everything (made by water I guess?). And she used all her weight to support my research by sitting/sleeping on my typing wrist all the time. Overall, she proves that cats are more than cuteness but can be seriously sleepy in front of any scientific papers. Another friend is Doug, a handsome and quiet Basset Hound, owned by Joe Hicks. Doug and I became good friends in Athens when Joe and Doug generously let Violin and I stay with them while I had trouble with finding a proper apartment. Even after he was diagnosed with cancer earlier this year, he was still a happy friend and so determined to keep us well-accompanied to finish our dissertation and Joe's relocation to Santa Fe. Unfortunately, since September he is no longer with us. He is a wonderful friend that will always be remembered.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Literature Review

Viral respiratory infections commonly affect the upper or lower respiratory tract, occur in epidemics and spread rapidly across communities in the globe (1,2). The symptoms may vary from mild to severe respiratory diseases with life-threatening outcomes (3–8). Among the etiologies of respiratory infections, influenza and respiratory syncytial virus (RSV) are the leading viral pathogens to cause morbidity and mortality in all ages, with significantly higher impacts on children younger than 5 years and the elderly (9). Every year, influenza leads to respiratory tract infections in 5–15% of the population, severe illness in 3–5 million individuals (2), and seasonal influenza-associated respiratory deaths in 0.3–0.65 million people (10). Another impact of viral respiratory infections on disease burden is that they increase disease susceptibility and severity of secondary viral/bacterial infections or other co-morbidities (3,11), since the primary viral infections might change the immunity and pathology in the host (3,12). For example, the estimated deaths from 1918 H1N1 influenza pandemic are 40-50 million individuals, of which many are attributable to secondary bacterial pneumonia with *Streptococcus pneumoniae* (13).

Respiratory viruses can spread rapidly due to the ease of transmission, usually via contact, fomites, droplets, or aerosols (14). Especially, airborne transmission via droplets and aerosols enables some of these viruses to spread efficiently among humans, causing outbreaks that are difficult to control (14). Studies on influenza viruses have shown that some

environmental and ecological factors can affect the size, stability, and inhalation of both droplets and aerosols (15–18). Such factors include temperature, humidity, and transportation connections (15–18), which may affect the seeding and geographic diffusion of influenza viruses. While some geographic diffusion patterns of influenza viruses have been reported (18–22), further characterizing the global and local transmission dynamics and their associated epidemiological and ecological factors can facilitate effective outbreak prevention measures (20).

Another important reason for the continuous outbreaks and rapid spread of influenza viruses is that viral genetics are highly diverse and host populations usually only have naïve or partial immune protection due to rapid viral genetic and antigenic changes (23,24). The most common mechanisms of genetic and accumulated antigenic changes in RNA viruses are high point mutation rates (usually cause "antigenic drift"), immune selection and migration (25,26). Additionally, frequent reassortments of influenza viruses due to their segmented genome can result in "antigenic shift" leading to novel antigenic strains that can cause influenza pandemics (26–28). Accurate estimation and reconstruction of the evolutionary history of influenza viruses can enhance the understanding of pathogen changes and outbreak dynamics, identify significant prevention measures, and facilitate effective vaccine design (28,29).

The growing importance of statistical phylogenetic methods to study the epidemiology, evolution and ecology of rapidly evolving viral pathogens has been driven by the increasing availability of genetic sequences, development of advanced algorithms and rapid improvement of computing capability (30). In recent decades, viral phylogenetic analysis has become a powerful tool to reveal the evolutionary history of pathogens and understand viral diffusion and disease outbreaks (28,30). In this literature review, I will present an overview of influenza viruses and vaccines, basic phylogenetic modeling tools, and recent studies of viral diffusion and disease

dynamics. It aims to lead to the knowledge gaps and research questions I intend to answer in this dissertation.

Pathogen: Influenza Viruses

Influenza, mostly referred to as "flu", is an acute respiratory disease caused by the influenza viruses belonging to the family of *Orthomyxoviridae* with a negative-sense, single stranded, enveloped, and segmented RNA genome (31,32). Influenza viruses have very diverse antigenic characteristics, where they have been classified into 4 types: A, B, C, and D (31,33). Type D, first isolated from swine in 2011, only circulates in some major livestock like cattle, sheep, and goats (33), which is not known to infect humans. Type C has been observed to only cause sporadic mild upper respiratory symptoms and is not thought to cause epidemics (31). The epidemics and outbreaks of influenza are mainly caused by types A and B (34), attracting the most attention in research and healthcare. Compared to influenza B, which only circulates in humans (35), influenza A viruses can infect a wide range of host species, including avian, humans, and other mammals (36).

The RNA genome of influenza A and B viruses consists of eight discrete gene segments, each coding for at least one protein (31). The surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA), function for host cell entry and release, and contain the antigenic determinants recognized by the adaptive immune system of the host (37). Due to the distinctive serology of HA and NA, influenza A viruses have been subdivided into subtypes with 18 highly variable HAs (H1 to H18) and 11 distinct NAs (N1 to N11) (38). Three polymerase proteins – polymerase basic 1 (PB1), polymerase basic 2 (PB2), and polymerase acidic (PA) – and the nucleoprotein (NP) form the ribonucleoprotein (RNP) complex that is essential for virus

transcription and replication (39). The matrix protein (M1) underlies the virion lipid envelope, and the membrane protein (M2) is inserted into the host-derived lipid envelope and conduct ion channel activity (40). The non-structural proteins NS1 and NS2/nuclear export protein (NEP) have multiple functions like assisting in RNP nuclear export but other specific roles are still under investigation (36,41). Influenza B viruses have no subtypes due to their fixed antigenic characteristics of HA and NA, though their viral structure is very similar to type A. But influenza B viruses have diverged into two antigenically distinguishable lineages – B/Yamagata-like and B/Victoria-like, based on the accumulated antigenic variabilities since 1970 (31,42).

Epidemiology: pandemic, seasonal and emerging influenza viruses

The zoonosis of influenza A viruses has significant public health implications, where infecting both animal reservoirs and humans results in continuous cross-species transmissions (43–45) and frequent reassortments to generate antigenic shifts and potential pandemic strains (46–49). Pandemic influenza is a global-scale epidemic characterized by the rapid spread of an antigenically novel virus through human populations (35), which often causes high morbidity and mortality (43). Four influenza pandemics have occurred in the last century: the 1918 Spanish H1N1, the 1957 Asian H2N2, the 1968 Hong Kong H3N2, and the 2009 pandemic H1N1 (50). Although there have been controversies on the exact origins or animal source of genes for the early pandemic viruses, studies agree on that these pandemics were originated by the successful adaptation of a novel HA subtype to humans from an animal source (50). The abrupt and major change in HA (also called "antigenic shift") results in new HA protein to which most humans have no immunity. It is estimated that about 50 million people were killed during the 1918 Spanish H1N1 pandemic (31). The most recent 2009 H1N1 pandemic, which emerged in early April 2009 and ended in April 2010, has provided large viral samples and sequencing dataset to

explore its emergence. Previous studies via phylogenetic analysis have inferred that genes of the 2009 H1N1 pandemic viruses were derived from H3N2 and/or H1N2 triple reassortant viruses originally isolated from swine in North American in 1998 (51) and H1N1 avian-like swine virus first isolated from swine in Europe in 1979 (50). This indicates that close surveillance, investigation of the emergence and accurate reconstruction of evolutionary history of the pandemic viruses can provide important implications on early preparation and control of future potential pandemics.

After a novel pandemic influenza A virus was introduced into human population the virus continues to circulate with seasonal annual patterns with continuous antigenic evolution. Currently circulating seasonal influenza viruses are influenza A/H1N1, A/H3N2, and two influenza B lineages – B-Yamagata and B-Victoria. They have clear seasonality with epidemic peaks during winter months in temperate regions (October through the next May in the Northern Hemisphere and April to September in the Southern Hemisphere) but have year-round occurrence in tropical regions (52). The global disease burden of seasonal influenza viruses is one of the highest among viral pathogens to cause respiratory diseases (53). Seasonal influenza affects all age groups, but usually causes higher incidence and more severe symptoms in young children and the elderly (9). Based on the systematic analysis of seasonal influenza in 195 countries during 1990 - 2016, the incidence of influenza-associated lower respiratory infections was 5.3 per 1,000 persons in all age groups, 9.1 per 1,000 persons in children younger than 5 years, and 15.8 per 1,000 persons in adults older than 70 years, respectively (9). Seasonal influenza vaccines have been the most effective prevention measure to control disease outbreak, though the effectiveness of vaccines can vary for each season. The estimates of vaccine effectiveness during flu season from 2004 to 2018 range between 10%-60% (54).

The World Health Organization (WHO) also highlights some subtypes of influenza A viruses to keep close surveillance and risk assessment as they have the potential to cause pandemics, for example, highly pathogenic avian influenza (HPAI) H5Nx (including N1, N2, N6, and N8) and H7N9 (55). These two subtypes can infect humans directly from close contacts with avian populations and cause high case fatality in both birds and humans (56,57). Not only are the disease burden and public health risks from H5Nx and H7N9 significant, but also the economic impacts are high due to the tremendous loss of protein products during outbreaks in the poultry population (58). The pandemic potential and severe health risks in both humans and poultry population make these two subtypes important to monitor and study.

HA hypervariability and protein structure

The HA is usually the target to study immune responses and develop effective vaccines, due to its immunodominance. This results in the HA sequences being the most abundant genetic data out of all other influenza virus gene segments. Correspondingly, the epidemiological, immunological and ecological data associated with these infections are also the most abundant for inferring viral dynamics and evolutionary history. Currently, the HA glycoprotein of influenza A viruses has been classified into 18 subtypes based on serological tests (38). Even within one subtype, the interaction between continuous immune selections and rapid mutations allowing the virus to escape from host's immune responses results in highly diverse HAs (59). The hypervariability and continuous evolving of HA protein are the main reason to require annual surveillance and updates for seasonal influenza vaccines (60,61).

The HA protein with a spike-shaped structure, binding the virus to the sialic acid on the host's cell membranes, is the major target of protective antibody response against influenza virus infection, especially against the globular head domain of HA (62). The mature form of the HA

glycoprotein is a homotrimer of three HA monomers. The monomer is composed of signal peptide, globular head domain, stalk domain, transmembrane domain and cytoplasmic domain (62,63). Due to their functions of each domain, evolutionary or immune selection pressure is not uniform across the protein (64,65). Evolutionary models that account for the conservation of functional structures and distinct selection pressure on each domain may be critical to more precisely understand the evolutionary history of HA (66).

Disease Prevention: Influenza Vaccines

Vaccines are amongst the most cost-effective prevention measures for many infectious diseases (67). In the U.S., the development and use of influenza vaccines has started in the 30's after Influenza types A and B were isolated in 1933 and 1936, respectively (68). With the discovery that influenza viruses could grow in embryonic chicken eggs, the first inactivated influenza vaccine was developed in 1938 and administered to the U.S. soldiers during World War II, which ultimately made no difference on clinical outcomes in vaccinated and unvaccinated populations (69). Early influenza vaccines only contained inactivated type A (monovalent) but became a bivalent vaccine in 1942 with both types A and type B (69). These vaccines also resulted in high incidence of side effects due to the crude and impure preparations manufactured by early production methods (70). Another important change in the history of vaccine development is the discovery and introduction of cell culture for influenza virus growth in 1949 to produce more purified vaccine components (71). Eventually, the protective efficacy of these inactivated vaccines was first confirmed in the 1950s through surveillance studies and continuous vaccine development (68,69).

In order to regularly monitor and report the virus-vaccine mismatches, in 1952, WHO initiated the first surveillance system of circulating seasonal influenza strains in several countries worldwide (71). In 1973, based on the improved surveillance system covering the majority of the global regions, WHO begun the routine of issuing annual recommendations for the composition of seasonal influenza vaccine (69). In 1978, the first trivalent vaccine containing two type A strains and one type B strain was developed (71–73). During 1960s – 1980s, most of the vaccine side effects have been reduced by the design of split and subunit vaccines and the engineering of genetic reassortment, where chemically or physically inactivated virions treated with detergent are split vaccines and further purification of the HA and NA proteins of these viruses are subunit vaccines (71,73). The first live attenuated influenza vaccine (LAIV), called "FluMist®", was authorized by the U.S. Food and Drug Administration (FDA) in 2003, which is intranasally administered in healthy population of 5-49 years old (74). In the following years, the updated vaccines were recommended for a broader age range, including infants. In 2012, the first quadrivalent LAIV called "Fluarix®" was approved by FDA, which contains two type A strains (A/H1N1 and A/H3N2) and two type B lineages (B-Yamagata-like and B-Victoria-like) (75). Meanwhile, cell-based influenza vaccine instead of chicken embryo-cultured vaccine has been developed since 2012, which eliminates allergens introduced by egg culture and expands the usage of different cell types to make high productivity of cell-cultured vaccine possible (73). A more complete history of vaccine design, application and future directions has been summarized in these reviews (70,71,73).

Behind these important changes in influenza vaccine development, the core effort is to select vaccine candidates that match the currently circulating strains. These vaccine candidates from natural influenza virus strains are recommended by WHO based on the characterization and

prediction of circulating strains that are likely to dominate in the upcoming epidemic season. Twice a year, the expert panel from the WHO Collaborating Centers and essential laboratories/academies reviews the evidence of global surveillance, laboratory and clinical studies, and evaluates the availability of vaccine strains to make recommendation on the components of influenza vaccine for North and South Hemisphere, respectively (76). The evaluation procedure is mainly based on viral antigenic and genetic characterization, which requires tremendous annual surveillance efforts and laboratory tests. After the selection of vaccine strains, it takes at least 6-8 months to produce sufficient global supplies of influenza vaccine via current vaccine production technologies with egg-based, cell-based or recombination-based approaches (77).

Some critical problems have demonstrated the limits of traditional vaccine design for the rapidly evolving viruses. The influenza vaccines selected from wild strains predominantly elicit specific antibodies against the globular head domain of HA glycoprotein for each subtype or lineage, which is only effective to protect against closely-matched antigenic variants (60). The HA, however, undergoes rapid antigenic drift, allowing the virus to escape neutralizing antibody responses (78) and resulting in imprecise prediction of circulating strains. Even with tremendous efforts of continuous surveillance to update the vaccine strain and evaluate vaccine effectiveness, vaccine mismatch has occurred many times (79). In addition to potential antigenic mismatch from selection procedure and delays in production, egg-adapted mutations accumulated during egg-based vaccine production can further exacerbate this issue, where the vaccine strain cultured in chicken embryo acquires amino acid change in the HA protein, resulting in low vaccine effectiveness (80–83). Studies investigating the impacts of vaccine mismatch have reported broad ranging efficacy (10% to 60%) for these annual vaccines, demonstrating severely low and

unstable immune protection from influenza infection (84). Furthermore, the seasonal vaccines offer little or no protection to emerging zoonotic influenza viruses with pandemic potential. For instance, when the surface glycoproteins, HA and/or NA, are replaced through reassortment, the human population has no pre-existing immune protection and the vaccines in use are not cross-reactive with these new strains (47,50,85). This is the mechanism whereby influenza pandemics or novel zoonotic outbreaks with high case fatality have emerged. To overcome these significant challenges, novel approaches, like computational design, have been employed to rationally and promisingly develop vaccine candidates that can induce broadly (ideally universally) cross-protective and durable immunity for all seasonal even emerging pre-pandemic strains (79,86,87).

Tools: Bayesian Phylogenetic Framework

Phylogenetic analysis is a powerful tool that uses information from genetic data to reconstruct the shared ancestry among organisms and provides insights about organism origins and evolutionary dynamics (88,89). The Bayesian statistical framework has been largely used in phylogenetic inference due to powerful computing capacity and advanced model development (30,90). It has the advantage of simultaneously co-estimating important evolutionary parameters with major basic model components, and the capability of inferring the posterior distribution of reconstructed phylogenetic trees rather than only relying on one or several maximum likelihood tree(s) (30). The major model components in Bayesian framework to reconstruct the time-resolved phylogenies include nucleotide substitution model, molecular clock model and coalescent model (91). The nucleotide substitution models are to describe the process of nucleotide changes based on different probabilistic models and assumptions. For example, the popular Hasegawa, Kishino, and Yano (HKY) substitution model developed in 1985 considers

unequal nucleotide base frequencies and allows for the unequal rates of transitions and transversions (92). The most parameter-rich substitution model is the generalized time reversible (GTR) model developed in 1986, which allows for unequal nucleotide base frequencies and unequal rates of any exchanging pairs of the nucleotide bases (93,94). Substitution model computes the genetic changes between nucleotide sequences and provides the branch length on the phylogenetic tree as numbers of substitutions per site. To calibrate the absolute substitutions with real calendar time information, the molecular clock model was introduced where the evolutionary rate and divergence dating can be inferred (95,96). The simplest and usually unrealistic clock model, called a strict clock, assumes that every branch in a phylogenetic tree evolves at the same evolutionary rate (97). The commonly used molecular clock model in Bayesian framework is the uncorrelated relaxed clock allowing rate heterogeneity across the phylogenetic trees with branch rates drawn from a log-normal distribution (96,98,99). Lastly, the coalescent model is used to incorporate the organism population dynamics that maintain the observed genetic diversity under the estimated evolutionary rates (28,100). This model is based on any two dated genetic samples to trace back to a common ancestor (i.e. a coalescent event) and to infer the effective population size to sustain the divergence of the two samples (101). Different parametric or non-parametric coalescent models have been developed to describe distinctive scenarios of population dynamics. Which to use can be decided based on the prior knowledge or model fitting and selection procedure (102).

The pillar of Bayesian phylogenetic framework to compute posteriors from selected models, priors, and data is the Markov chain Monte Carlo (MCMC) usually constructed by the Metropolis-Hastings algorithm (30,103). This algorithm calculates the ratio of posterior densities from two chain steps to move forward with accepted higher posterior, which avoids the direct

calculation of the posterior density itself, significantly saving the computing cost. Details of applying this algorithm can be found in the review (30).

Viral Phylodynamics

In 2004, the term "viral phylodynamics" was coined. It is defined as "the study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies" (28,100). Due to the high mutation rates of RNA viruses, the ecological and epidemiological processes that shape their phylogenies likely occur at the same time scale as their evolution, that is, the imprints from immunological and epidemiological process can be inferred on the evolution tree (100). Therefore, with the impacts from transmission dynamics and immune selection on viral genetic variation, viral phylogenies can be used to explore the interaction of important epidemiological, immunological, and evolutionary processes, such as spatiotemporal dynamics and epidemic spread among different populations, transmission barriers of host species, avian virus zoonotic transmissions, and antigenic drift (28). Of particular interest, host movement and viral migration are examples of cryptic processes that contribute to the spread and seasonality of influenza epidemics. The application of phylodynamics into the studies of viral geographic diffusion, termed as "phylogeography" (104,105), has been revealing very important geographic spread patterns for avian and seasonal influenza viruses. For example, phylogeographic analysis has been conducted on multiple avian subtypes with pandemic potential following human and mass poultry outbreaks to decipher the viral origins and transmission patterns between different locations, including H9N2 (106,107), HPAI H5N1 (108,109), and HPAI H7N9 (110). Phylodynamic models have also been applied to explore whether the seasonal influenza viruses are spread globally from potential source populations during each season or whether local persistence is the main pattern (19–21,46).

Bayesian phylogenetic models applied in the phylogeographic framework are not limited to only geographic traits but also can be applied to other discrete epidemiological factors, like host species, specific antigenic traits, etc. (109). This allows for the simultaneous reconstruction of the evolution of both the virus population and the discrete trait in the context of each other. Furthermore, the generalized linear model (GLM) has been introduced into this framework to explore the tree diffusion processes under the impacts of epidemiological and ecological factors (111). The GLM diffusion approach has the advantage of efficiently exploring multiple epidemiological and ecological factors that can be either discrete or continuous variables (111). This model framework powerfully unifies the evolutionary and epidemiological processes, which can identify the relevant predictors in the epidemiological process and more precisely predict influenza spatial spread (98). For instance, in the predictive model of the known pandemic expansion of H1N1 during 2009, when using the population migration rates generated with both genetic sequence data and passenger flux data via GLM into the epidemiological susceptible-infected-recovered (SIR) model simulation can provided the best predication of the global spatial spread, when compared to only using genetic data or only passenger flux data (18,98).

Structured Coalescent

Population structure can have strong influence on the shape of viral phylogeny. Besides discrete trait analysis in the source-sink model, another way to take population structure into consideration is to distinguish the mutation and migration events in the coalescent procedure, which is not considered in most of the current coalescent models (112–114). To incorporate the heterogeneity of host populations, structured coalescent model comes into use with the development of sophisticated algorithms and increased computing power (114). With structured coalescent model, the host structure or geographic location of isolation is not used as discrete

traits outside coalescent model but can be defined directly in the coalescent model to estimate the viral movement between these populations with important parameters to be reported. For example, a recent study of Middle Eastern Respiratory Syndrome Coronavirus (115) defines human and camel populations as two separate populations, referred to as “demes”, in the coalescent model. Then the coalescent rate can be estimated separately for each deme, and the movement rates between two demes can also be estimated. Compared to the model considering the hosts as discrete traits, even though the sequence sampling schemes have strong biases, the structured coalescent gives more reasonable inferences on the viral transmissions and dynamics (115). Another study using both simulated data and real-world observations from the 2013-2016 Ebola virus epidemic in West Africa also reports that structured coalescent can give precise estimation of evolutionary rates even when only a small amount of data are available at the early stage of the outbreak, where other tree prior coalescent models always overestimate the evolutionary rates (116).

Viral Evolution and Diffusion Dynamics

Applied phylodynamic modeling and comparative genomic analysis of respiratory viruses can be applied to tackle important unanswered questions including developing novel models to improve the accuracy of inferred viral evolutionary parameters, exploring geographic diffusion of seasonal viruses and identifying significant epidemiological and ecological predictors for the transmission dynamics. These questions can address how the epidemiological and ecological processes can impact viral genetic diversity and thereby enhance the understanding of viral evolvability (117). This information can be very valuable for epidemic prediction, prevention and vaccine selection.

Protein Structure in Evolutionary Models

Understanding how the vaccine targeted protein evolves can help with effective vaccine design. Since different functional domains of viral protein can be under disparate immunologic pressure, they may have different evolutionary characteristics (118,119). As mentioned previously, the globular head domain of HA protein acts as immunodominant region to induce neutralizing antibodies and are under higher immune pressure, compared to the stalk domain. Current seasonal influenza virus vaccines induce humoral immune responses mainly targeting the immunodominant globular head domain, which can have strong specific protection from infecting virus that is similar with vaccine candidates, but have weak cross-reactivity to antigenic drift variants (120,121). Conversely, studies in the ferret experimental model demonstrate that the stalk domain was highly conserved across influenza A subtypes and stalk-specific antibodies could provide cross-reactive protection to a high diversity of influenza A subtypes (120,122). In Chapter 2 of this dissertation a phylogenetic model that utilizes protein structure to inform a partitioning strategy will be presented to study how substitution rates and patterns vary across functional domains and provide valuable insights to understand viral evolution (123). Chapter 3 is a case study to extend this strategy to develop a phylogenetic model for respiratory syncytial virus (RSV).

Some studies have developed phylogenetic models with simple representation of protein structure and stability, for example, a scoring system to represent protein structure (124). Kmiecik et al. has explored protein structure with coarse-grained models for protein structure prediction and molecular dynamic simulations of protein folding (124). This provides the statistical scoring system for sequence-structure compatibility, which can be used to evaluate the probability of fixation of a given mutation and improve the precision of ancestral reconstruction

(125). However, this model does not consider the functional structure of the protein. Furthermore, in a phylogenetic context, structurally informed models are still outperformed by some site-independent models in terms of model fitting (125). The simple representation of protein physical structure has been used so far, but novel models with including information of protein structure to reflect biological realism may be expected to yield a better fit.

The integration of protein functions and structures into evolutionary models has two main challenges: 1) published viral protein structural and functional information may not be available or sufficiently resolved based on current studies; 2) The assumption of nucleotide site independence in the model cannot capture the biological reality that some sites are linked due to shared function (126). The first challenge has been partially resolved for some important surface proteins. For example, the functional domains of influenza HA protein and the fusion (F) protein of RSV have been well defined (127,128). For other proteins of interest, high-throughput experiments termed "deep mutational scanning" (129,130) can quantify the effects of all single mutations on gene function. Then evolutionary model can adequately capture the heterogeneity of selection at different sites, which may improve phylogenetic inference and ancestral sequence reconstruction. This technique has been used to quantify the impacts of codon changes on several proteins or functional domains (129,131–135). The second challenge can be resolved with a partitioning strategy, where the rational is that sites in the same partition have similar evolutionary characteristics while different partitions have distinctive evolutionary characteristics (30,136,137). The characteristics could be substitution rates, base composition, synonymous changes, branch lengths, or even the tree topology (30). Bayesian phylogenetic framework can separately estimate parameter values for these defined partitions under one phylogenetic tree, thus accounting for their heterogeneity in the evolutionary process. For

example, the approximate codon model (usually called SRD06 model) incorporates information of the genetic code, where it partitions the third codon position separately from the first and second codon positions. This model allows for a rate variation between these two partitions, because the changes on the third codon position usually does not result in an amino acid change (synonymous mutations) but changes in the first and second codon positions usually cause an amino acid change (nonsynonymous mutations) (137). This codon portioning strategy thus indirectly captures different selection pressure on codon positions. It provides insights on partitioning a protein based on its structural and functional domains to explore how the evolutionary process acts distinctively on these domains.

Spatiotemporal diffusion of seasonal influenza viruses

Seasonal influenza viruses, causing outbreaks on a worldwide scale, have clear pattern of annual epidemics with global impacts (138,139). Mapping the spatiotemporal diffusion pattern of the pathogen is very important to understand the seasonality and factors that affect outbreaks (111,140), which can lead to effective prevention measures. Previous studies either based on the analysis of epidemiological data alone or combined with genetic data (phylodynamic modeling) have provided valuable inferences on spatiotemporal diffusion dynamics (19–22,46,141). Bayesian phylodynamic modeling has been well developed as a powerful tool to understand viral diffusion dynamics (98,142). A distinct advantage of Bayesian phylodynamic modeling is that models of rapid viral genetic evolution can be unified with epidemiological and ecological predictors to both improve the accuracy of evolutionary reconstruction and reveal statistical associations between the predictors and viral spread rates (98,104).

Recent applications of Bayesian phylodynamic modeling and newly developed analysis techniques have provided a clear depiction of global dynamics and the source populations for the

diffusion of seasonal influenza viruses (18,19,21,46). Rambaut et al. (46) studied how genomic processes related to global influenza dynamics with A/H3N2 and A/H1N1 in temperate regions during 1992-2005. It showed that seasonal influenza A viruses had a complicated interplay between periodic selective sweeps and ressortments (46). They also compared the different dynamics between A/H3N2 and A/H1N1, demonstrating A/H1N1 viruses have weaker antigenic drift. Taken together, this study via a source-sink phylogeographic modeling procedure hypothesized that the persistence reservoir may be located in the tropics and spread to the temperate regions. This can be explained by that the year-round transmission and consequently its constantly larger effective population size of influenza A virus in the tropical population allow more efficiently natural selection for antigenic diversity than in the sink populations which experience major seasonal bottlenecks. But some contradictory conclusions have appeared in later studies. Bedford, et al. (21) developed Bayesian framework and phylogenetic tree trunk reconstruction models to reconstruct the evolutionary history and global dynamics of A/H3N2 collected during 1998 to 2009. They found that China and Southeast Asia played the largest role in seeding the seasonal strains, but they also found that some strains with mutations harbored in the temperate regions can emigrate to more favorable climes and persist for multiple seasons in the global virus population (21). Later, Bahl et al. (19) investigated the seasonal dynamics and migration patterns of influenza A/H3N2 viruses isolated from global urban centers during 2003-2006 with Bayesian phylogeographic analysis. It revealed that tropical regions may not maintain a source for annual A/H3N2 influenza dynamics, but each region may function as a potential source viral population in temperate regions and spread the virus via population migrations. To explore what drove the viral migration, Lemey et al. (18) developed novel phylodynamic models which unified viral genetics and human air transportation data to explore the global transmission

dynamics of A/H3N2, with isolates collected from 2002 to 2007. This study showed that global transmissions of A/H3N2 were driven by human migration via air transportation and the scales of local spread were negatively correlated with geographic distance. They emphasized the central role of mainland China and Southeast Asia to maintain a source for global A/H3N2 diversity.

The most complete study to date on global diffusion patterns of seasonal influenza has been conducted by Bedford et al. with large and long-term datasets (20). Phylodynamic analysis was performed on A/H3N2, A/H1N1, B-Victoria, and B-Yamagata viruses collected during 2000-2012 to explore and compare global circulation patterns of these major influenza virus lineages (20). The analyses showed that genetic variants of A/H3N2 viruses did not persist locally between epidemics and were reseeded from East and Southeast Asia. In contrast, A/H1N1 and type B viruses persisted across several seasons and exhibited complex global dynamics, where East and Southeast Asia played limited roles of disseminating new variants. They concluded that there may be viral, host and ecological factors that complicated the global dynamics (20). In sum, these studies via phylodynamic modeling integrate both genetic and epidemiological information to understand global diffusion dynamics of seasonal influenza viruses, which offers potential for improving epidemiological surveillance through phylodynamic reconstructions.

Many studies have reported that epidemiological and ecological factors may impact the diffusion of seasonal influenza viruses. Global or local studies mainly based on epidemiological data (141,143–145) have reported that temperature, humidity, precipitation, host movement, population size, and air transportation can affect the diffusion and dynamics of seasonal influenza viruses. In a recent study, Dalziel et al. (22) used weekly incidence data of influenza-like illness from 603 cities in the U.S. to explore important spatial variations, which revealed that

population size and age structure, humidity and the peak climatic conditions of urban centers can drive the incidence of influenza infections and its spatiotemporal dynamics. While this study provided evidence that environmental characteristics may alter herd immunity for different levels of urbanized population, viral genetic data was not incorporated into the analysis. Phylogenetic approaches conducted on a more recent and systematic dataset could quantitatively evaluate the effects of environmental factors on viral diffusion patterns.

As summarized above, previous phylodynamic analyses have provided insights to global spatiotemporal diffusion of seasonal influenza viruses. However, local dynamics, for example, the specific source populations of seasonal influenza viruses into the U.S. and the viral diffusion patterns within the U.S., have not been well explored and quantified. By identifying the close geographic region sources from global settings and transmission connections in the U.S., a better understanding of viral dynamics can improve the prediction of circulating strain and further assist vaccine selection. Chapter 4 of this dissertation will explore the within country (i.e. within the U.S.) diffusion patterns of all seasonal influenza subtypes/lineages. It will first identify important external global introductions into the U.S. via a global dataset. It will also compare the sources of influenza viruses for different seasonal influenza subtypes or lineages. The second part of this chapter will focus on understanding the viral dynamics amongst the U.S. regions. It will further explore the important epidemiological and ecological factors of the U.S. regions to potentially explain the dynamics, which may provide valuable information for viral spread prediction and disease prevention. This dissertation fully takes advantage of a large and complete global and U.S. local genetic and epidemiological datasets of all four seasonal influenza subtypes/lineages to answer the important questions listed above.

Vaccine impacts on seasonal influenza viruses

Since the introduction of influenza vaccines into the population, vaccine efficacy and safety are always the main concerns in the healthcare and research communities (146,147). But evidence has shown that vaccines may shape the dynamics of the pathogen populations (148,149). One explanation is that the host immune landscape, modulated by vaccination, can put strong selection pressure on the pathogen (150). Furthermore, imperfect vaccine design or vaccination procedure may also drive the evolution and genetic diversity of pathogens (151). Questions regarding the impact of influenza vaccine on viral diversity and diffusion dynamics within and between communities have not been adequately studied due to limited information regarding vaccination status of infected individuals. Chapter 5 will utilize samples collected through vaccine trials in central Texas to evaluate the impacts of vaccination on viral genetic diversity during an epidemic season. This study will address the following specific questions: 1) Does the vaccinated population have a higher genetic diversity? 2) Does the vaccinated population have a lower transmission rates to the unvaccinated population and to the global? Answers from these questions can be very valuable to understand local viral source and transmissions, and quantitatively evaluate the impacts of LAIV on viral diversity and diffusion dynamics.

Purpose of the Study

This dissertation will develop and apply phylogenetic models in the Bayesian framework to enhance the understanding of viral evolution and spatiotemporal dynamics. The targeted pathogens are primarily multiple influenza A virus subtypes and two influenza B lineages, which continue to pose global public health threats, including several pandemics in history, heavy disease burdens in humans with annual seasonal outbreaks, and high economic loss with highly pathogenic viral strains causing large outbreaks in avian hosts. Understanding viral evolution and diffusion dynamics by integrating large genetic and epidemiological data is critical for disease prevention and control. In this dissertation I aim to examine the evolutionary and epidemiological dynamics across ecological and biological scales in order to provide insights that could improve disease control. To achieve this, specific studies are conducted to provide insights on: whether integrating viral protein structure can improve phylogenetic inference and better understand the evolution of vaccine targets; how the seasonal influenza epidemic spread among U.S. regions; and whether increased vaccination rate within a community can impact viral diversity and diffusion dynamics.

Specific Aims

Aim 1. To develop structurally informed evolutionary models of HA protein for emerging, seasonal and pandemic influenza viruses. Novel phylogenetic model to incorporate protein functional structure is developed and compared with other phylogenetic models via model selection procedure to evaluate the model fitting. Evolutionary rates and approximate selection pressure from the new model are estimated for each functional domain of HA protein to provide insights on how the vaccine targets evolve.

Sub Aim 1. A case study extends the structurally informed partitioning scheme to improve phylogenetic inference of respiratory syncytial virus (RSV) and assess the generalizability of the application of this new model to other viral respiratory pathogens. It focuses on a specific protein partitioning scheme for RSV, where it considers the different conformations of Fusion protein (F protein) structure before and after the fusion process. The structurally informed evolutionary models are conducted with F protein of both RSVA and RSVB to explore the evolutionary history and approximate selection pressure of each functional domain.

Aim 2. To systematically infer global introductions of seasonal influenza viruses into the U.S. and estimate viral diffusion amongst U.S. regions. The main global regions that introduce seasonal influenza viruses into the U.S. are inferred via phylogeographic models. Furthermore, viral diffusion patterns among the Health and Human Services (HHS) regions in the U.S. are explored with further examining what are the main drives for viral diffusion, where potential ecological and epidemiological predictors are constructed in the phylodynamic model via generalized linear model. The viral diffusion patterns and significant predictors are compared across different seasonal influenza subtypes or lineages.

Aim 3. To evaluate the impacts of influenza vaccine on genetic diversity and diffusion dynamics of H3N2 during 2004-2006 in Central Texas, USA. The viral sources of Texas local cities are inferred via phylogeographic models to first understand whether the vaccinated Texas cities and unvaccinated Texas cities have different viral sources and diffusion patterns. The hypothesis is that vaccinated population needs more external introductions compared to

unvaccinated population. Then the impacts of influenza vaccine on viral genetic diversity and diffusion dynamics are quantified via structured coalescent model, where the tested hypothesis is that higher migration rate from unvaccinated population to vaccinated population can be inferred and larger coalescent time can be observed in the vaccinated population to indicate a higher viral diversity.

CHAPTER 2

STRUCTURALLY INFORMED EVOLUTIONARY MODELS IMPROVE PHYLOGENETIC RECONSTRUCTION FOR EMERGING, SEASONAL, AND PANDEMIC INFLUENZA VIRUSES [1]

[1] Qiu, X. and Bahl, J. To be submitted to *Molecular Biology and Evolution*, 11/2019

Abstract

Precise estimation of genetic substitution patterns is critical for accurate reconstruction of pathogen phylogenies. Few studies of viral evolution account for variations of mutation rate across a single gene. This is especially true when considering evolution of segmented viruses where individual segments are short, encoding for few proteins. However, the structural and functional partitions of these proteins could provide valuable information for more accurate inference of viral evolution, due to the disparate immune selection pressure on different functional domains. Accurately reconstructed evolutionary features on specific functional domains can in turn provide biological information on viral protein and immune targets for vaccine design. In this study I developed and evaluated a structurally informed partitioning scheme that accounts for rate variation among immunogenic head and stalk domains of the surface protein hemagglutinin (HA) of influenza viruses. I evaluated the model fit and performance of four different models - HKY, SRD06 codon, HKY with a structurally informed partitioning scheme, SRD06 with a structurally informed partitioning scheme on pandemic A/H1N1pdm09, seasonal A/H1N1postpdm, A/H3N2, B-Yamagata-like and Victoria-like lineages, and two highly pathogenic avian influenza A viruses H5Nx and H7N9. Results showed that structurally informed partitioning with SRD06 performed better for all datasets with decisively statistical support. Significantly faster nucleotide substitution rates for head domain, compared to stalk domain was observed and may provide insight for stalk derived broadly reactive vaccine design. Taken together, integrating a functionally informed partitioning scheme based on protein structures of immune targets allows for significant improvement of phylogenetic analysis and providing important biological insights.

Introduction

The importance of statistical phylogenetic methods to study the epidemiology, evolution and ecology of rapidly evolving viral pathogens has been driven by the growing availability of whole genome sequences (152). Precisely estimating the pattern of genetic variations is critical to reconstruct the accurate pathogen phylogenies (136,153). Nucleotide substitution models have been developed to describe the process of change from one nucleic state to another among viral isolates, where they often allow for rate variations between transitions and transversions (92), or incorporate nucleotide base frequencies with substitution rate parameters (154,155). Segmented viruses, such as influenza A virus (IAV), contain short gene segments that encode for few proteins (36). Consequently, the literature regarding model development of phylogeny reconstruction is dominated by the analysis on complete protein coding regions. For example, the complete viral surface glycoprotein hemagglutinin (HA) is used the most when studying influenza viral evolution and diffusion (98). However, assuming natural selection acts uniformly across the protein domains may not be justified (156). Some partitioning strategies have been developed to account for variations inside one gene segment (136,137). For example, approximate codon-models, such as the SRD06, incorporate information about the genetic code by allowing for a rate variation by defining substitution models for codon positions 1 and 2 and an independent model for codon position 3. This model accounts for the rapid accumulation of synonymous mutations in the third codon position (137).

HA glycoprotein with a spike-shaped structure binds to the receptors on the targeted host cell membrane when the virus begins the infection (62). This protein has two main domains – the globular head domain and stalk domain. Functionally different, the globular head domain contains the receptor binding sites, while the stalk domain is a main structure responsible for

membrane fusion machinery (62,63). Even though immune selection strongly drives antigenic drift, allowing for accumulation of mutations in the head domain, the stalk domain is functionally conserved across viral subtypes. Current seasonal influenza virus vaccines induce humoral immune responses primarily targeting the immunodominant globular head domain (120,121), which can provide strong protection from an infecting virus that is similar with vaccine candidates, but have weak cross-reactivity to antigenic drift variants. Conversely, studies in the ferret experimental model demonstrate that the stalk domain is highly conserved across IAV subtypes and stalk specific antibodies could provide cross-reactive protection to a high diversity of IAV subtypes (120,122). Despite the importance of understanding how substitution rates and patterns vary across functionally conserved domains in universal vaccine design (123,157), few studies have incorporated protein structure in phylogenetic models.

In this study I proposed a novel phylogenetic model that incorporates a protein structure informed partitioning scheme to account for variable evolutionary rates resulting from immune selection. I aimed to evaluate the appropriateness of the novel model to reliably to estimate biologically informative parameters from available genetic data. I evaluated and compared four different models: HKY, SRD06 codon, HKY with a structurally informed partitioning scheme, SRD06 with a structurally informed partitioning scheme. I also evaluated the model performance across multiple viral subtypes, including pandemic H1N1pdm09, seasonal A/H1N1postpdm, A/H3N2, the two seasonal Influenza B virus lineages – Yamagata-like and Victoria-like, and two highly pathogenic avian Influenza (HPAI) H5Nx and H7N9 viruses. I further conducted sensitivity analysis to determine whether the new model was sensitive to sample size, data distribution and epidemic stage resulting in biased estimation. Separate estimation of viral evolutionary rates for the head and stalk domain may be informative for vaccine design. I

therefore formally tested the hypothesis that evolutionary rates in each domain were significantly different and further estimated the relative selection pressure for each functional domain of HA protein.

New Approaches

This study proposed a new model to incorporate a structurally informed partitioning scheme on a single protein into phylogenetic reconstruction. Decisive statistical support from the model selection procedure and stably supported across multiple influenza subtypes validated the superiority of the new model. The model can provide biological insights for viral evolution, with reporting domain-specific evolution rates and approximate selection pressure. The tree branches with domain-specific rate ratio $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ can inform the approximate selection pressure on each strain, suggesting some biological explanations related to antigenic drift and emerging strains. The structurally informed phylogenetic model may reveal novel biological insights of viral evolution and have the potential to reveal more biological realism without over-parameterization.

Methods

Datasets

Multiple datasets representing pandemic, seasonal and emerging influenza viruses were used to develop and broadly test the proposed structurally informed partitioning models, including pandemic H1N1pdm09, seasonal influenza viruses A/H1N1postpdm, A/H3N2, B/Yamagata-like and B/Victoria-like, and two HPAI H5Nx and H7N9. Except for HPAI H7N9, H1N1pdm09 and seasonal A/H1N1postpdm that have been directly downloaded from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/), all other datasets

were from published research, including seasonal A/H3N2 (20), HPAI H5Nx (158), B/Yamagata-like and B/Victoria-like (159). To avoid improper weights of duplicate isolates on the partitioning models, sequences with 100% similarity were removed for all datasets (final datasets can be found in the corresponding xml files via the GitHub link: https://github.com/XuetingQiu/FluPartitioningModels).

Briefly, the final datasets contained 505 global isolates for H1N1pdm09 during 2009/04 – 2009/12, 635 global isolates for HPAI H7N9 during 2013-2018, and 554 global isolates for seasonal A/H1N1postpdm during 2012-2018. HPAI H5Nx contained 1,095 isolates during 1996-2016, after removing the vaccine candidates (158). The two influenza B virus lineages were full genome datasets sampled during 2002-2013 in eastern Australia and New Zealand (159). For B-Victoria, the dataset finally retained 214 isolates after randomly sampling down to 50% of the original dataset and removing the duplicates. For B-Yamagata, 241 isolates were included after removing the duplicates. The global dataset of seasonal A/H3N2 (20) contained 906 isolates during 2000-2012 after the duplicates were removed and further randomly subsampled.

Definition of HA protein structural partitions

HA glycoprotein has been defined as different functional domains, including signal peptide, stalk domain, globular head domain, transmembrane domain and cytoplasmic domain (62,63). These protein domains labeled with sites of amino acid were translated into the linear sites on the nucleotide sequences of HA gene (Fig. 2.1a and 2.1b). To use the full length of HA sequences and avoid some partitions being too short and lack of information, signal peptide, transmembrane domain, and cytoplasmic domain were combined as one partition named STC, given the similarly evolutionary characteristics of these functional domains. Though the stalk

domain on the linear nucleotide sequence was separated by the head domain, the two gene regions of stalk domain were merged together for analysis (supplemental table I-1 and xml files in GitHub https://github.com/XuetingQiu/FluPartitioningModels).

Structurally informed phylogenetic models

As shown in Fig. 2.1b, four different Bayesian phylogenetic models were conducted using the same molecular clock model and coalescent model for each dataset. Only the partitioning strategies on the nucleotide sequences were different. The base model applied HKY substitution model (92) with Gamma invariant distribution, where this model accounts for base frequencies and allows for rate variations between transitions and transversions. SRD06 codon model (C model in Fig. 2.1c) (137) partitioned all codons of the HA gene reading frame into codon positions 1+2 and codon position 3, where these two partitions also used HKY substitution model. Protein structure partitioning model (P model in Fig. 2.1c) contained STC, stalk and head functional domains with HKY substitution model for each domain. The last model (CP model in Fig. 2.1c) combined the partitioning strategy in both C and P models, that is, codon positions in each protein functional domain were further partitioned into codon positions 1 + 2 and codon position 3.

Tree likelihood construction in the partitioning models

After partitioning with codon positions or/and protein structural domains, the overall tree likelihood calculations need to be explained. The general Bayesian inferences are based on the posterior probability of a hypothesis, for example, a given tree τ. Then the posterior probability of the tree based on the data can be obtained using Bayes theorem,

$$Pr(\tau|x) = \frac{Pr(x|\tau)Pr(\tau)}{\sum Pr(x|\tau)Pr(\tau)} \quad [1]$$

In which, *Pr(τ)* is the prior probability of the tree hypothesis τ based on data x; *Pr(x|τ)* is the probability of observing the data at a specific site.

Then assuming independence of substitution across sites, the general likelihood based on a specific phylogenetic model, i.e. the probability of observing the aligned matrix X of all j sequences is,

$$f(X|\tau_j, v_j, \theta, \phi) = \prod_{i=1}^{c} f(x_i|\tau_j, v_j, \theta, \phi) \quad [2]$$

In which, c is the number of total sites along the sequences; τ represents the tree hypothesis, v represents the branch lengths, θ is the substitution model parameters, and ϕ represents all other model parameters.

The proposed structurally informed partitioning model based on HA protein functional domains extends the formula [2] of the tree likelihood estimation. Each of the defined partition has its tree likelihood estimated independently, therefore, the overall tree likelihood is the products of each partition-specific likelihood, that is,

$$f(X|\tau_j, v_j, \theta, \phi)$$

$$= \prod_{i=1}^{c_1} f(x_i|\tau_j, v_j, \theta, \phi)$$

$$* \prod_{i=1}^{c_2} f(x_i|\tau_j, v_j, \theta, \phi) * \ldots * \prod_{i=1}^{c_n} f(x_i|\tau_j, v_j, \theta, \phi)$$

In which, the sum of c_1, c_2, …, c_n is equal to c, the total number of nucleotide sites along the sequences; and n is the number of partitions in the model.

Phylogenetic model simulations

All models were simulated in the package of Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.8.4 (160). To allow for rate variations across lineages, the uncorrelated lognormal relaxed molecular clock was used with an initial mean of 0.0033 with a uniform prior ranging from 0.0 to 1.0. Based on prior knowledge, a smooth and time-aware Gaussian Markov random field (GMRF) process prior on the population sizes was applied in the Skyride calescent model (161). Each model was performed at least four independent runs of 100 million Markov Chain Monte Carlo (MCMC) generations. To report the substitution rates and phylogenetic trees, four runs for each model were combined after removal of burn-in to achieve an Effective Sample Size (ESS) of >200 as diagnosed in Tracer v 1.5. Visualized violin plots for substitution rates were generated via personal R scripts (GitHub https://github.com/XuetingQiu/FluPartitioningModels).

Model selection procedure and criteria

The goal of model selection procedure is to compare model superiority and identify a model that is sufficiently complex to capture the biological realism and evolutionary processes that have occurred but to avoid overparameterized models with more parameters than these can be reliably estimated from the available data (102,136). In this study, both path-sampling (PS) (162) and stepping-stone sampling (SS) (163) were used to compute the marginal likelihood estimation to perform the model selection procedure. PS and SS approaches account for both the number of parameters and the appropriateness of prior distributions for these parameters.

The marginal likelihood is the probability of the data (that is, likelihood) given the model type, not assuming any particular model parameters (102). The marginal likelihood estimation

for both PS and SS with Beta path step distributions is set to run 100 path steps and length of chains as 1 million. Bayes Factors (BF), the ratio of marginal likelihood estimations from two models, are used to evaluate the significance of the model comparison. When converted to a log scale, BF becomes the difference between two log marginal likelihood estimations. The mathematical conversion is below (164):

For a model with parameters Θ, the marginal likelihood for the model M with data X is,

$$p\,(X \mid M) = \int p\,(X|\Theta, \mathrm{M})\, p(\Theta|\mathrm{M}) d\Theta \quad [3]$$

Then the Bayes factor can be calculated for model M1 against model M2,

$$BF_{12} = \frac{p(X|M_1)}{p(X|M_2)} = \frac{\int p\,(X|\Theta_1, M_1)\, p(\Theta_1|M_1) d\Theta_1}{\int p\,(X|\Theta_2, M_2)\, p(\Theta_2|M_2) d\Theta_2} \quad [4]$$

The log scale of BF is derived as below,

$$logBF_{12} = log\ p(X|M_1) - log\ p(X|M_2) = dB \quad [5]$$

With log scale BF as dB, the criteria for model selection are: dB <0 means the selection supports M2; 0 ≤ dB <5 means no evidence for supporting M1; 5 ≤ dB < 10 means substantial strength of evidence for supporting M1; 10 ≤ dB <15 means strong evidence of supporting M1; 15 ≤ dB < 20 means very strong evidence of supporting M1; dB ≥ 20 means decisive evidence of supporting M1.

Model performance, validation and sensitivity analysis

To compare model performance among four models for each dataset, two basic evolutionary parameters were estimated and reported: the root heights and overall substitution rates. If the parameters for each dataset were similar across these four models, then it indicated that the model can provide accurate and stable estimation no matter how many partitions it has included.

To evaluate whether the structurally informed partitioning model was sensitive to sample size, sensitivity analysis was performed on HPAI H5Nx from 1996-2017, the largest final dataset with 1,095 isolates in this study. With retaining the oldest isolate of H5Nx, the rest was randomly subsampled to keep 80%, 60%, 40% and 20% of the dataset, respectively. Four runs of 100 million MCMC chain length of the four models for the full dataset and these subsampled datasets were conducted with marginal likelihood estimations for model selection. Tree height parameters, domain-specific substitution rates, and model selection BFs were reported to evaluate the data sufficiency for model stability.

To further conduct model validation, two more approaches were applied. Firstly, to evaluate whether the structurally informed partitioning model was sensitive to data distribution, the full seasonal A/H3N2 dataset was randomly and repeated sampled 40% of the full dataset into two subsets. These two subsets had different geographical and temporal distributions of A/H3N2 isolates. Secondly, to assess whether the new model was sensitive to the epidemic stage. With the H1N1pdm09 dataset, the Early stage of the pandemic was defined as April to middle of July, 2009, and the Later stage of the pandemic was July-December, 2009. The same model variations and model selection procedure were conducted to all these subsets.

Domain-specific dC$_{1+2}$/dC$_3$ recorded to tree branches

The new model can report domain-specific rates and dC$_{1+2}$/dC$_3$ values. To visualize these specific rates or ratios on the tree branches, the BEAST xml files (GitHub link: https://github.com/XuetingQiu/FluPartitioningModels) were coded to compute and load the specific branch rates for stalk domain codon positions 1+2, stalk domain codon position 3, head domain codon positions 1+2, and head domain codon position 3. Python script (GitHub link:

https://github.com/XuetingQiu/FluPartitioningModels) was constructed to compute and write the dC$_{1+2}$/dC$_3$ for stalk and head domain from the median rate of each partition to the final summarized tree branches. Seasonal A/H3N2 and all Egypt H5Nx data were used as examples to show the domain-specific dC$_{1+2}$/dC$_3$ changes in A/H3N2 with different vaccine isolates introduced and the changes in different hosts (human v.s. avian) in Egypt H5Nx.

Results

Model selection

To compare the four models, the log scale Bayes Factor (BF) was used, which was the difference of log marginal likelihood estimations from two models. Marginal likelihood estimation was conducted with two approaches – path-sampling and stepping-stone sampling, where both generated consistent results. Based on the model selection procedure for all subtypes (Fig. 2.1), the model with mixed partitioning on both codon positions and HA protein structure (CP model in Fig. 2.1c) had decisive BF support for better model fitting in all datasets compared to HKY and HKY with protein structure partitioning model (HKY model and P model in Fig. 2.1c). Compared to SRD06 codon model (C model in Fig. 2.1c), CP model performed significantly better for all datasets, with different levels of BF supports. HKY with protein structure model (P model) performed significantly better than HKY model for human influenza A and B viruses, but it was not significant in highly pathogenic avian influenza H7N9 and H5Nx datasets. This may indicate that these emerging viruses with avian-dominated host populations have a different immune selection pressure pattern compared to human influenza viruses.

Tree root height and substitution rates

To evaluate the accuracy and stability of model performance, tree parameters and estimated substitution rates were used to compare different models. Results from four models showed very similar tree root height and 95% Bayesian Credible Interval (BCI) for each influenza subtype (Table 2.1). The estimated mean substitution rates (Fig. 2.3) estimated from all models were similar for each dataset suggesting that the CP model generated reliable estimations for important evolutionary parameters that were comparable to other models. Furthermore, since both CP and P model incorporated partitioning scheme on protein structure, these two models provided domain-specific substitution rates for STC (including signal peptide, transmembrane domain and cytoplasmic tail), stalk and head domains, respectively. However, the domain-specific rates reported by P model were underestimated, compared to CP model (results shown in Supplemental Table I-2). Such an observation was unsurprising since the HKY substitution model on each protein structure partition cannot account for more realistic variations among codon positions, same as Shapiro et al. reported in (137). One thing to note, though STC domain was not partitioned on codon positions, its small amount of information from short nucleotide length (about 150 nucleotides) resulted in high uncertainty of the STC substitution rates, showing long tails in the violin plots.

Comparing the overall substitution rates for each influenza subtype, I observed that the pandemic H1N1pdm09 had the highest substitution rate (~7.5×10^{-3} substitution/site/year). HPAI H7N9 and H5Nx had the substitution rates around $4.0 – 4.8 \times 10^{-3}$ substitution/site/year. Seasonal influenza viruses had the range as $2.0 – 3.5 \times 10^{-3}$ substitution/site/year, where the two type B lineages had lower rates than the two A subtypes.

Furthermore, I compared the separate substitution rates for stalk and head domains from the CP model. The absolute rates calculated directly from each step of the Bayesian Markov Chain Monte Carlo (MCMC) simulation were plotted into violin plot (Fig. 2.3). Results showed that the head domain had a significantly higher substitution rate than stalk domain for all influenza subtypes, under the unit of substitutions per site per year. The Bayes Factors (BF) for the statistical tests between the substitution rates of head and stalk domains were 1137 (H1N1pdm09), 6 (H1Npostpdm), 577 (A/H3N2), 197 (B-Victoria), 191 (B-Yamagata), 303 (H7N9), +infinity (H5Nx), respectively, with the significant criteria as BF>3.

Approximate selection pressure

With partitioning on codon positions, the CP model can estimate the ratio of substitution rate of codon positions 1 and 2 (majority of non-synonymous changes occur in these codons) over the rate of codon position 3 (majority changes are synonymous), that is, $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, for stalk domain and head domain separately. The relative rate of C_3 to C_{1+2} is a good predictor of the ω parameter of selection pressure measure (137) and thus could reflect the domain specific selection pressure. As shown in Table 2.2, the range of $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ was 0.10 – 0.32 for stalk domain and 0.25 – 0.87 for head domain from all subtypes of influenza viruses included in the study. The overall $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ for the whole HA protein calculated from the c model was lower than head domain but higher than stalk domain for each influenza subtype. I also observed that the stalk domain had a significantly lower $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ than head domain for each influenza subtype (Bayes Factors shown in Table 2.2), which indicated that stalk domain experienced stronger purifying selection to maintain its conservation. Further to compare pandemic H1N1pdm09 and seasonal A/H1N1 post pandemic, I found that the seasonal H1N1 post

pandemic had lower dC$_{1+2}$/dC$_3$ in both stalk and head domains, indicating the approximate selection pressure changed as disease dynamics changed from pandemic to seasonal epidemic patterns.

To observe different patterns of selection pressure on the viral populations, phylogenetic tree branch-specific stalk dC$_{1+2}$/dC$_3$ and head dC$_{1+2}$/dC$_3$ were calculated and visualized via mapping the upper and lower quantiles of the ratio values onto the phylogenetic tree. I used seasonal H3N2 (Supplemental Fig. I-1) and Egypt H5Nx (Supplemental Fig. I-2) as examples. The dC$_{1+2}$/dC$_3$ had a range of 0.63 – 0.75 for H3N2 head domain, 0.21 – 0.23 for H3N2 stalk domain, 0.12 – 1.87 for Egypt H5Nx head domain and 0.09 – 0.32 for Egypt H5Nx stalk domain. Again, results from the branch-specific dC$_{1+2}$/dC$_3$ demonstrated that the head domain had significantly higher dC$_{1+2}$/dC$_3$ compared to stalk domain, which indicated that stalk domain was under stronger purifying selection pressure to maintain its conservation. To note, the majority of the dC$_{1+2}$/dC$_3$ values were less than 1, which indicated purifying selection, but occasionally some diversifying selection pressure (dC$_{1+2}$/dC$_3$ >1) was observed in Egypt H5Nx head domain in chicken-dominated population. This high dC$_{1+2}$/dC$_3$ in H5Nx chicken population may indicate that the viruses circulating in chicken may experience less purifying selection pressure to generate higher diversity with keeping more non-synonymous changes compared to that in human population. This suggests that viruses circulating in chickens potentially have greater evolvability and generate more variants, which may have pandemic potential.

Sensitivity analysis and model validation

With the largest dataset HPAI H5Nx in this study, sensitivity analysis was performed to estimate the stability of model performance under different sample sizes. The same model fitting

and selection procedure was tested on randomly selected 80%, 60%, 40% and 20% of the H5Nx-full dataset. The model selection procedure (Fig. 2.4) showed that all these datasets had decisive BFs to favor the structurally informed partitioning model with SRD06 (CP model). Tree root height estimation (Table 2.3) from these datasets had similar estimates and 95% BCI, where it was only slightly overestimated in the 20% of H5Nx dataset. The substitution rates of protein structure partitions (Fig. 2.5) also showed stable estimations across different sample sizes but only had slight overestimation with the 20% H5Nx dataset. Taken together, though the new model with partitioning on both codon positions and protein functional domains introduced more parameters, it performed stably with different amount of data and thus was not sensitive to sample size.

With two subsets of random selection from seasonal H3N2, model selection procedure and evolutionary parameter estimations generated very similar results. It indicated that the new model validated through different subsets and was not sensitive to the different distributions of epidemiological factors, including geographic location and isolation time (shown in Supplemental Table I-3).

With dividing the H1N1pdm09 dataset into the Early stage and Later stage of an epidemic outbreak, model selection procedure reported that the CP model was significantly better than all other models in both datasets. The overall and domain-specific substitution rates in the Early stage was slightly higher compared to the Later stage of the outbreak (shown in Supplemental Table I-4). And the Early stage had a higher $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, indicating it experienced weaker purifying selection pressure to be relatively diverse. This may be related to host adaptation in the Early stage, where it generated more variants for the advantage of widely spreading in the host population.

Discussion

In this study a novel phylogenetic model with integrating both codon positions and protein structure was developed and tested with HA protein of influenza viruses. Results showed that the new model was statistically supported to fit the data better and provided novel biological information that can reveal new insights into viral evolution. The substitution rates of stalk and head domains of HA protein were significantly different, with a higher rate in head domains for all the tested influenza types. Approximate selection pressure constructed from codon position partitions showed that stalk domain experienced higher purifying selection pressure to maintain its conservation. Model validation from subsets of multiple influenza subtype reported the model performed stably and was not sensitive to sample size, data distribution and epidemic stage. Taken together, this new model could provide potential biological explanations related to antigenic drift, vaccine escape and pandemic emergence to understand viral evolution and risk assessment.

Partitioning scheme on genetic data allows for capturing similar evolutionary features inside one partition of data and comparing different partitions to generate new insights on viral evolution. For example, SRD06 codon model categories the first and second codon positions together since they behave similarly to mainly generate non-synonymous changes, while the third codon position is treated as a separate partition due to mostly generating synonymous changes (136,137). Some models conduct partitioning scheme on multiple proteins to reconstruct the co-evolution and estimate protein-specific rates under one model (165). These models function well for some purposes of observing the co-evolution of multiple proteins, however, it assumes evolutionary processes act uniformly across one protein, which may not be the biological reality due to distinctively functional domains on one protein.

When developing the new model in this study, the primary tasks are to efficiently capture the information regarding HA protein structural and functional domains and to properly incorporate this information into the evolutionary models. Therefore, how to partition on the protein structure with genetic data is critical. One solution is to compare all possible partitioning schemes for a given genetic sequence dataset to find the best partitioning strategy (136). However, this approach is computationally inflexible because the combination of possible partitioning schemes is tremendous even for a very small number of data blocks (166). Consequently, many previous studies ended up either arbitrarily choosing a single partitioning scheme or just selecting the best scheme based on statistical model selection (136,167,168). With these limitations on proper partitioning, the accuracy of the inferences from partitioned phylogenetic analysis remains uncertain, even though advanced developments in phylogenetic analysis have been achieved in recent decades (136). Due to advanced progress in viral protein structure (127), partitioning scheme in this study is based on biological functions of different protein domains, where HA protein has been defined to distinguish the immunogenic roles and other molecular-level functions during viral infection. Molecular mechanism of influenza viruses, for example, the initiation of viral fusion or antigenic-immune stimulation, is determined by the head domain of the protein. Nucleic acid changes observed in this domain usually result from a balance between immune selection and conserved functionality (169–171). The new partitioning model considers the distinctive functions of head domain to mainly bind to host-cell receptors and the stalk domain to perform membrane fusion.

Another advance in this study is that a formal model selection procedure with reliable approaches (172) to quantitatively evaluate the model performance and accuracy of inferences for pandemic, all seasonal and two emerging influenza viruses. Previously the harmonic mean

estimate (HME) and Akaike's information criterion through MCMC (AICM) (173) have been commonly used. In recent years, several new approaches to perform model selection in the field of phylogenetics, such as path- sampling (PS) (162) and stepping-stone sampling (SS) (163). Baele et al. (102) have tested HME, AICM, PS, and SS approaches for demographic and molecular clock model comparison, which "confirmed that HME systematically overestimates the marginal likelihood and fails to yield reliable model classification, and PS and SS substantially outperform HME estimators" (174). Therefore, in this study both PS and SS approaches have been used to perform the model selection procedure to generate reliable results of model comparison.

The new model with both partitioning on codon positions and protein structure incorporates biological functional domains inside one protein to allow for a quantitative exportation of the evolutionary process as a function of approximate host immune selection. The model did capture the different evolutionary process of these functional domains, which showed that the stalk domain had a significantly lower substitution rates and lower $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, indicating that the stalk domain experienced stronger purifying selection and maintained its conserved functionality. The estimated domain-specific $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ can reveal insights of selection pressure on different viral isolates in different host populations. For example, diversifying selection pressure on the head domain has been observed only in chicken population in Egypt. One possible explanation could be that the failed efficacy of poultry influenza H5 vaccines implemented in Egypt (175,176) may drive the genetic diversity of the viruses in the chicken population.

Even though $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ estimated by the model is not an exact measurement of the ω parameter, it is a good predictor of the ratio of non-synonymous changes over synonymous

changes (137). The new model can provide an informative approximation and avoid heavy computational demands of a full codon or amino acid models (137). When assuming synonymous changes do not affect viral fitness, the dC_{1+2}/dC_3 value can provide corresponding estimation for selection pressure compared to mutation-selection models which assess the strength of natural selection on specific mutations (177). Furthermore, the partitioning scheme and additional parameter estimation did not significantly increase the analysis run time compared to other models in this study under the same computing environment.

The structurally informed evolutionary model may provide insight for universal vaccine design. The higher rates observed in the head domain are probably due to strong immune selection targeted on the globular head, where some sites are mostly targeted by the neutralizing immunity induced by seasonal influenza vaccine (120). Supported by many laboratory studies, stalk domain could be used to induce more broad-reactive vaccines (120,122,123). This model can provide important insights into the evolvability and potential durability of stalk-based vaccines. Accurate reconstruction of ancestral stalk domain sequence under a shared viral phylogeny has potential to induce broad protection against influenza A virus Group 1 (H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, and H17) and Group 2 (H3, H4, H7, H10, H14 and H15) (170), respectively, which might be a promising approach to design a universal influenza vaccine.

While this model has been developed with influenza in mind, there is potential to apply this approach to other viruses. Proteins, such as the respiratory syncytial virus F and G protein or the coronavirus spike protein that have multiple functional domains including immune stimulators, may be well suited to the novel approach that incorporates protein structural and functional domains into the nucleotide substitution models. Incorporating this type of

information into comparative genomic analyses has the potential to provide important biological information and improve both vaccine design and evaluation.

Tables and Figures

Table 2.1. Tree root height estimation and its 95% Bayesian credible interval (BCI) from each model for each influenza subtype. Results from four models show very similar tree root height and 95% BCI for each influenza subtype. The unit of the root height is in years.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

Models	Pandemic			Seasonal													Emerging					
	H1N1pdm09			H1N1postpdm			A/H3N2			B-Victoria			B-Yamagata			H7N9			H5Nx-full		
	Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI	
HKY	1.12	0.99	1.29	8.01	7.70	8.34	16.06	14.46	18.86	11.44	11.10	11.84	13.81	12.88	14.86	6.31	5.34	7.89	21.25	21.08	21.48
c	1.12	0.99	1.29	8.01	7.68	8.34	16.06	14.45	18.67	11.46	11.13	11.87	13.79	12.87	14.86	6.30	5.37	7.90	21.28	21.08	21.48
p	1.12	0.98	1.30	8.00	7.71	8.35	16.07	14.38	18.80	11.47	11.13	11.88	13.79	12.87	14.83	6.29	5.32	7.86	21.27	21.08	21.49
cp	1.12	0.99	1.29	8.01	7.69	8.31	16.06	14.44	18.80	11.46	11.12	11.87	13.80	12.84	14.81	6.29	5.34	7.88	21.27	21.08	21.49

Table 2.2. Substitution rate and dC_{1+2}/dC_3 of stalk and head domain for each influenza subtype. Compared to the head domain, the stalk domain with lower value of dC_{1+2}/dC_3 experiences stronger purifying selection to maintain its conserved functionality.

Datasets	Substitution rates (subs/site/year)				dC_{1+2}/dC_3*			
	Stalk 1+2*	Stalk 3**	Head 1+2	Head 3	Stalk	Head	BF++	Overall^
H1N1pdm09	3.80E-03	1.18E-02	6.78E-03	1.42E-02	0.32	0.48	53	0.42
H1N1postpdm	1.26E-03	6.37E-03	2.16E-03	6.07E-03	0.20	0.36	+Inf	0.28
A/H3N2	1.39E-03	6.28E-03	3.66E-03	5.42E-03	0.22	0.68	+Inf	0.43
B-Victoria	5.44E-04	4.04E-03	1.51E-03	4.63E-03	0.13	0.33	8804	0.21
B-Yamagata	5.12E-04	4.93E-03	1.39E-03	5.56E-03	0.10	0.25	8875	0.18
H7N9	1.48E-03	7.48E-03	3.05E-03	7.80E-03	0.20	0.39	+Inf	0.30
H5Nx-full	1.74E-03	7.23E-03	6.11E-03	7.03E-03	0.24	0.87	+Inf	0.43

+: dC_{1+2}/dC_3 means the ratio of substitution rate of codon positions 1 and 2 over the rate of codon position 3.

++: BF is Bayes Factor. BF>3 represents statistical significance.

*: 1+2 represent the mean substitution rates of the codon positions 1 and 2.

**: 3 means the codon position 3.

^: overall dC_{1+2}/dC_3 is calculated from the SRD06 codon model for the whole HA gene.

Table 2.3. Tree root height estimation and its 95% Bayesian credible interval (BCI) from each model for all H5Nx datasets. The statistical significance of the new model holds across different subsampling datasets. Only 20% dataset has the trend to slightly overestimate the root height and would not report the precise evolution history of H5Nx. In sum, the new model is not sensitive to sample size. The unit of root height is in years.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

Models	H5Nx-full			H5Nx-80%			H5Nx-60%			H5Nx-40%			H5Nx-20%		
	Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI		Height	95% BCI	
HKY	21.27	21.08	21.48	21.31	21.08	21.57	21.34	21.07	21.62	21.42	21.07	21.79	21.64	21.07	22.19
c	21.29	21.08	21.48	21.30	21.08	21.57	21.32	21.07	21.61	21.43	21.07	21.81	21.65	21.08	22.20
p	21.28	21.08	21.49	21.32	21.08	21.57	21.32	21.07	21.59	21.41	21.07	21.79	21.65	21.08	22.20
cp	21.28	21.08	21.49	21.33	21.08	21.58	21.32	21.07	21.59	21.42	21.07	21.80	21.66	21.09	22.23

Figure 2.1. Conceptual diagram for influenza virus Hemagglutinins (HA) domains and four models tested

1a. HA protein 3D structure: Stalk and Head domains. Head domain is the immunodominant with high plasticity, which contains the receptor binding sites. Stalk domain is immune-subdominant and more conserved. ***1b. Linear diagram for functional domains on HA nucleotide sequence***. The stalk domain has two parts separated by head domain on the linear sequence. To compensate for short length, signal peptide, transmembrane domain, and cytoplasmic tail are combined into one partition, referred to as "STC" in the analysis. ***1c. Model conceptual design.*** Four models are applied and compared for each dataset. **HKY model** is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. **C model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. **P model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. **CP model** combines both C and P models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

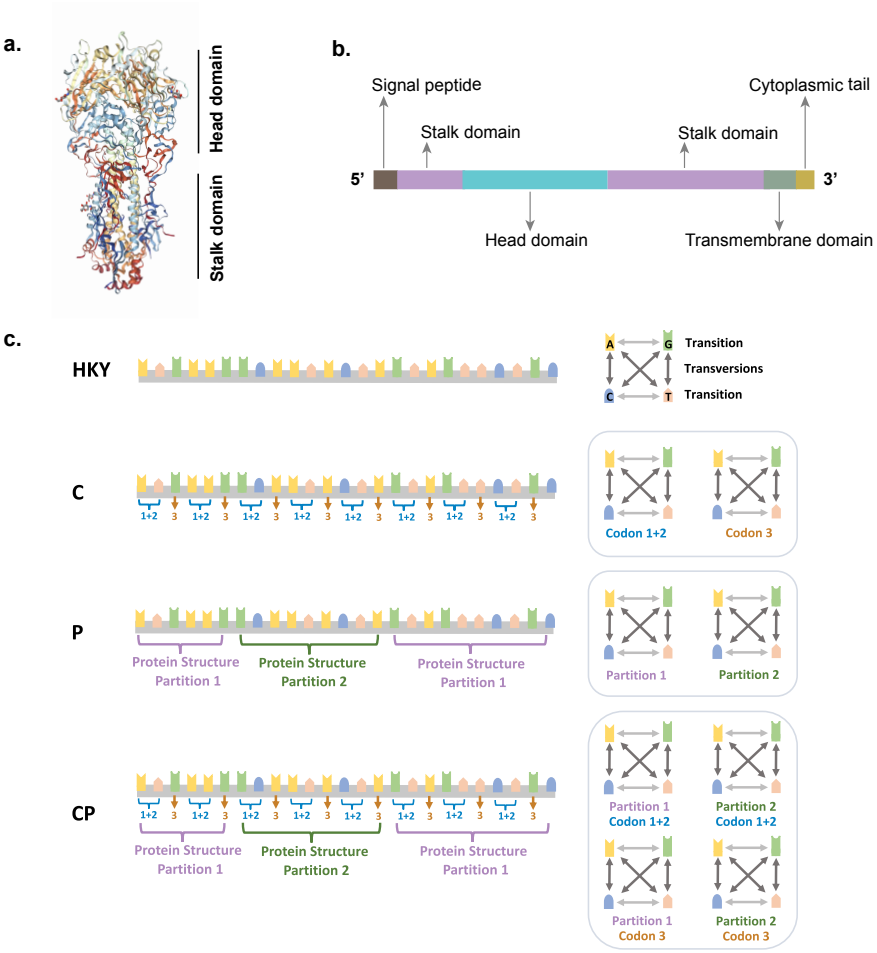

a.
Head domain
Stalk domain
b.
Signal peptide
Cytoplasmic tail
Stalk domain
Stalk domain
5'
3'
Head domain
Transmembrane domain
c.
HKY
A
G
C
T
Transition
Transversions
Transition
C
1+2
3
Codon 1+2
Codon 3
P
Protein Structure
Partition 1
Protein Structure
Partition 2
Protein Structure
Partition 1
Partition 1
Partition 2
CP
Partition 1
Codon 1+2
Partition 2
Codon 1+2
Partition 1
Codon 3
Partition 2
Codon 3

Figure 2.2. Log scale Bayes Factors for model selection procedure for each influenza subtype. All these Bayes Factors (BF) are calculated to compare with HKY via path-sampling (upper panel) and stepping-stone sampling (lower panel) approaches to compute the marginal likelihood estimates for each model. The asterisk * means a significant supportive BF for the structurally informed codon model (cp), compared to SRD06 codon only model (c model). Both cp and c models perform better than HKY model with decisive BF supports in all datasets. The HKY plus structurally informed model (p model) perform better in all datasets except for H7N9.

Log scale BF criteria: BF< 5 represents no significance; 5 ≤ BF< 10 means substantial support; 10 ≤ BF< 15 means strong support; 15 ≤ BF < 20 means very strong support; BF ≥ 20 means decisive support.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

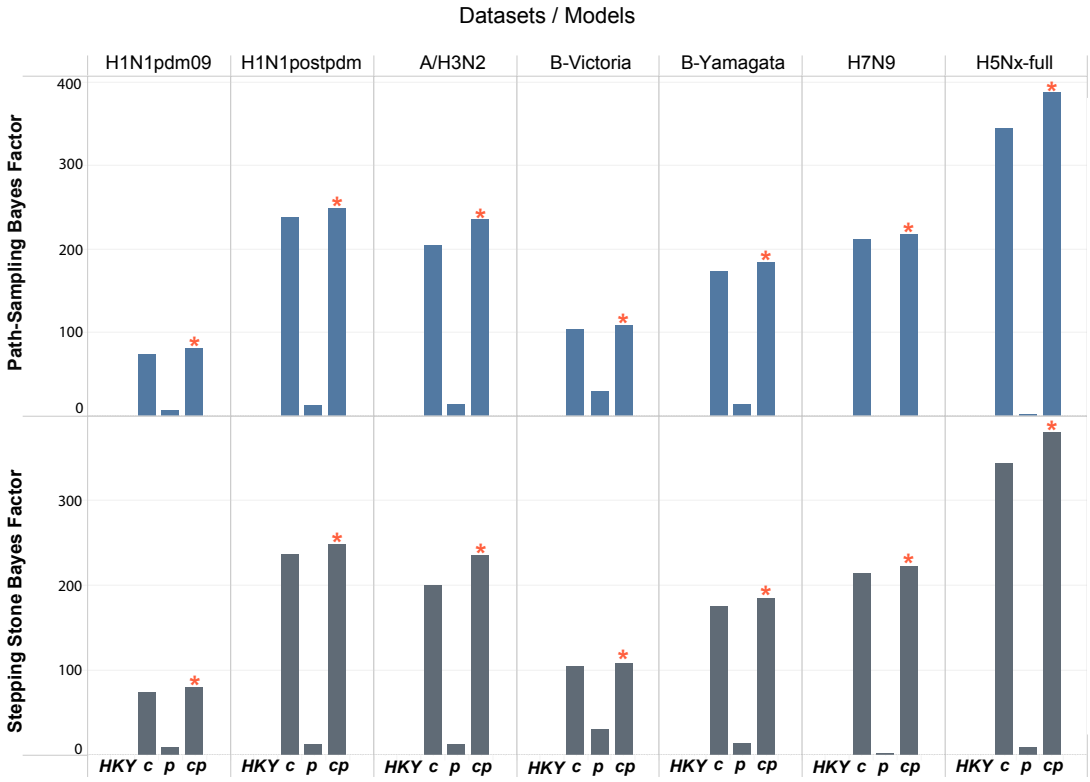

Datasets / Models
H1N1pdm09
H1N1postpdm
A/H3N2
B-Victoria
B-Yamagata
H7N9
H5Nx-full
Path-Sampling Bayes Factor
Stepping Stone Bayes Factor
0
100
200
300
400
HKY c p cp

Figure 2.3. Overall mean substitution rates estimated from four models and domain-specific substitution rates from structurally informed partitioning model. Violin plots show the substitution rates from each step of Bayesian simulation. The results show overall mean substitution rates from four models for each dataset are similar, but the head domain has a significantly faster rate than that of stalk domain based on the new model (cp model). Generally, STC domain has more uncertain estimation with long tails of the violin plot, which is probably due to short length of nucleotides. Bayes Factors for the statistical tests on the differences of head and stalk domain substitution rates are 1,137 for H1N1pdm09, 6 for H1N1postpdm, 577 for A/H3N2, 197 for B-Victoria-like, 191 for B-Yamagata-like, 303 for H7N9 and +Infinity for H5Nx.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

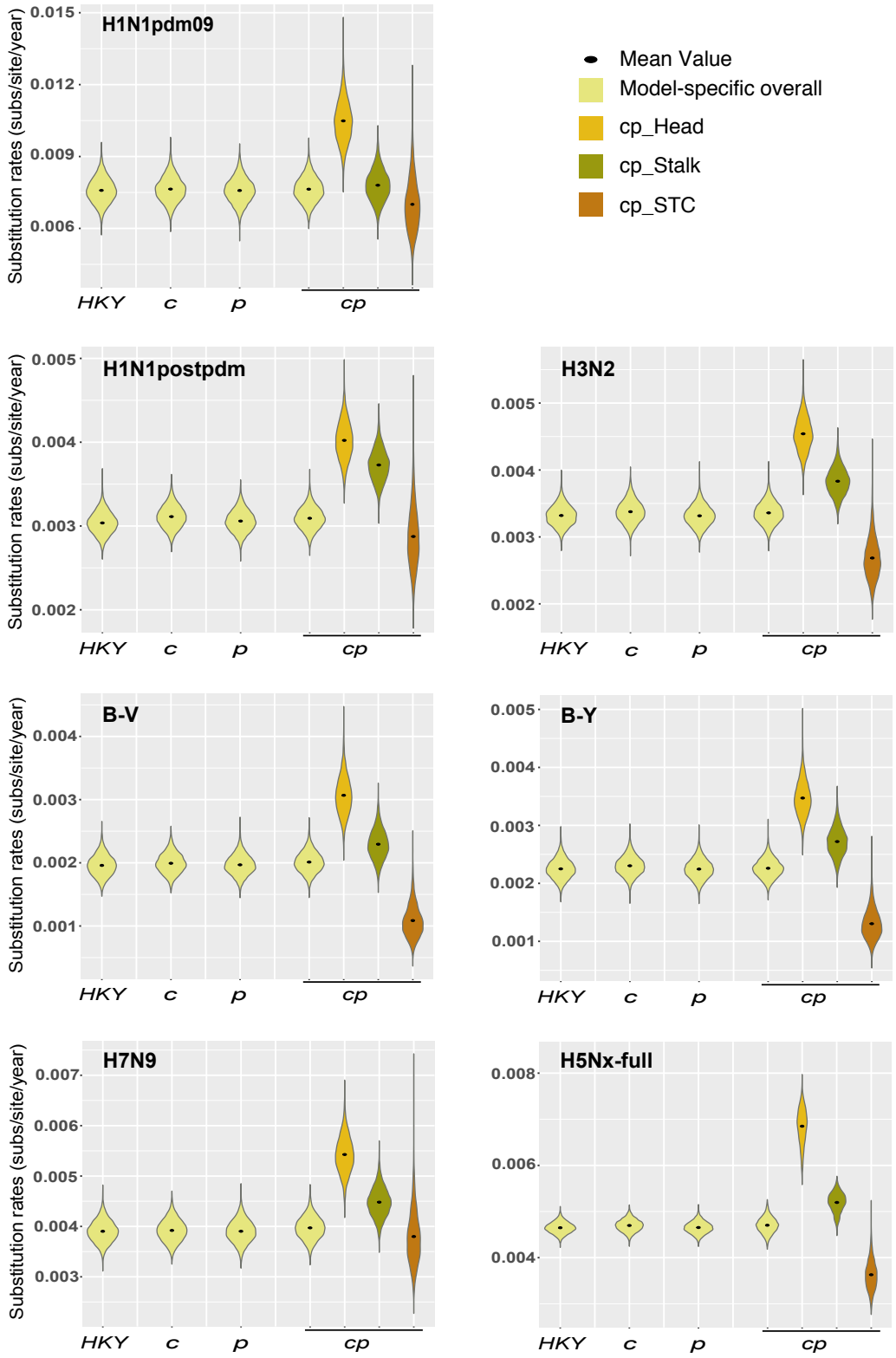

H1N1pdm09
H1N1postpdm
H3N2
B-V
B-Y
H7N9
H5Nx-full
Substitution rates (subs/site/year)
HKY
c
p
cp
Mean Value
Model-specific overall
cp_Head
cp_Stalk
cp_STC

Figure 2.4. Log scale Bayes Factors for sensitivity analysis on full and subsampled HPAI H5Nx datasets. All these Bayes Factors (BF) are calculated to compare with HKY via path-sampling (upper panel) and stepping-stone sampling (lower panel) approaches to compute the marginal likelihood estimates for each model. The asterisk * represents a significant supportive BF for the structurally informed codon model (cp), compared to SRD06 codon only model (c model). Both cp and c models perform better than HKY model with decisive BF supports. The HKY plus structurally informed model (p model) performs better with supportive BFs in all datasets.

Log scale BF criteria: BF< 5 represents no significance; 5 ≤ BF< 10 means substantial support; 10 ≤ BF< 15 means strong support; 15 ≤ BF < 20 means very strong support; BF ≥ 20 means decisive support.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

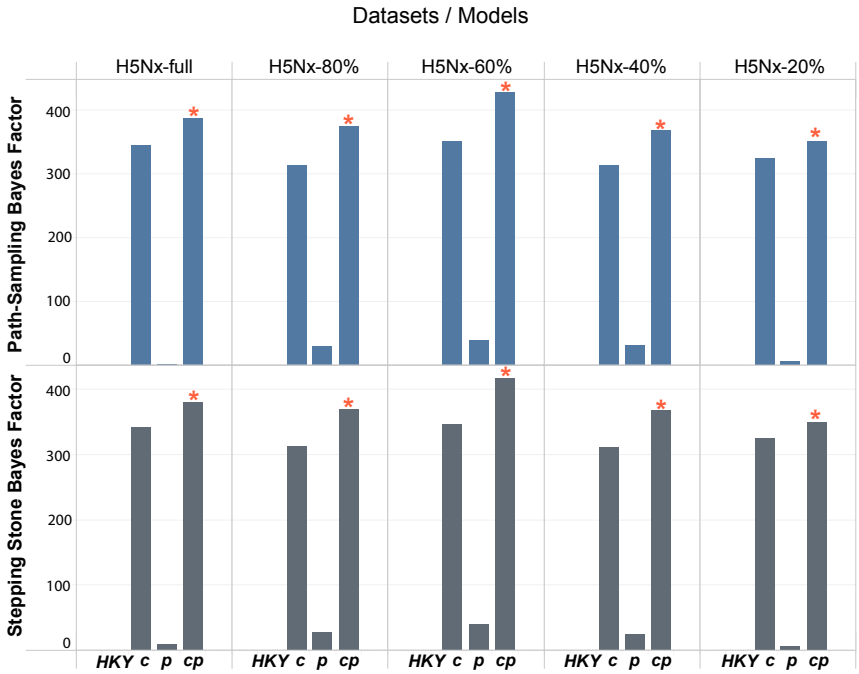
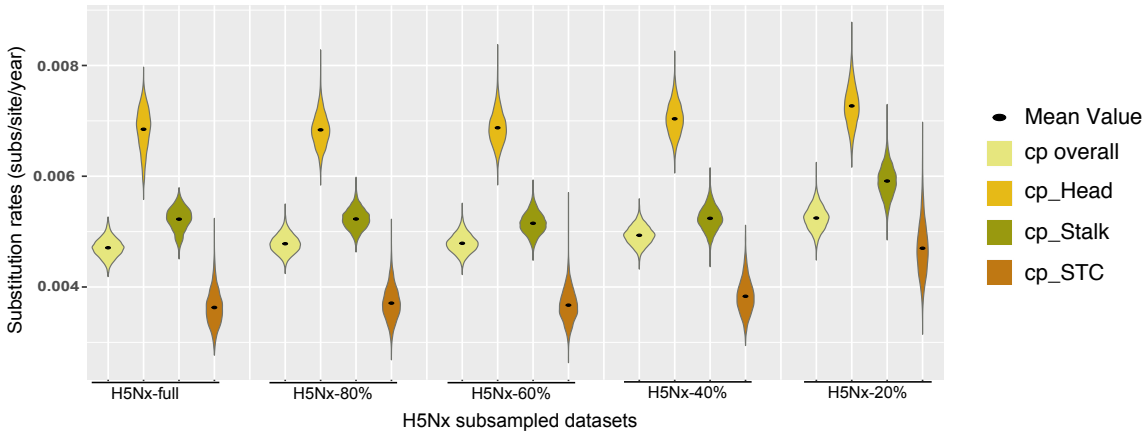

Datasets / Models
H5Nx-full
H5Nx-80%
H5Nx-60%
H5Nx-40%
H5Nx-20%
Path-Sampling Bayes Factor
Stepping Stone Bayes Factor
0
100
200
300
400
HKY c p cp

Figure 2.5. Sensitivity analysis: overall mean substitution rates and domain-specific substitution rates from structurally informed partitioning model for HPAI H5Nx subsampled datasets. Violin plots show the substitution rates from each step of the Bayesian simulations. Only 20% dataset has the trend to slightly exaggerate the rates and would not report the precise evolution history of H5Nx. The model superiority generally holds in the sensitivity analysis and the model is not sensitive to sample size.

Models: The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. It uses HKY substitution model.


Substitution rates (subs/site/year)
0.008
0.006
0.004
H5Nx-full
H5Nx-80%
H5Nx-60%
H5Nx-40%
H5Nx-20%
H5Nx subsampled datasets
Mean Value
cp overall
cp_Head
cp_Stalk
cp_STC

CHAPTER 3

SUB AIM 1. A CASE STUDY OF STRUCTURALLY INFORMED EVOLUTIONARY MODELS ON RSV F GENE

Introduction

As elaborated in Chapter 2, the novel structurally informed evolutionary model has been developed with influenza viruses. Intuitively, there is potential to apply this approach to other viruses with spike-shaped protein that have multiple functional domains including immunogenic stimulators. Therefore, to test the generalizability of the structurally informed partitioning scheme to improve phylogenetic inference, Human Respiratory Syncytial Virus (RSV) is used as a case study to extend the application of partitioning phylogenetic modeling to other respiratory viral pathogens. Below is a brief introduction of this pathogen.

RSV is the most important viral agent of severe acute pediatric lower respiratory tract illness (LRI) worldwide, commonly infecting infants and children under 5 years (9,53,178,179). RSV infections mostly cause mild flu-like symptoms but some can develop severe respiratory diseases, such as pneumonia and bronchiolitis (4,178). Epidemic seasonality of RSV is an annual or two-yearly epidemic in the temperate regions with unclear pattern for the tropics (178). Besides infants and children under 5 years, the elderly and immunosuppressed individuals also suffer higher risk of RSV infections and severe outcomes (53). The annually global disease burden of RSV includes about 3.2 million hospital admissions, 59,600 in-hospital deaths in children under 5 years, and the overall RSV acute LRI mortality is about 118,200 cases (53,180).

Evidence has shown severe RSV disease early in life is associated with lingering abnormalities in pulmonary function, which indicates vaccine for preventing from RSV infection is the best approach to protect susceptible populations (181). Unfortunately, due to the poor growth in vitro and physical instability of RSV, research on effective vaccine and antiviral therapeutic drugs has not been successful (181–183).

RSV is the member of family *Paramyxoviridae*, an enveloped and cytoplasmic virus (178). Serologically, it has been classified into two antigenic subgroups, RSVA and RSVB, which can co-circulate in one epidemic season (178,184,185). Genetically, it has a single-stranded, non-segmented, and negative-sense RNA genome (182). The 10 genes in the genome encode for 11 separate functional proteins (182). Fusion (F) glycoprotein, a viral surface protein with a globular head domain and the stem stalk domain, initiates the infection by fusing the virion membrane with a targeted host cell membrane. It is also the major antigen, inducing protective immune responses to neutralize infectivity efficiently or inhibit viral fusion process (127,184,186). Therefore, F protein plays an important role on vaccine design and candidate selection (187), where the antigenic sites and functional domains of RSV F glycoprotein have been defined for understanding the protein structure (127,186–188). One special feature of F protein is that it has significant rearranged conformations before and after the viral fusion process (for details refer to the paper (127)). Briefly, the most obvious change between the two conformations is the protein packing of the heptad repeat A and B (HRA and HRB) regions. Before the fusion process, the HRA region is packed into the globular head domain, exposed on the protein surface, and HRB itself forms the stalk domain (127). In the post-fusion conformation, HRA moves down to the stalk domain, together with HRB to form the coiled coil structure of the stalk domain (127). Understanding how immune selection can act differentially

on the pre- and post-fusion orientations of the F gene may provide valuable insights for antigen selection and vaccine development.

With knowing which domains of RSV F protein are located in the pre- and post-fusion conformations, questions can be asked: what are the evolutionary rates of these different structural domains of F protein given that they function differently before and after the fusion process? Can these differences be inferred and explained by immune selection and how this may inform F-protein based vaccine design? To answer these questions, I have developed a structurally informed evolutionary models to analyze the F protein of RSVA and RSVB to explore the evolutionary history and approximate selection pressure of each domain, while considering the different conformations of F protein in pre- and post-fusion process.

Methods

Publicly available data of RSVA and RSVB F gene sequences were downloaded from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). After sequence alignment and removing the duplicates, 518 RSVA (1977-2015) and 364 RSVB (1977-2016) were included in two separate datasets. Based on the pre- and post-fusion structure of F protein, the structural partitioning scheme included 4 partitions: STC (signal peptide, transmembrane domain and cytoplasmic tail), HRA (heptad-repeat A, located on the head domain of pre-fusion F protein but on the stalk domain post-fusion), HRB (heptad-repeat B, located on the stalk both pre- and post-fusion) and Other-Head (located on the head domain both pre- and post-fusion). The model simulations and model selection approaches were the same as what have been developed in Chapter 2. Briefly, four models were compared: HKY substitution model (HKY model), SRC06 codon position model (c model), HKY substitution model with protein structure partitioning scheme (p model),

and SRD06 codon position model with protein structure partitioning scheme (cp model). All partitioning models applied HKY substitution model for each partition. The uncorrelated relaxed molecular clock with a log-normal distribution was applied to infer the evolutionary rate from the dated samples (172). A smooth and time-aware Gaussian Markov random field (GMRF) process prior on the population sizes in the Skyride calescent model was used to incorporate population demographic dynamics (161). Model selection procedure and reports of parameters resembled the steps in Chapter 2. All xml files can be found here: https://github.com/XuetingQiu/RSVPartitions.

Results

Model selection procedures with two sampling approaches (path-sampling (PS) and stepping-stone (SS) sampling) reported similar Bayes Factors (BF) which statistically supported the cp model. The cp model performed the best for both RSVA and RSVB (Figure 3.1). This indicates that the cp model both statistically fitted the data better and captured more biological reality without over-parameterization. Checking the stability of the model performance, phylogenetic tree root heights (Table 3.1) and nucleotide substitution rates (Figure 3.2) from each model were reported. It demonstrated very similar tree root heights and 95% Bayesian Credible Interval (95% BCI) across different models for RSVA and RSVB, respectively. Violin plot to show the distribution of substitution rates from each Markov Chain Monte Carlo (MCMC) simulation in Figure 3.2 indicated that the overall mean substitution rates from each model were similar. Together, these evidences indicated that the cp model had stable performance and provided accurate estimates on important evolutionary parameters compared to other models.

With the cp model, partition-specific rates were reported (Figure 3.2). Focusing on the domain-specific rates, the model estimated that in RSVA, the highest rate was Other Head region, followed in order by the rate of HRA and then of HRB (i.e. Other Head > HRA > HRB). In RSVB, the order was slightly different, where HRB had the highest rate, followed in order by the rate of Other Head and then HRA (i.e. HRB > Other Head > HRA). Furthermore, the cp model can estimate the ratio of substitutions rate of codon positions 1 and 2 (majority of non-synonymous changes occur in these codons) over the rate of codon position 3 (majority changes are synonymous), that is, $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, for each functional partition separately. The relative rate of C_3 to C_{1+2} is a good predictor of the ω parameter (137) and could reflect the partition specific selection pressure. As shown in Table 3.2, the values of $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ were 0.10, 0.18 and 0.13 for RSVA Other Head, HRA, and HRB region, respectively. For RSVB, the values of $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ were 0.12, 0.15 and 0.09 for Other Head, HRA, and HRB region, respectively. Though the evolutionary rates and selection pressure showed differences between RSVA and RSVB, the HRA region in both RSVA and RSVB had the highest $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, indicating that HRA experienced weaker purifying selection, which was statistically significant (BF>3).

Discussion

This extended case study on RSV further confirms that the structurally informed codon position model statistically fits the sequence data better and captures more biological reality without over-parameterization. Stable performance on inferring evolutionary parameters and extra information (partition-specific substitution rate and $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ to indicate approximate selection pressure) on the evolutionary history make this new model valuable for vaccine design. Distinctive partition-specific rates of RSVA and RSVB F gene were estimated from the model,

which may indicate different evolutionary history of these two types and further different strategies needed for each vaccine design. The different estimated selection pressure can be used to understand the evolution and biological behaviors of each partition, which may provide new insights for vaccine design. HRA was inferred to experience weaker purifying selection pressure compared to other regions, which may be due to its unique feature of migrating from head domain to stalk domain during the fusion process (127). It indicates that the fusion process may affect the behavior of the partition, which should be taken into consideration for vaccine design.

One main limitation of this study is data sampling biases, where the amount of RSVB isolates is much less than RSVA, though both types need more efforts on viral sampling and sequencing (189). The large uncertainty due to small sample size are also reflected from the wider 95% BCI on the root height. Better surveillance efforts to collect sufficient data on the distribution of global RSV and its molecular epidemiology will improve the inferences on viral evolution (189), including more accurate partition-specific rates estimated by the new model. Another limitation is that the biological mechanisms to explain the behavior and evolutionary characteristics of each partition need to be explored. This process may lead to identify critical antigenic sites for vaccine design. Nevertheless, the structurally informed model can be extended to other respiratory viruses with distinctive structures and functions of protein domains and provide additional information on the partition-specific evolutionary history.

Tables and Figures

Table 3.1. Tree root height and its 95% Bayesian Credible Interval (BCI) from each model for RSVA and RSVB. Results from four models show very similar tree root height and 95% BCI for RSVA and RSVB, respectively. The unit of root height is in years.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.

Models	RSVA			RSVB		
	Height	95% BCI*		Height	95% BCI*	
HKY	46.11	42.40	50.43	48.55	44.26	53.58
c	46.11	42.48	50.66	48.50	44.30	53.33
p	46.10	42.27	50.50	48.55	44.31	53.55
cp	46.06	42.01	50.53	48.46	44.02	53.13

Table 3.2. Substitution rate and dC_{1+2}/dC_3 of each partition for RSVA and RSVB. Lower value of dC_{1+2}/dC_3 indicates higher purifying selection pressure on this partition to maintain its conserved functionality. The differences between partition-specific rates are statistically supported with Bayes Factor >3. The unit of the substitution rate in each column is substitutions/site/year.

Partitions	RSVA			RSVB		
	Codon3$^+$	Codon1+2$^+$	dC_{1+2}/dC_3*	Codon3	Codon1+2	dC_{1+2}/dC_3
STC	2.18E-3	7.23E-4	0.33	2.05E-3	8.10E-4	0.40
Other Head	1.92E-3	2.00E-4	0.10	2.03E-3	2.48E-4	0.12
HRA	1.57E-3	2.81E-4	0.18	1.83E-3	2.81E-4	0.15
HRB	1.40E-3	1.77E-4	0.13	2.31E-3	2.13E-4	0.09
Overall F[#]	1.85E-3	3.23E-4	0.18	2.04E-3	3.60E-4	0.18

*dC_{1+2}/dC_3 means the ratio of substitution rate of codon positions 1 and 2 over the rate of codon position 3.

+ Codon3 means the nucleotide substitution rate of codon position 3; Codon1+2 means the nucleotide substitution rate of codon positions 1 and 2.

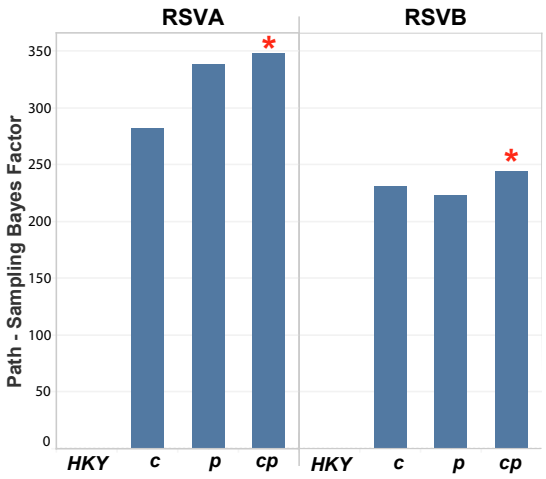
Overall F protein. The rates for this overall come from the c model. The partition-specific rates are generated in cp model.

Partition Abbreviations: STC represents the combination of signal peptide, transmembrane domain and cytoplasmic tail; HRA represents heptad-repeat A, which locates on the head domain of pre-fusion F but on the stalk domain post-fusion; HRB represents heptad-repeat B, which locates on the stalk both pre- and post-fusion; and Other Head represents other regions except for prefusion HRA that form and locate on the globular head domain both pre- and post-fusion.

Figure 3.1. Log scale Bayes Factors for model selection procedure for RSVA and RSVB. All these Bayes Factors (BF) are calculated to compare with HKY via path-sampling (left panel) and stepping-stone sampling (right panel) approaches to compute the marginal likelihood estimates for each model. The red asterisk * means a significant supportive BF for the structurally informed codon model (cp), compared to SRD06 codon only model (c model). Both cp and c models perform better than HKY model with decisive BF supports in all datasets.

Log scale BF criteria: BF< 5 represents no significance; 5 ≤ BF< 10 means substantial support; 10 ≤ BF< 15 means strong support; 15 ≤ BF < 20 means very strong support; BF ≥ 20 means decisive support.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.


RSVA
RSVB
Path - Sampling Bayes Factor
0
50
100
150
200
250
300
350
*
*
HKY
c
p
cp
HKY
c
p
cp

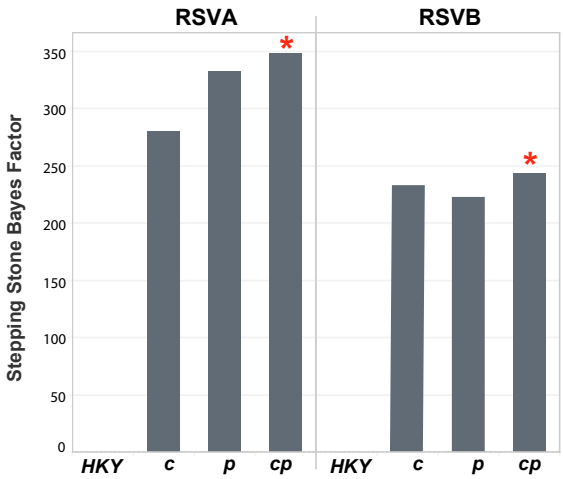
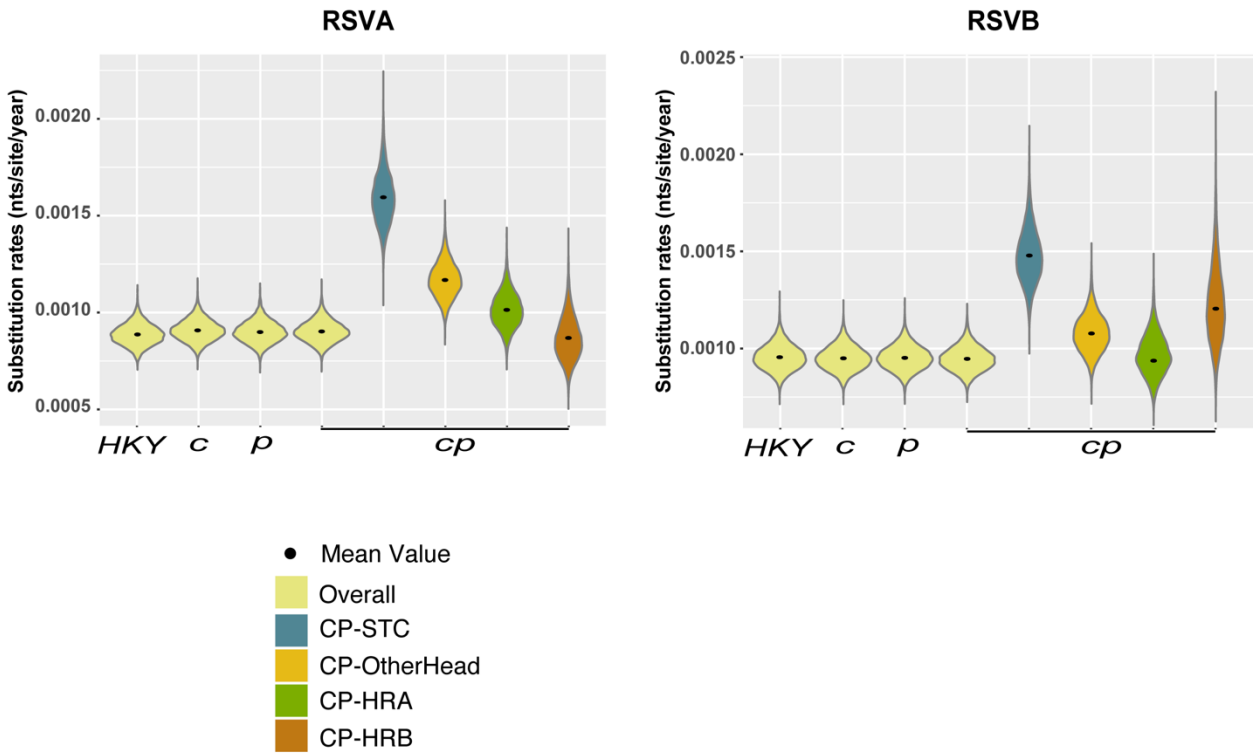

RSVA
RSVB
Stepping Stone Bayes Factor
0
50
100
150
200
250
300
350
*
*
HKY
c
p
cp
HKY
c
p
cp

Figure 3.2. Overall mean substitution rates estimated from four models and domain-specific substitution rates from structurally informed partitioning model. Violin plots show the substitution rates from each step of the Bayesian simulations. The results show overall mean substitution rates from four models for each dataset are similar, but domain-specific rates from the cp model vary within each dataset.

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. All models use HKY substitution model.


RSVA
RSVB
Substitution rates (nts/site/year)
0.0005
0.0010
0.0015
0.0020
0.0025
HKY
c
p
cp
Mean Value
Overall
CP-STC
CP-OtherHead
CP-HRA
CP-HRB

CHAPTER 4

DIFFUSION DYNAMICS OF SEASONAL INFLUENZA VIRUSES IN THE U.S. DRIVEN BY FLIGHT CONNECTIONS AND HIGH-RISK POPULATIONS [2]

[2] Qiu, X., Chen, J., Hicks, J.T., et al. To be submitted to *PLOS Computational Biology*, 12/2019.

Abstract

Understanding the viral diffusion patterns and environmental factors that affect the dynamics is critical for the prevention and control of influenza outbreaks. Building upon the knowledge learned from other global studies on seasonal influenza virus diffusion, in this paper we explored the global introductions into the U.S. and studied the diffusion dynamics inside the U.S. We took advantage of a large and complete genetic and epidemiological dataset of all four seasonal influenza virus subtypes to answer important questions regarding diffusion patterns and ecological/epidemiological predictors of viral spread. Analysis demonstrated that different global regions contributed viruses to seasonal epidemics in the U.S., where East Asia and Pacific (EAP), Europe and Central Asia (ECA), and Latin America and Caribbean (LAC) were the main sources of transmission to the U.S. Each U.S. epidemic season may have a different region acting as the major viral contributor to the outbreak. Dynamics amongst different Health and Human Services (HHS) Regions in the U.S. demonstrated a primary hub pattern: Region 5 for A/H3N2 and B-Victoria, Region 4 for A/H1N1, and Region 8 for B-Yamagata. With detailed epidemiological and ecological factors for these regions, important predictors for disease transmission were identified as geographic distance, flight connections, different population age structures and average precipitation. Longer geographic distance is a barrier for viral diffusion. More flight connections, lower proportion of adult (18 to 64 years old), and higher precipitation strongly correlates with viral spread. These results suggest that increased hygiene in the high traffic occasions and targeted vaccine administration to the high-risk populations may impact viral spread among communities at national scales.

Introduction

Seasonal influenza viruses infect the respiratory tract, causing epidemics that spread rapidly within communities across the globe (1,2). These viruses cause heavy disease burden in the United States (U.S.), with especially high impacts on children and the elderly (9) (190). While influenza vaccines are amongst the most effective ways to prevent infections, understanding the spatiotemporal diffusion patterns of seasonal influenza viruses is critical for tracking the disease spread and exploring the factors that affect outbreaks (111,140). Previous studies either based on the analysis of epidemiological data alone or combined with genetic information (phylodynamic modeling) have provided valuable inferences on spatiotemporal diffusion dynamics (19–22,46,141). Bayesian phylodynamic modeling has been well developed as a powerful tool to understand viral diffusion dynamics (98,142). A distinct advantage of Bayesian phylodynamic modeling is that models of rapid viral genetic evolution can be unified with epidemiological and ecological predictors to both improve the accuracy of evolutionary reconstruction and reveal statistical associations between the predictors and viral spread rates (98,104). The rational of unifying genetic and epidemiological data is that the evolutionary and epidemiological processes of rapidly evolving pathogens are on the same time scale (28,153). Bayesian framework phylodynamic approaches can connect genetic sequence data to trait evolution for rapidly evolving pathogens, where the traits can be host, geographic location, or antigenic information. Furthermore, the technique of Bayesian stochastic search variable selection (BSSVS) (98) and generalized linear model (GLM) (111) can incorporate and test the significance of the covariates relating to the pathogen evolutionary and epidemic processes in the phylogenetic reconstruction procedures (98).

Recent years, the application of Bayesian phylodynamic modeling and new developed analysis techniques have provided a clear depiction of global dynamics and the source populations for the diffusion of seasonal influenza viruses (18,19,21,46). Via source-sink phylogeographic diffusion modeling, Rambaut et al. (46) studied how genomic processes relate to global influenza dynamics with A/H3N2 and A/H1N1 in temperate regions during 1992-2005, demonstrating that seasonal influenza A viruses have a complex interaction between periodic antigenic drifts and reassortments. Bedford et al. (21) later developed a phylogenetic tree trunk reconstruction models to reconstruct the evolutionary history and global dynamics of A/H3N2 collected from 1998 to 2009, revealing that China and Southeast Asia seed global outbreaks. With Bayesian phylogeographic analysis applied to influenza A/H3N2 viruses isolated from global urban centers during 2003-2006, Bahl et al. (19) have reported that tropical regions did not function as a source for A/H3N2 yearly dynamics, but that each outbreak region could have multiple external sources and could also play as a potential source viral population. The important role of population migration to diffuse viruses has been further studied by Lemey et al. (18) via GLM to unify viral genetics and human air transportation data to explore the global transmission dynamics of A/H3N2. They found that human migration via air transportation drives global transmission of A/H3N2 and longer geographic distance is a key barrier for the scales of local spread (18). The most complete study to date on global diffusion patterns of seasonal influenza has been conducted by Bedford et al. with large and long-term datasets (20). Phylodynamic analysis was performed on A/H3N2, A/H1N1, B-Victoria, and B-Yamagata viruses collected during 2000-2012 to explore and compare global circulation patterns of these major influenza virus lineages. The analyses showed that genetic variants of A/H3N2 viruses did not persist locally between epidemics but were reseeded from East and Southeast Asia. In

contrast, A/H1N1 and type B viruses persisted across several seasons and exhibited complex global dynamics. They concluded that there may be viral, host and ecological factors that are complicating the global dynamics. These studies integrated both genetic and epidemiological information, which offered potential for epidemiological surveillance through phylodynamic reconstructions.

Many studies have reported that epidemiological and ecological factors may impact the diffusion of seasonal influenza viruses. Global or local studies mainly based on epidemiological data (141,143–145) have reported that temperature, humidity, precipitation, host movement, population size, and air transportation can affect the diffusion and dynamics of seasonal influenza viruses. In a recent study, Dalziel et al. (22) used weekly incidence data of influenza-like illness from 603 cities in the U.S. during 2002 – 2008 to explore important spatial variations, which revealed that population size and age structure, humidity, and peak climatic conditions of urban centers can drive the incidence of influenza infections and its spatiotemporal dynamics. While this study provides evidence that environmental characteristics may alter herd immunity for different levels of urbanized population, viral genetic data was not incorporated into the analysis. Phylogenetic approaches conducted on a more recent and systematic dataset could quantitatively evaluate the effects of environmental factors on viral diffusion patterns.

As summarized above, previous phylodynamic analyses have provided insights to global spatiotemporal diffusion of seasonal influenza viruses. However, local diffusion dynamics, for example the specific source populations of seasonal influenza viruses into the U.S. and the diffusion patterns amongst important U.S. regions, have not been well explored and quantified. By identifying the close geographic region sources from global settings and transmission connections in the U.S., a better understanding of viral dynamics could improve circulating strain

prediction and vaccine selection. Therefore, the major goals in this study were to characterize viral introductions and quantitatively estimate diffusion patterns of each subtype of seasonal influenza in the U.S. with the most abundant and complete genetic and epidemiological data to date (period 2011-2018). Furthermore, phylodynamic approach with generalized linear model (GLM) was applied to explore the association between epidemiological and ecological predictors and seasonal influenza transmission among U.S. regions to reveal insights for disease prevention.

Methods

Data management

To explore seasonal influenza virus introductions into the U.S., viral HA sequence datasets of the four major seasonal influenza subtypes/lineages were collected, including A/H1N1 (2011-2018), A/H3N2 (2011-2018), B-Victoria-like (2011-2018), and B-Yamagata-like (2011-2018). Sequence data were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/). The following inclusion and exclusion criteria were applied: a) the minimum amount of epidemiological data for each sequence must include collection date (at least year and month information) and country information; b) the minimum length of the sequences should be more than 50% of the full gene length; c) vaccine, derivative, recombinant and laboratory sequences should be excluded; d) duplicate sequences with the exactly same location and 100% similarity should be excluded with the oldest dated strain retained.

Sequence datasets were aligned separately with the high accuracy and high throughput software, MUSCLE v3.8.31 (191). Manual alignment was used to manage unreasonable insertions and deletions. Primary phylogenetic analysis was conducted with maximum-likelihood

(ML) approaches in Randomized Axelerated Maximum Likelihood (RAxML) v8.0 (192), which has the advantage of handling large datasets. ML trees were examined in TempEst (193), employing root-to-tip divergence regression and distribution of residuals to assess for temporal signal of these heterochronous sequences and outlier removal.

Global datasets

After alignment, to guarantee sufficient sample size in each location and to reduce the dimension of the transmission matrix, these datasets were categorized by global geographic regions, where the U.S. was kept as a separate geographic location to infer the introductions into the U.S. from the global community. The global regions were coded into seven regions defined by the World Bank (194): East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA), South Asia (SAS), and Sub-Saharan Africa (AF). Details of the data distribution by year and by region can be found in the Supplemental Table II-1.

U.S. Region datasets

To explore the diffusion dynamic modeling between U.S. regions defined by the U.S. Department of Health and Human Services (HHS) (https://www.hhs.gov/about/agencies/regional-offices/index.html), the HA sequences of U.S. state level were identified. Modifying the 10 HHS regions (specific geographic components of each region were listed in the HHS website), Alaska from the Seattle region and Hawaii from the San Francisco region were treated as a separate region given their distinctive climate and

geographic locations, resulting in 12 geographic regions for the U.S. region-level analysis. Details of data distribution by year and by region can be found in Supplemental Table II-2.

Subsampling procedure

After categorized into regions, identical sequences collected from the same region were removed to reduce the sampling weight on the identical viral isolates. To examine potential over-representation of some regions, descriptive analysis by year and region was conducted to develop a subsampling strategy. After conducting descriptive statistics of metadata distributions, random sampling was used to generate a representative subset in each region (Details in Supplemental Text II-1). Multiple sets of subsampled data were generated to conduct a preliminary analysis by checking the estimated root height, evolutionary rates and potential outliers to confirm the subsampling strategy was proper. This process was conducted for both global data and U.S. domestic region-level data.

Covariates

To explore which epidemiological and ecological factors have impacts on the transmission patterns in the U.S., publicly accessible covariate measures were collected, including demographic information, flight connections, climate data, vaccination rates, epidemic peak time and epidemic size.

Demographic information

Total population size and population age structure were based on the 2010 population census from the United States Census Bureau (https://www.census.gov/). Total population in each state was collected and then combined into the total population for each HHS region. Age

structure of the populations was also collected, where it divided the population into 3 categories: children/youth (<18 years), adults (18-64 years) and the elderly (≥65 years). Regional population density (persons/square miles) was calculated by dividing total regional population by HHS region area calculated by ArcGIS Pro (https://www.esri.com/en-us/arcgis/products/arcgis-pro/resources).

Total flight connections

The flight connections were collected from a publicly available data source from United States Department of Transportation (https://www.transtats.bts.gov/DL_SelectFields.asp). Since the scheduled flights in each month in each state had little variation during 2011-2018, the 2015 flight counts were used. With the airline data, each fight in the U.S. was identified at the state-level, and then counted into the monthly numbers of inbound and outbound flights for each state. All monthly state-level data during flu season (September to the next May) in the U.S. were summed up into HHS regions to generate a total number of region-level flights.

Climate data and geographic distance

For the climate covariates, average temperature, precipitation and relative humidity during 2011-2018 were included. These data were extracted from the National Oceanic and Atmospheric Administration (NOAA, https://www.ncdc.noaa.gov/). For temperature and precipitation, the monthly data were summarized from the website (https://www.ncdc.noaa.gov/cag/statewide/mapping/110/tavg/201002/1/value) to get the average temperature and precipitation for all states from November to the next year February as winter measures and from May to August as summer measures. The mean temperature (in Fahrenheit) and precipitation (in inch) from state-level data was used as the region-level data. For humidity data, the U.S. census shapefile data with 1:20,000,000 resolution level were used to define the

HHS regions, in which, Alaska and Hawaii were kept separately. The monthly mean data for relative humidity at 1:20,000,000 resolution dataset (https://www.esrl.noaa.gov/psd/data/gridded/data.narr.monolevel.html) was used to summarize the monthly mean of relative humidity from November to February during 2011-2018 for each HHS region, Alaska and Hawaii, with the unit as percentage. Similar process for summer relative humidity was calculated from May to August. The distance between regions was defined as the distance between the region centroids identified in ArcGIS (http://kb.mit.edu/confluence/pages/viewpage.action?pageId=10977309) and the unit was in kilometers.

Region vaccination rate and epidemic curve

Vaccination rates were collected from the U.S. Centers for Disease Control and Prevention (CDC; www.cdc.gov/flu/fluvaxview) during 2011-2018. The average vaccination rate was used for each HHS region, Alaska, and Hawaii. The epidemic curve data of clinical influenza positive samples were gathered from CDC FluView Interactive (https://www.cdc.gov/flu/weekly/fluviewinteractive.htm) for each state during 2011-2018 and then summarized into HHS regions. New epidemic curves were generated with the summarized region data. The epidemic size for each region was ranked based on the magnitude of the positive samples, where 1 represented the smallest epidemic size and 12 represented the largest. The epidemic peak week (the week with the highest count of positive samples) of each geographic category was identified for each season. Region with the earliest peak week was set as 0 and other regions took the differences for each season. The peak time during 2011-2018 was summed up and ranked from earliest to the last as 1-12.

Correlations between these covariates were calculated with Pearson Correlation (195). If the correlation coefficient was larger than 0.70 indicating high co-linearity between two covariates, one of the covariates was not included in the model. With the correlation test, it ended up with removing total population size and summer climate data but keeping population density and winter climate data in the model.

Phylogenetic analysis and empirical trees

To evaluate how many introductions into the U.S. from global regions, phylogenetic analysis was conducted in Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.10.3 (160). The statistically well-supported general time reversible (GTR) substitution model (93,94) was applied with gamma-distributed rate variation among sites. The lognormal relaxed molecular clock (98) was used with an initial mean of 0.0033 with a uniform prior ranging from 0.0 to 1.0, given that the estimates of substitution rates for influenza viruses were power to -3. Based on prior knowledge, a smooth and time-aware Gaussian Markov random field (GMRF) process prior on the population sizes was applied in the Skyride coalescent model (161).

Due to the large sample size, seven independent Markov Chain Monte Carlo (MCMC) chains of 200 million generations were simulated, sampling every 20,000 generations to yield 10,000 trees per run. The convergence of all seven runs was diagnosed in Tracer v1.7 (http://tree.bio.ed.ac.uk/software/tracer/) for all parameters to ensure a sufficient effective sample size (ESS > 200). LogCombiner v1.10.3 as part of the BEAST software package was used to combine the multiple runs to generate log and tree files after appropriate removal of the burn-in from each MCMC chain to guarantee convergence of these runs. The Maximum Clade Credibility (MCC) tree was summarized in TreeAnnotator v1.10.3 from the combined tree file.

The MCC tree was visualized in FigTree v1.4.4 (https://github.com/rambaut/figtree/releases), where the posterior probability for each node (>0.50) and the 95% Bayesian credible intervals (BCI) of node age were checked as indicators of phylogenetic estimation uncertainty. To further conduct discrete source-sink model and generalized linear model for covariates, an empirical distribution of phylogenetic trees was estimated from these converged runs.

Source-sink model

With dated samples and tip-associated geographic region characters, the continuous-time Markov chain (CTMC) model of discrete traits was applied to infer how these region traits have evolved with the viral population over the sampling time (98,109). In addition, the number of state changes (i.e., state transition events or Markov jumps) across the phylogeny for each state was estimated at the internal tree nodes. These Markov jumps were counted along each sampled phylogenetic tree (196) to represent the overall transitions between regions during the sampling timeframe. An asymmetric CTMC matrix model was used to estimate the transition rates and absolute Markov jump counts for both directions between two trait character states. The Bayesian stochastic search for variable selection (BSSVS) approach was employed to provide the most parsimonious diffusion process to randomly turn on (indicator = 1) and turn off (indicator = 0) the transmission between two geographic traits (98,109). Further, BSSVS allowed for the calculation of the Bayes Factor (BF) as a measure of statistical support for each transition rate. The BF was calculated for each transition rate using the posterior probability of non-zero indicators and the prior probability. The strength of the statistical support was interpreted with these criteria: BF < 3 indicated no significance; 3 ≤ BF <10 indicated substantial support, 10 ≤ BF < 30 indicated strong support, 30 ≤ BF <100 indicated very strong support, and BF ≥ 100

indicated decisively statistical support (106). BF calculation was done by the program SpreaD3 (197). Transmission rates were extracted and calculated using personalized Python scripts (found in the Github link: https://github.com/XuetingQiu/FluDiffusionUS).

Generalized linear regression

The generalized linear model (GLM) (18,198) was incorporated into the previously defined diffusion model to evaluate the impacts of multiple epidemiological and ecological factors on viral diffusion patterns across U.S. regions. The logarithm scale of the covariates was calculated and standardized by a normal distribution with mean as 0 and standard deviation as 1, to avoid the effects from the different magnitude of these covariates.

The GLM parameterized each rate of among-location transition in the diffusion model as a log-linear function of various potential predictors as shown in the following equation:

$$Log\gamma_{AB} = \theta_1\beta_1 X_{1AB} + \theta_2\beta_2 X_{2AB} + \cdots \theta_j\beta_j X_{jAB}$$

In which, A and B represented geographic regions, γ represented the estimated transition rate between A and B given the diffusion model, $X_{i,\,(1\le i\le j)}$ were the ecological/epidemiological factors of interest, $\beta_{i,(1\le i\le j)}$ were the coefficients on a log scale quantifying the effect size of the factor on the transmission rates, and $\theta_{i,(1\le i\le j)}$ were the binary indicators introduced by the BSSVS allowing the factor to be included or excluded from the model at each MCMC generation. Exploratory GLM analysis was applied separately to each subtype/lineage of seasonal influenza viruses, where each subtype had been fit with two models – one without sample size of each region and one with sample size to explore whether the sample size of each region had impacts on the model fitting. Then assuming all seasonal influenza A viruses in humans were transmitted by the same mechanisms a joint GLM estimation was used to evaluate the effects of a single set

predictors shared by all seasonal influenza A subtypes on viral diffusions, where the predictors were fit with four sets of rate matrices for all seasonal influenza subtypes/lineages. The same BF criteria as the source-sink model described in the previous section were used to identify statistically supported epidemiological and ecological factors. The coefficient, 95% BCI and BF were reported to quantitively estimate the conditional effect size of the factor and its statistical significance. Xml codes, python and R scripts can be found via the Github link: https://github.com/XuetingQiu/FluDiffusionUS.

Ethics Statement

All the genetic and epidemiological data used in this study are publicly available without the host's identification. Publicly available data from GISAID do not include protected health information or personal identifiers of patients from whom the samples were isolated.

Results

Global introductions of seasonal influenza into the U.S.

Data distribution and the phylogenies

After the subsampling scheme by region and by year, the final global data distribution can be found in Table 4.1. Since one of the goals was to capture the introductions into the U.S., higher proportion of seasonal samples in the U.S. was subsampled compared to other global regions. The phylogeny of each influenza virus subtype was reconstructed with Bayesian phylogenetic framework (Supplemental Figure II-1a-d). These trees showed clear ladder shape and bushy leaves, which corresponded to the strong seasonal patterns of outbreaks and immune selection. The medians of estimated substitution rates were 4.78E-3 (95% BCI: 4.46E-3, 5.13E-

3) and 4.45E-3 (95% BCI: 4.16E-3, 4.76E-3) for A/H1N1 and A/H3N2, respectively. B-lineages had lower substitution rates with medians as 2.33E-3 (95% BCI: 2.14E-3, 2.52E-3) for B-Victoria lineage and 2.52E-3 (95% BCI: 2.32E-3, 2.73E-3) for B-Yamagata lineage. The medians of the estimated most recent common ancestor time (tMRCA) were about 0.75 – 1.25 years prior to the oldest isolate in the dataset with very narrow 95% BCI (Supplemental Table II-3a), indicating less uncertainty in these estimations. The dynamics of effective population size for each subtype were reconstructed with Skyride coalescent model (Figure 4.1). We found that the effective population sizes of A/H1N1 and B-Victoria lineage showed clear seasonality, where each seasonal peak had slight variations on the magnitude. B-Yamagata lineage had a flat curve with weak signals of seasonal peaks. The 2018 B-Yamagata seasonal epidemic had the highest effective population size but with higher uncertainty. The most interesting pattern was found in A/H3N2, where it had earlier peak time compared to other subtypes and peak seasons were usually followed by a relatively flat season.

Diffusion dynamics

To estimate global introductions into the U.S., a discrete state source-sink phylogeographic model was applied to each dataset. For A/H1N1 (Figure 4.2a and Table 4.2), the statistically supported main global introductions into the U.S. were from Europe and Central Asia (ECA), East Asia and Pacific (EAP), South Asia (SAS) and North America (NA) with decisive Bayes Factors (BFs > 100). Though Middle East/North Africa (MENA) and Latin America/Caribbean (LAC) were also statistically supported, they had very low migration rates. Sub-Saharan Africa (AF) was not a supported source for the migration to the U.S. When breaking down U.S. data into seasons during 2011-2018, it was reported that each season had 2-6 supported sources, but for the most part, only one or two main sources (ECA or NA) contributed

to the majority of introductions for that season. The pie charts in Figure 4.3 showed the proportional source populations into the U.S. for each season. Two seasons (season 2013-2014 with Bayesian posterior probability as 0.24 and season 2016-2017 with posterior probability as 0.80) had introductions from U.S. previous season, which indicated that some of the isolates of A/H1N1 may either be circulating locally at low levels or circulating in a region that was not sampled but later transmitted virus into the U.S.

For A/H3N2 (Figure 4.2b and Table 4.2), the main global sources into the U.S. were from EAP, ECA, SAS and LAC with decisive BFs. NA and AF had lower migrations into the U.S., while MENA was not supported as a source. When looking into each season, it was reported that each season had different main sources and the majority were from ECA and EAP (Figure 4.3). Specifically, though NA had low migrations overall, but it had higher contributions for several seasons, for example, season 2012-2013 and season 2014-2015. No local persistence was found for A/H3N2, which indicated that A/H3N2 always required external introductions to sustain the outbreak in the U.S.

For B-Victoria lineage (Figure 4.2c and Table 4.2), the main global sources into the U.S. were EAP, LAC and ECA. AF was statistically supported, but it had very low migration rate. MENA, NA and SAS were not supported as viral sources into the U.S. When I broke down the U.S. data into seasons, I found that the main sources for each season were different and local persistence could be a major source (Figure 4.3). For example, the main source was EAP for the season 2011-2012, but in the next season, the main source was from U.S. prior season with posterior probability = 1.0, with only some minor introductions from EAP and SAS. Supported local persistence was observed for all studied seasons, where part of the viral sources for the season was from its prior season. The local persistence had supported posterior probability

ranging from 0.32 to 1.0, where two of them generated decisively supported BFs (BF>100) and one had a very strong BF (BF>30). No more than two continuous seasons of local persistence was found, for example, the 2011-2012 viral population circulated in 2012-2013 season but did not circulate in 2013-2014 season or later.

For B-Yamagata lineage (Figure 4.2d and Table 4.2), with the exception of NA, all other regions were identified as sources for the U.S. outbreak, in which ECA and EAP were the main sources with decisive BFs. When divided into seasons, the main sources were still EAP or ECA, except for 2012-2013 season, where local persistence was the main source with posterior probability = 0.98 (BF=520) (Figure 4.3). We also found some local persistence in season 2016-2017 with posterior probability as 0.61 (BF=19).

In sum, for different subtypes of seasonal influenza, the main sources of transmission to U.S. were slightly different, and regions with the highest transmission rate were ECA for A/H1N1 and B-Yamagata, EAP for A/H3N2, and LAC for B-Victoria (Supplemental Figure 4.2). The full transmission matrices with and without specific U.S. seasons for each influenza type can be found in Supplemental Table II-4a-d and Supplemental Table II-5a-d, respectively.

Viral diffusion dynamics among U.S. regions

Data distribution and the phylogenies

The final data to study diffusion dynamics within the U.S. were generated via random sampling scheme by each HHS region with Alaska and Hawaii separately. The distribution of each dataset can be found in Table 4.3. The phylogenies (Supplemental Figure II-3a-d) and estimated substitution rate for each viral subtype were very similar with global dataset to explore viral introductions into the U.S. Similarly, the medians of estimated tMRCAs were about 0.63 –

0.84 years prior to the oldest isolate in the dataset with very narrow 95% BCI (Supplemental Table II-3b). The estimated patterns of effective population size (Supplemental Figure II-4) were close to the global pattern, but B-Yamagata lineage in the U.S. showed more seasonal peaks than the global dataset.

Diffusion dynamics and significant predictors in the U.S.

Phylodynamic modeling with 12 discrete traits (10 HHS regions, Alaska, and Hawaii) was applied to explore the viral diffusion dynamics for each subtype within the U.S. To capture the main diffusion dynamics among these regions out of the complicated dynamics within the U.S., Figure 4.4a-d showed the transmission rates greater than the mean of all the significant rates (BF≥3) or the rates that have a decisive support (BF≥100) for each viral type, while Supplemental Table II-6a-d demonstrated the full matrices of the transmission rates and BF support levels. Furthermore, the separate and joint generalized linear model (GLM) was used to evaluate what ecological and epidemiological factors may affect the diffusion dynamics. The descriptive statistics of these factors can be found in Table 4.4. For the separate GLM for each subtype, one extra model was fit to include sample size of each region to evaluate the impacts of sample size. Results showed that the sample size did not affect the significance of these predictors (Figure 4.5 and Supplemental Figure II-5), which indicated that the subsampled data can represent the viral populations. The effective size and statistical support of each predictor reported below was based on the model with sample size (Figure 4.5 and the left panel of each subtype in Supplemental Figure II-5).

For A/H1N1, in the full transmission matrix, 87 rates out of 132 were statistically supported (Supplemental Table II-6a). HHS regions 4 and 5 were the main transmission hubs to

spread the viruses to other regions with decisive or strong BFs (Figure 4.4a). The two factors that significantly affected the viral diffusions identified by GLM were geographic distance and population over 65 years (Figure 4.5a). The geographic distance had an estimated effect size on a log scale as -0.24 with a decisive statistical support for its inclusion in the model (posterior probability = 0.97 and Bayes Factor = 914). The effect size indicated that viral lineage migration rates to the longest geographic distance were about 0.79 ($e^{-0.24}$) times the rates compared to the shortest distance, after controlling for all other factors in the model. Population of over 65 years at the origin region had an estimated effect size of 0.52 (posterior probability = 1.0 and BF = 281,086) on a log scale, which means higher proportion of the elderly at the origin region facilitated viral diffusion to other regions.

For A/H3N2, in the diffusion matrix, 83 rates out of 132 were statistically supported (Supplemental Table II-6b). HHS region 5 was the main transmission hub to connect all other regions with decisive BFs (Figure 4.4b). Furthermore, region 3 was another main source for region 1, 2, 4 and 5 with decisive BFs. Alaska was a main source for region 5. When checking these predictors, we found more factors that could have impacted A/H3N2 diffusion compared to other viral types (Figure 4.5b). It included geographic distance (conditional effect size on a log scale = -0.21, posterior probability = 0.45, BF=27), total flight number at origin (conditional effect size on a log scale = 0.84, posterior probability=0.84, BF=171), and total flight at destination (conditional effect size on a log scale = 0.18, posterior probability=0.42, BF=24), winter average relative humidity at origin (conditional effect size on a log scale = 0.44, posterior probability=0.17, BF=7), winter average temperature at origin (conditional effect size on a log scale = -0.74, posterior probability=0.99, BF=4,231), peak time rank (conditional effect size on a log scale = -0.27, posterior probability=0.14, BF=6), population of age between 18 and 64 at

origin (conditional effect size on a log scale = -0.42, posterior probability=0.16, BF=6) and population of age over 65 years at origin (conditional effect size on a log scale = 0.51, posterior probability=0.22, BF=9). Among these factors, shorter geographic distance, more flights connections that were at both origin and destination regions, higher winter humidity, lower winter temperature, earlier the epidemic peak time, lower percentage of adult between 18-64 years, and higher percentage of the elderly in the population at the origin region could facilitate the viral diffusions.

For B-Victoria lineage, complicated diffusion dynamics occurred between regions where 75 out of 132 transmission rates were statistically supported, but Alaska was not a source for any other regions (Supplemental Table II-6c). HHS region 5 was identified as the main transmission center showing the highest transmission rates to all other regions with decisive BFs (Figure 4.4c). Region 5 also received viral introductions from many other regions. When GLM was applied to check the ecological and epidemiological factors that may affect these migrations (Figure 4.5c), the significant predictors were geographic distance (conditional effect size on a log scale = -0.40, posterior probability =0.99, BF = 4,493), population density at origin (conditional effect size on a log scale = 0.98, posterior probability = 0.39, BF=21), vaccination rate at origin (conditional effect size on a log scale = -0.55, posterior probability = 0.11, BF = 4), total flight numbers at origin (conditional effect size on a log scale = 0.74, posterior probability = 0.55, BF = 40), winter average temperature at destination (conditional effect size on a log scale = 0.21, posterior probability = 0.20, BF = 8), and the population of age under 18 at origin (conditional effect size on a log scale = 0.65, posterior probability = 0.16, BF = 6). Briefly, higher population density, higher vaccination rate, more total flight connections, higher percentage of youth

population < 18 years in the origin region, closer geographic distance between two regions, and higher winter temperature at the destination may intensify the viral diffusions.

For B-Yamagata lineage, migrations between regions were more complicated than these in B-Victoria lineage. A total of 99 rates out of 132 in the transmission matrix were statistically supported (Supplemental Table II-6d). HHS region 8 was the main source for all other regions with decisive BFs (Figure 4.4c). The GLM (Figure 4.5d) reported significant predictors as geographic distance (conditional effect size on a log scale = -0.26, posterior probability = 0.76, BF = 103), winter average relative humidity at the origin (conditional effect size on a log scale = -0.43, posterior probability = 0.72, BF = 84), and peak time rank at the origin (conditional effect size on a log scale = -0.50, posterior probability = 0.18, BF = 7). Shorter geographic distance, lower winter humidity and earlier epidemic peak time at the origin region may increase viral migrations.

Joint estimation of significant predictors in the U.S.

Assuming seasonal influenza A viruses are transmitted under similar mechanisms regardless of subtype, we conducted a joint GLM analysis where a shared set epidemiological and ecological features could predict the migration patterns observed for each influenza subtype. Results (Figure 4.6) from this joint model highlighted the significant predictors including geographic distance (conditional effect size on a log scale = -0.28, posterior probability = 1, BF = 149,026), total flight numbers at the origin (conditional effect size on a log scale = 0.14, posterior probability = 0.13, BF = 4.5), the population of adults (18 to 64 years) at the origin (conditional effect size on a log scale = -0.38, posterior probability = 1, BF = 149,026), and winter average precipitation at the origin (conditional effect size on a log scale = 0.17, posterior probability = 0.09, BF = 3.1). It indicated that longer geographic distance, lower flight

connections, higher proportion of adults and lower winter precipitation could barrier viral diffusions. In this joint model, the epidemic size was included, compared to the model without epidemic size (Supplemental Figure II-6), significant predictors were reported as the same, but the values of their conditional effect sizes were slightly different.

Discussion

Understanding the viral diffusion pattern is critical for the prevention and control of influenza outbreaks (18). Building upon the knowledge and methodologies learned from previous global studies on seasonal influenza virus diffusion (18–20,46), this study explored the global introductions into the U.S. and inferred the diffusion dynamics inside the U.S. It took the advantage of large and complete (global and U.S. local) genetic and epidemiological datasets of all four types of seasonal influenza viruses. The main findings answered several important questions regarding diffusion patterns and significant epidemiological and ecological predictors. Firstly, global phylogeographic analysis identified that the U.S. seasonal epidemics were seeded with one or several external introductions each season, indicating that the viral diversity in the U.S. was high. Secondly, the viral diffusion inside the U.S. showed that a major transmission hub diffused virus to other regions: Region 5 for A/H3N2 and B-Victoria, Region 4 for A/H1N1, and Region 8 for B-Yamagata. Lastly, important predictors on disease transmission were identified, such as geographic distance, flight connections, winter precipitation and different population age structures.

Diffusion dynamics of seasonal influenza have been studied more thoroughly on the global scale. Though slightly different conclusions have been drawn from these studies (18–20), the importance of reseeding by each season and of no local persistence for A/H3N2 global

circulation has been consistently reported. But for A/H1N1 and B viruses, they could sustain local persistence for multiple seasons, especially in India and China with mean duration > 2 years, and their global dynamics were complicated without clear patterns of reseeding from certain regions (20). In the global region scale, Bedford et al (20) reported that the persisting time in different regions had an average of 6 month for A/H3N2, 9 months for A/H1N1, 13 months for B-Victoria and 12 months for B-Yamagata. Similarly, this study found comparable local persistence in the U.S., that is, A/H3N2 required external introductions by each season but A/H1N1 and B type viruses could persist to next season. This may be related to the viral biology features and patterns of population immunity, for instance, the clear antigenic drift of A/H3N2 and population immunity may be related to the pattern of no local persistence of A/H3N2 (20,199). It may also indirectly reflect that A/H3N2 has higher diversity with multiple viral sources into the U.S. compared to other types.

Viral population dynamics reconstructed from coalescent model could reveal some characteristics of epidemic seasons (200). A/H3N2 had earlier peak time compared to other subtypes and its peak seasons were usually followed by a relatively flat season, which may be related to antigenic drifts and vaccine mismatching for the peak years (84,201). For example, 2011-2012 had a new antigenic variant A/Texas/50/2012 (H3N2)-like; 2014-2015 had very low vaccine effectiveness (only 19%); 2016-2017 had a new antigenic variant A/Singapore/INFIMH-16-0019/2016; and 2017-2018 had A/Kansas/14/2017 (H3N2)-like virus. These variants were recommended as vaccine candidates for later seasons.

Viral migrations between U.S. regions showed one or two primary hubs to diffuse viruses to other regions. A/H1N1, A/H3N2 and B-Victoria had Region 5 (Chicago area) and/or Region 4 (Atlanta area) as the central hub, while B-Yamagata showed the important role of Region 8

(Denver area and its north) to spread viruses. Though geographic distance is a significant barrier for all viral types, flight connections may explain the central hub pattern of viral diffusions. Based on the average flight numbers during flu season, Regions 4 and 5 are among the highest flight numbers. Regions 4, 5 and 8 own the top busiest airports in Atlanta, Chicago and Denver, respectively (202). Though Dallas in region 6 is recognized as another top busy airport, region 6 is not identified as a major hub. The further predictor analysis suggested that climate conditions and population demographic factors could confound the viral diffusion role of flight connections. Local spatial diffusions of all subtypes in the U.S. were decisively determined by the processes that were related to geographic distance, where transmissions were more likely to occur between closer regions. This result is corresponding with what Lemey et al. reported in their global study (18). Furthermore, this study reported that total flight number was positively associated with viral diffusion rate, consistent with studies on global spread and transmission patterns (18). This may indicate that hygiene prevention measures in airports could reduce viral diffusion. The joint GLM estimation also reported that higher winter precipitation may facilitate viral diffusion, which corresponded to the role of precipitation in the previous epidemiological study (203).

The most interesting novel findings with the GLM were population age structure. Results from separate analysis of each subtype were generally consistent with the joint estimates, but the latter provided more robust estimation of the effective size on viral diffusions. The main difference was that in the separate analysis the youth population and the elderly were reported as potential drivers of virus spread for multiple influenza types; but the joint estimation reported that higher proportion of adults in the population was a decisively supported protective predictor. This may be explained by the confounding factors of vaccine effectiveness and relative weight of herd immunity from larger proportion of low risk populations: vaccines may have different

effectiveness for a certain subtype/lineage during a flu season, where if vaccine effectiveness in the high risk populations is relatively lower (either due to low efficacy of vaccination or lower vaccine coverage in the high risk populations) than the herd immunity from low risk populations (i.e. the adults), then the significance of the youth and the elderly on enhancing viral diffusion may be observed for this subtype/lineage. But when joint analysis combines the vaccine effectiveness for all subtypes/linages, the weight of adults becomes significant to prevent viral diffusion. This speculation has been studied in the past (204), but needs further validation by simulation studies. Taken together, the youth and the elderly are the high risk populations for influenza infections (9). Therefore, vaccinating these high-risk populations with effective vaccines and vaccination strategies could provide critical prevention on viral spread and transmissions.

This study had the advantages of using the large and complete dataset from recent years with both viral genetic data and well-recorded metadata. The integrated analysis on viral genetic data and epidemiological data improved the accuracy of viral evolutionary history reconstruction and revealed important predictors on viral spreading (18,98). However, several limitations exist in this study. One universal problem is that sampling biases or missing data could reduce the power of the inferences. For example, some locations had very scare data or no data for a certain year, which limited the inference on the transmission paths. But since samples from several years and multiple locations were combined into one global region, the overall representative of each global region was balanced. To control for sampling biases and to confirm the results, subsampling scheme was repeatedly applied to generate multiple datasets. Another limitation is that predictors were summarized over seasons and over states to represent one HHS region, which reduced the resolution and limited the ability to detect the seasonal or state-level

variations of these predictors on viral spread. Some factors did not have large variations over time, for example, population density, total flight number and the climate data. However, the peak time could be very different in each season and the role of this predictor may not be precisely inferred due to low resolution in this study. The categorization of some predictors may confound the importance of the predictor. For example, the youth population included all population under 18 years old rather than had a specific category of under 5 years old. This may be the reason why the population of under 18 years was only statistically supported for B-Victoria lineage but not found for other types.

Nevertheless, this study identified major global sources for different types of seasonal influenza viruses, which may indirectly indicate that A/H3N2 has higher genetic diversity from various external introductions each season. Within the U.S., complicated diffusion patterns were observed, but different primary diffusion hub was identified for each viral type. Furthermore, geographic distance, total flight connections, winter average precipitation and the population age structure can significantly affect the diffusion rates. Hygiene intervention inside flights or heavy traffic settings may help to prevent viral diffusion during flu season. Furthermore, developing highly effective vaccines and administration strategies for the high-risk populations including the youth and the elderly could be a major approach for interrupting viral spread and transmission.

Tables and Figures

Table 4.1. Distribution of global datasets by regions and by influenza types after subsampling via random sampling strategy.

Regions	A/H1N1	A/H3N2	B-Victoria	B-Yamagata
AF	180	170	144	152
EAP	230	300	200	220
ECA	270	320	190	230
LAC	210	210	160	161
MENA	170	180	103	127
NA	148	210	80	89
SAS	170	160	117	132
USA	340	380	347	350
Total	**1,718**	**1,930**	**1,341**	**1,461**

Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

Table 4.2. Count of introductions, migration rate and 95% BCI from each global region into the U.S. for each influenza type. The three items in each matrix cell in order are count of introductions, migration rate, and its 95% Bayesian Credible Interval (95% BCI). The rows of the table are the donor region and the columns are the recipient region. The cell color indicates Bayes Factor (BF) levels, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

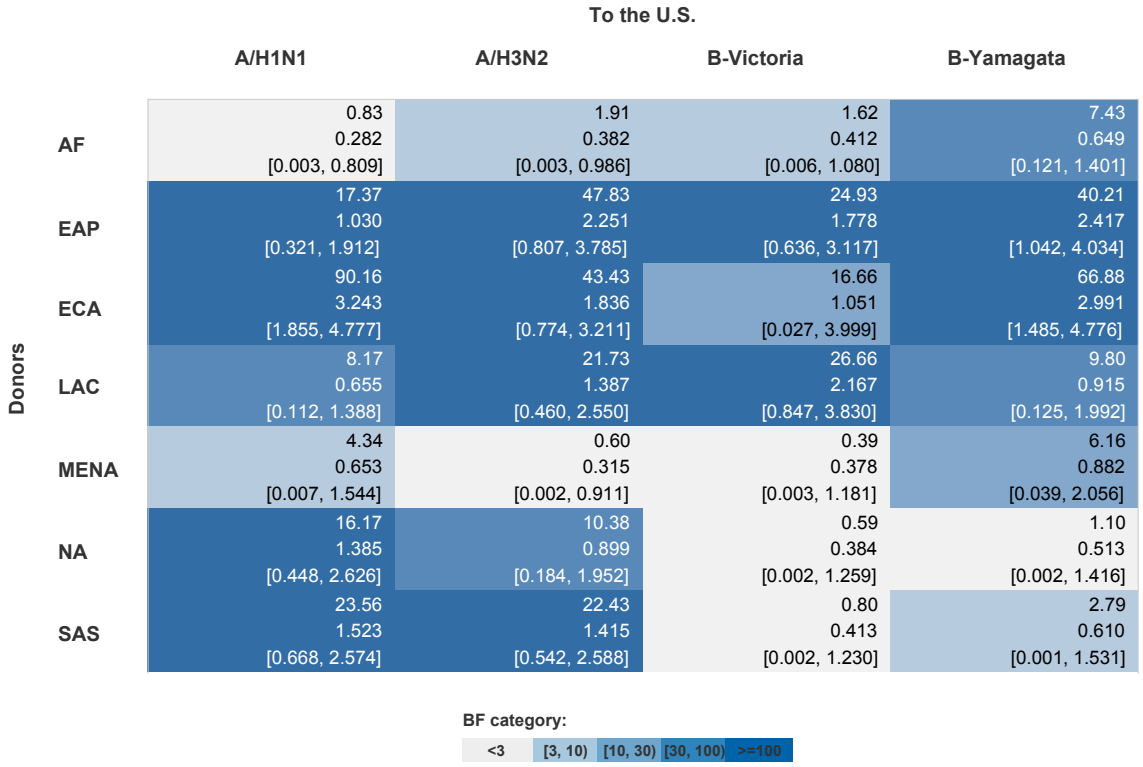


Table 4.3. Distribution of U.S. region dataset by locations and by influenza types after random sampling strategy.

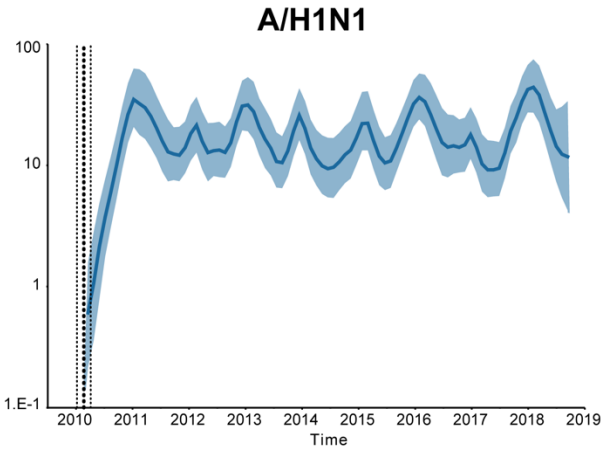
USA-region	A/H1N1	A/H3N2	B-Victoria	B-Yamagata
Region 1	90	110	80	90
Region 2	100	110	76	80
Region 3	100	120	80	90
Region 4	110	120	90	100
Region 5	110	140	90	100
Region 6	100	120	80	90
Region 7	80	100	65	80
Region 8	100	110	80	90
Region 9	100	120	80	90
Region 10	90	110	80	90
Hawaii	80	90	70	80
Alaska	68	90	20	65
Total	**1,128**	**1,340**	**891**	**1,045**

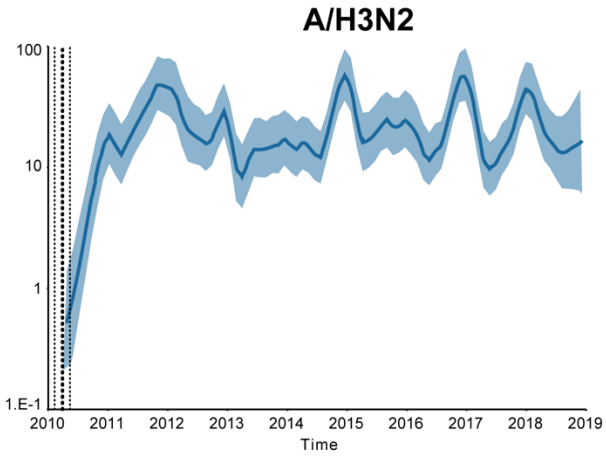
Table 4.4. Geographic, climate and demographic characteristics of the 10 Health and Human Services regions, Hawaii, and Alaska.

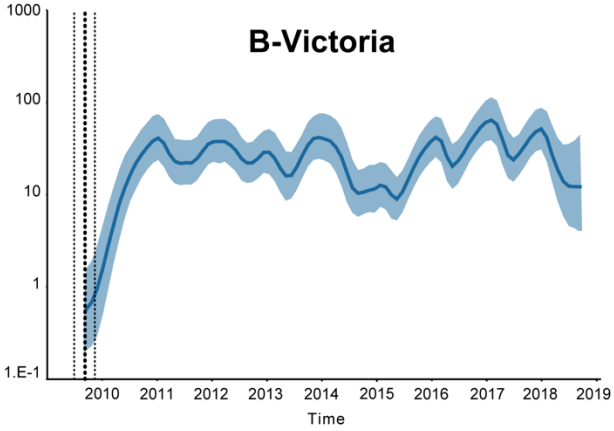
Predictors	Median	Mean	SD*	Minimum	Maximum
Geographic distance (km)	3014.0	2402.2	1993.5	426.7	8134.9
Population Density (per mile^2)	138.0	150.0	135.8	1.2	493.7
Vaccination Rate (%)	45.4	45.6	3.4	39.6	50.9
Total flights during flu season	548,396	709,754	597,187	48,094	1,790,528
Winter relative humidity (%)	76.5	74.1	8.8	54.4	84.5
Winter average precipitation (inch)	2.8	2.6	1.1	0.9	4.2
Winter average temperature (F)	32.8	36.9	16.0	10.0	75.5
Epidemic peak delay weeks	3.9	4.3	0.9	3.3	5.9
Age under 18 years (%)	23.9	23.8	1.9	20.7	26.4
Age between 18-64 years (%)	63.3	62.7	2.5	55.5	65.9
Age over 65 years (%)	13.4	13.5	3.7	7.7	23.7

SD*: Standard deviation.

Figure 4.1. Estimated effective population size by Skyride coalescent model for global samples. The x axis represents time in years and the y axis represents viral effective population size on a log scale. The bolder dashed vertical line indicates the estimated most recent common ancestor time (tMRCA) for the viral population. The thinner dashed vertical lines are the 95% BCI for the tMRCA. The horizontal blue line and blue shadow indicate the median and 95% BCI bands of estimated effective population size, respectively.


A/H1N1
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time


A/H3N2
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time


B-Victoria
1000
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time

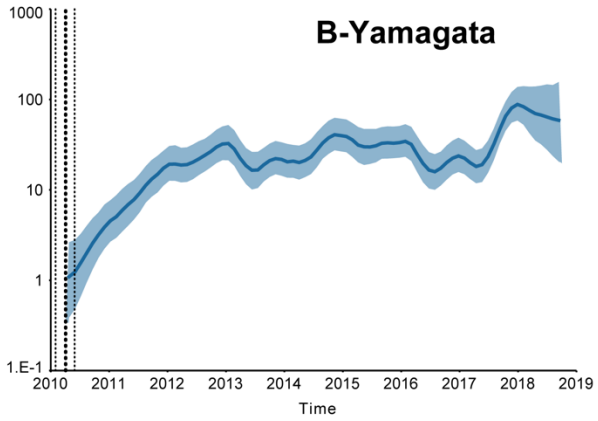

B-Yamagata
1000
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time

Figure 4.2. Global introductions into the U.S. for each influenza type. The stroke weights of the arrows are proportional to the values of the transmission rates. Only statistically significant rates (BF>3) from global regions to the U.S. were shown. The main sources were slightly different for different subtypes of seasonal influenza. The source region with highest migration rate was ECA for A/H1N1 and B-Yamagata, EAP for A/H3N2, and LAC for B-Victoria.

Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

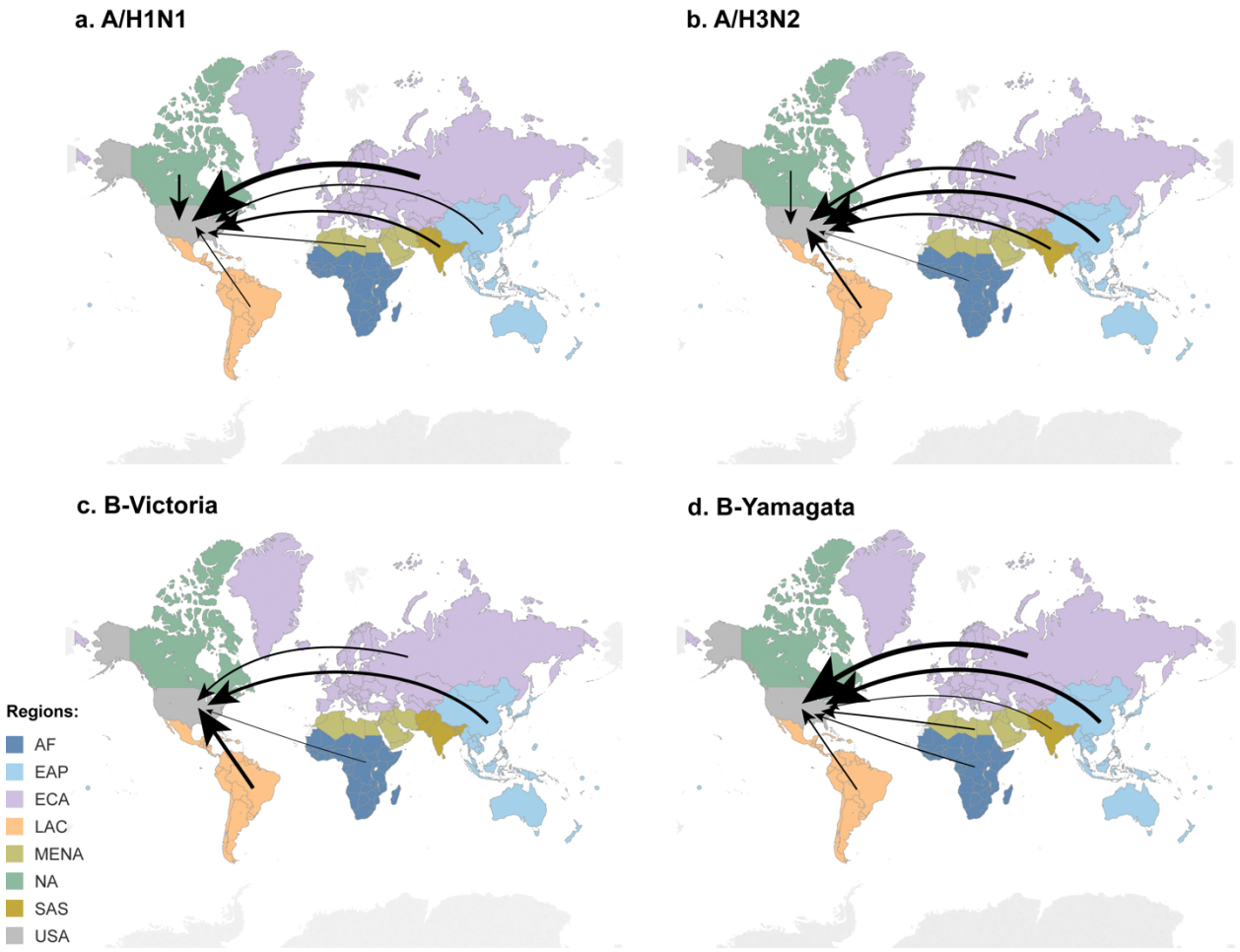

a. A/H1N1
b. A/H3N2
c. B-Victoria
d. B-Yamagata
Regions:
AF
EAP
ECA
LAC
MENA
NA
SAS
USA

Figure 4.3. Components of viral sources for each U.S. season. Each viral type had one or two different main sources for each season. A/H3N2 required external introductions for each season but A/H1N1 and B viruses could have local persistence. Introduction counts represent the counts of trait state transitions in the reconstructed phylogeny.

Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

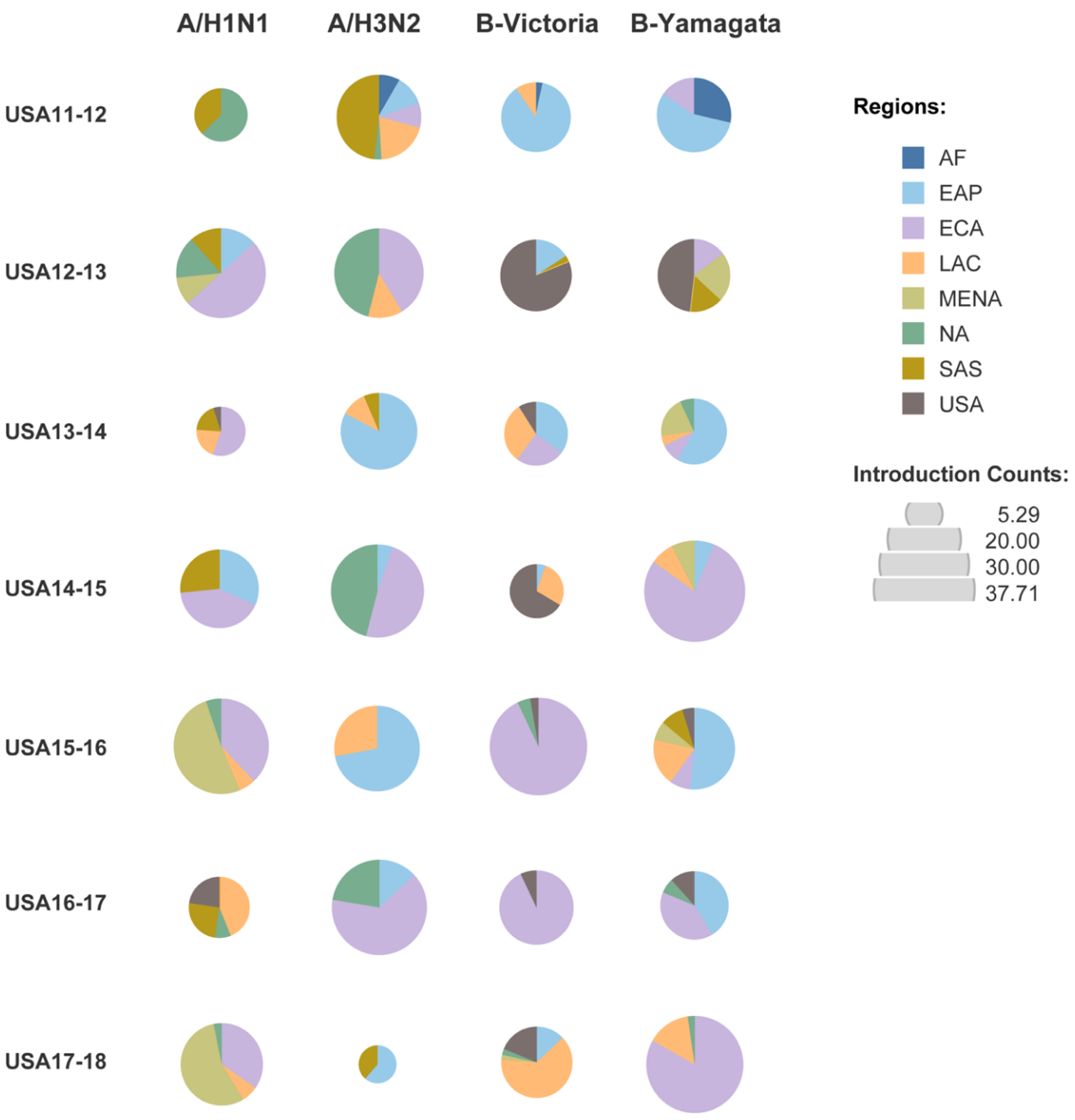
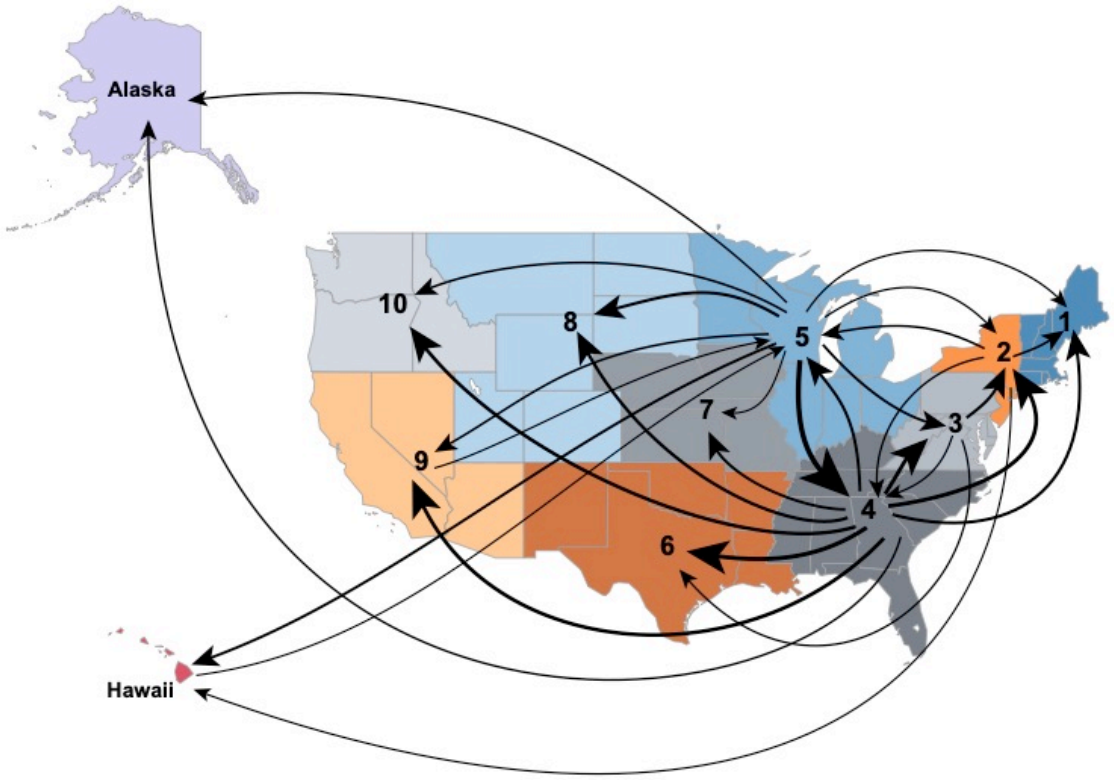
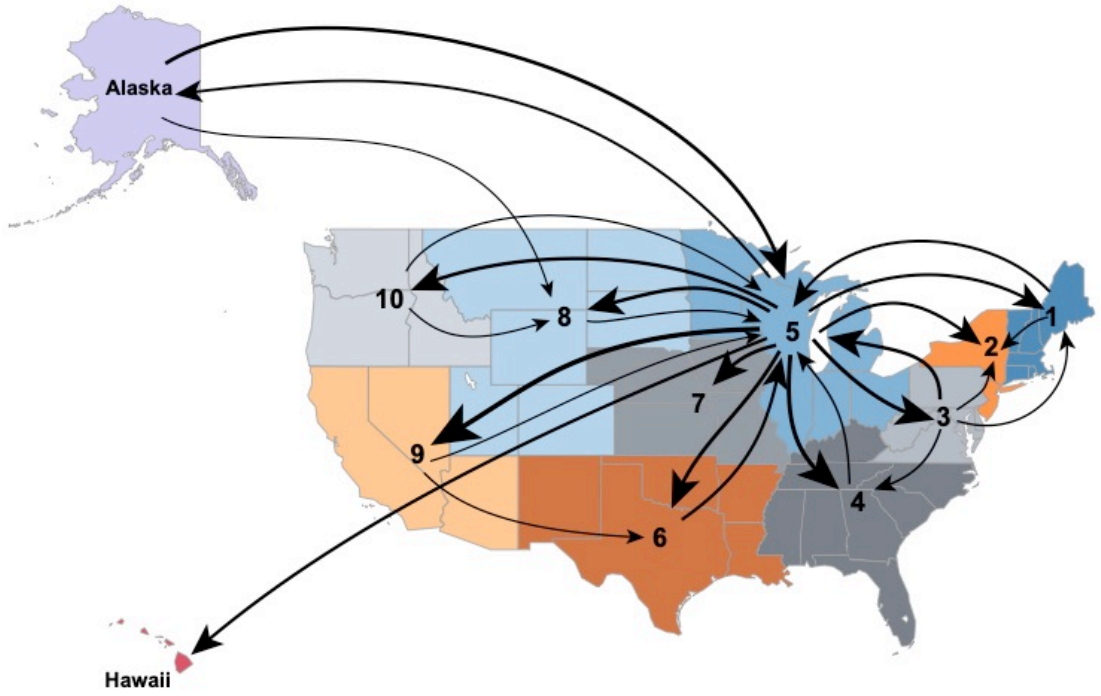

A/H1N1
A/H3N2
B-Victoria
B-Yamagata
USA11-12
USA12-13
USA13-14
USA14-15
USA15-16
USA16-17
USA17-18
Regions:
AF
EAP
ECA
LAC
MENA
NA
SAS
USA
Introduction Counts:
5.29
20.00
30.00
37.71

Figure 4.4. Major diffusion dynamics within U.S. regions for each influenza type. For each viral type, the arrows are only showing the migration rates greater than the mean of all the significant rates (BF≥3) or the rates that have a decisive support (BF≥100). The stroke weights of the arrows are proportional to the values of migration rates. Region 4 and/or Region 5 have been identified as the primary transition hub for A/H1N1, A/H3N2 and B-Victoria, but Region 8 is recognized as the major hub for B-Yamagata.

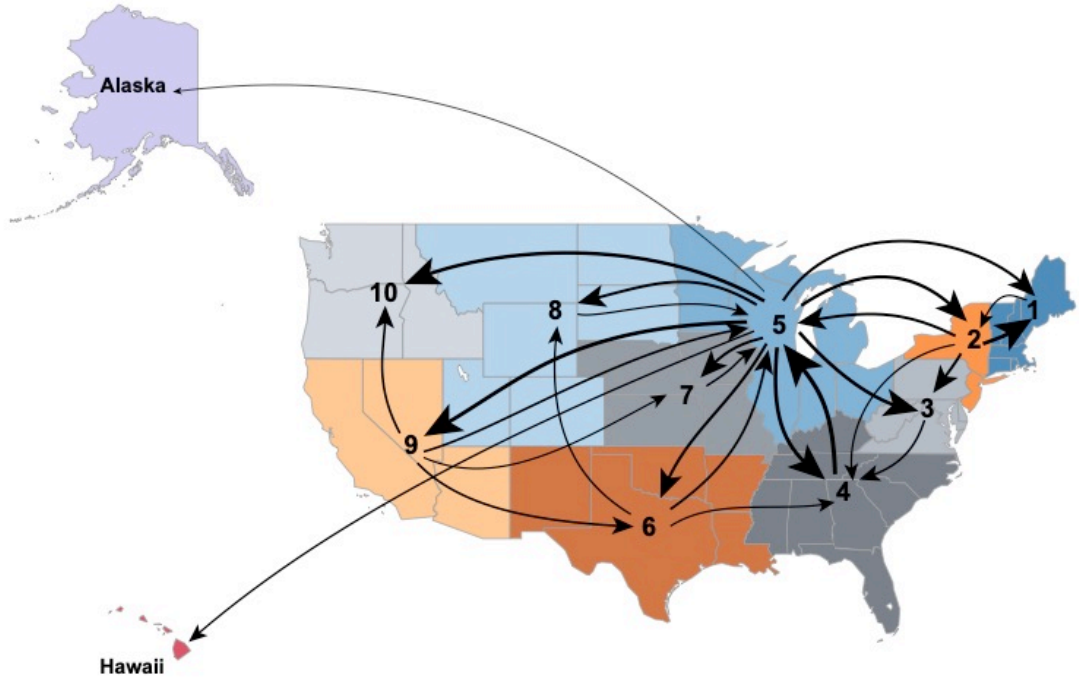
4.4a. A/H1N1



4.4b. A/H3N2



4.4c. B-Victoria



4.4d. B-Yamagata

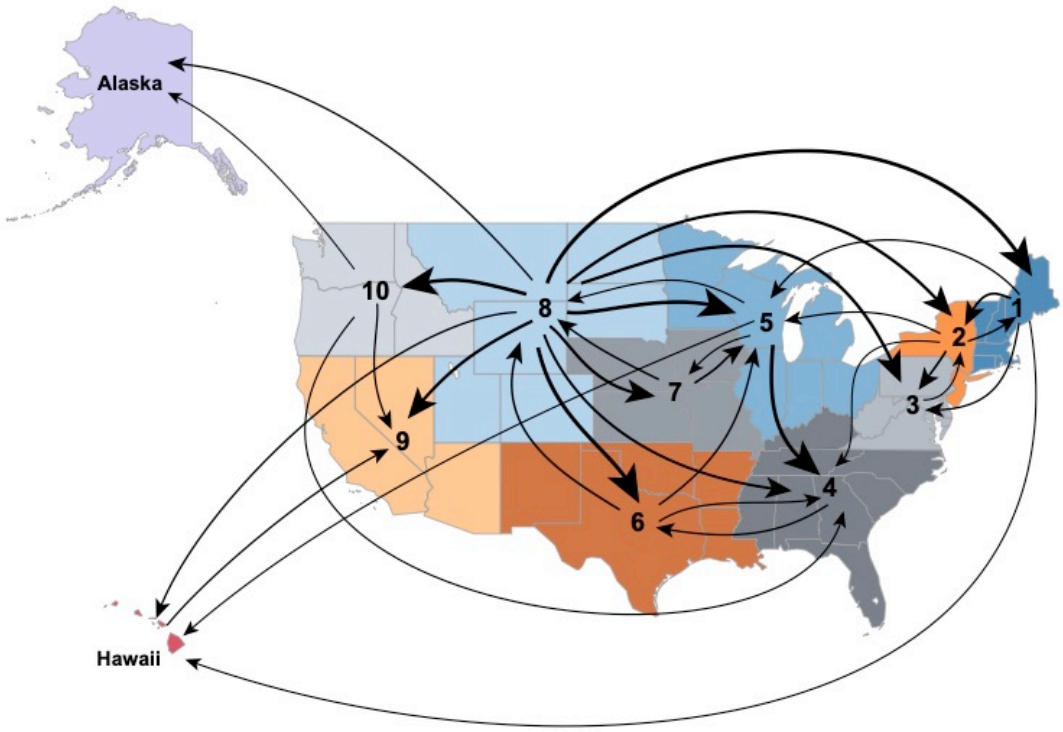


Figure 4.5. Generalized linear model with sample size reported significant predictors to affect diffusion dynamics for each influenza type in the U.S. The conditional effect size represented by the blue dot is the median of the coefficient when this predictor is included in the model. The bar on the blue dot indicates the 95% Bayesian Credible Interval (BCI). The posterior probability is the probability of this factor being included in the model, which is used to calculate Bayes Factor (BF). The dash vertical blue line indicates BF=3, where if the posterior probability is beyond the dashed line, it represents this indicator has statistical support to be included in the model. The solid vertical blue line represents BF=100, where if the posterior probability is beyond the solid line, then the indicator has decisively statistical support to be included in the model. Black dot with "na" indicates the conditional effect size is not available since the predictor is never included in the model during the simulation process. _O: geographic region as origin location; _D: geographic region as destination location. Abbreviation: PopDens – population density; VaccRate: average vaccine rate. For all subtypes, the sample size did not affect the significance of these predictors.

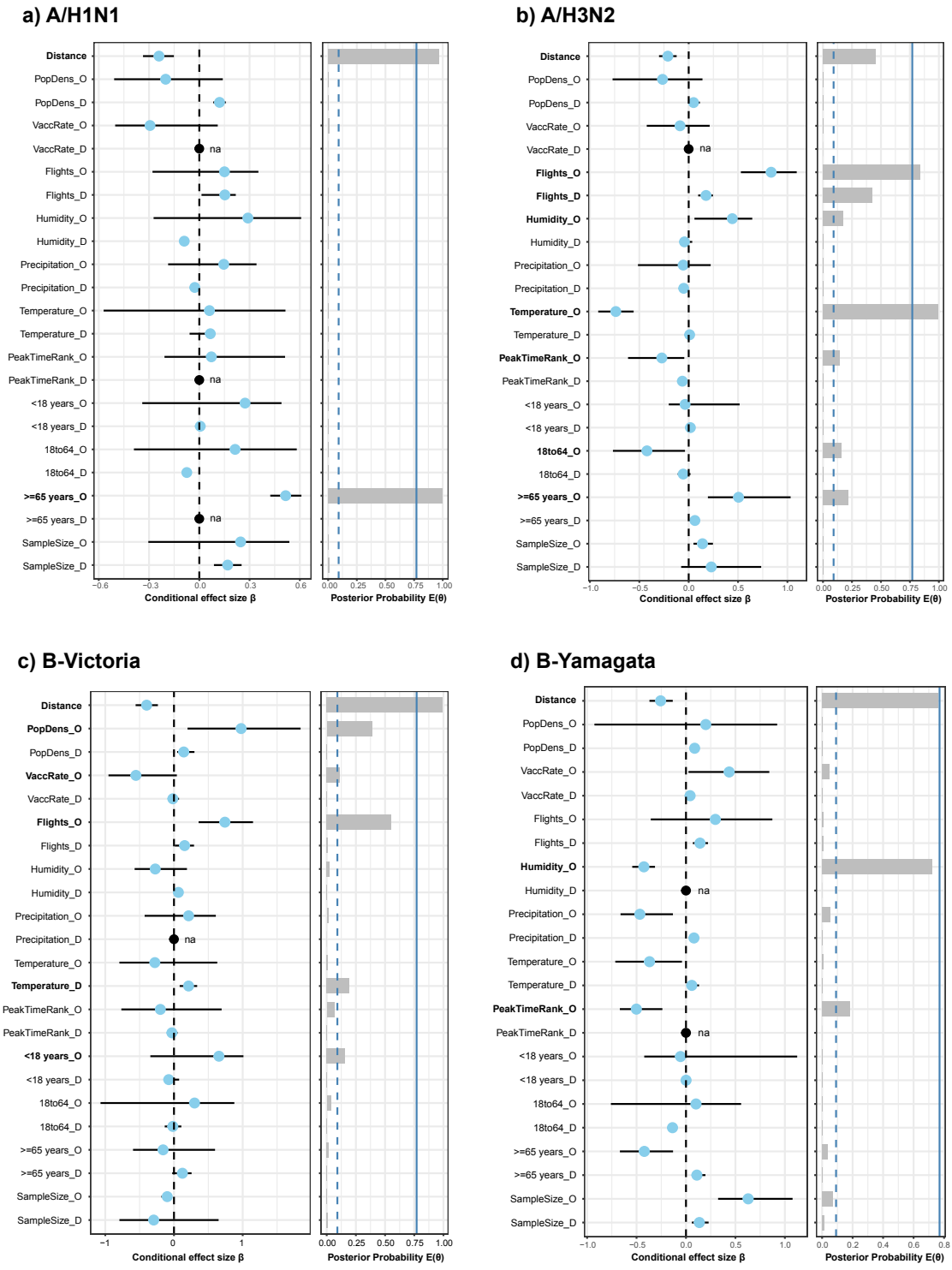
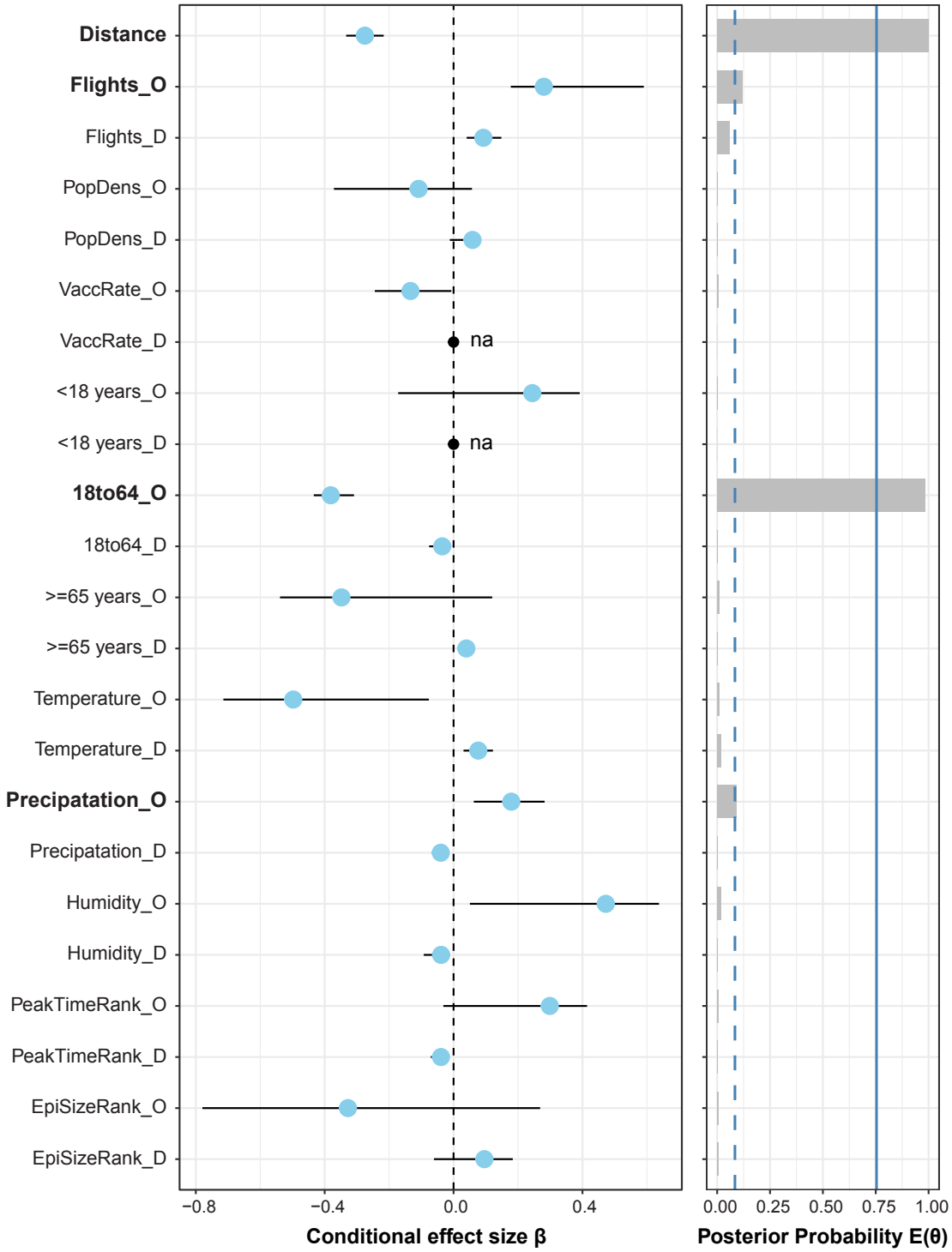

a) A/H1N1
Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
Conditional effect size β
Posterior Probability E(θ)
b) A/H3N2
Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
Conditional effect size β
Posterior Probability E(θ)
c) B-Victoria
Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
Conditional effect size β
Posterior Probability E(θ)
d) B-Yamagata
Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
Conditional effect size β
Posterior Probability E(θ)

Figure 4.6. Joint generalized linear model with epidemic size reported significant predictors to affect diffusion dynamics in the U.S. The joint estimation unified four transmission rate matrices of each influenza type into one model to estimate the joint effect size of the predictor. With epidemic size being included in the model, the same significant predictors were reported but the impacts of these predictors were slightly different compared to Supplemental Figure II-6. The conditional effect size represented by the blue dot is the median of the coefficient when this predictor is included in the model. The bar on the blue dot indicates the 95% Bayesian Credible Interval (BCI). The posterior probability is the probability of this factor being included in the model, which is used to calculate Bayes Factor (BF). The dash vertical blue line indicates BF=3, where if the posterior probability is beyond the dashed line, it represents this indicator has a statistical support to be included in the model. The solid vertical blue line represents BF=100, where if the posterior probability is beyond the solid line, then the indicator has a decisively statistical support to be included in the model. Black dot with "na" indicates the conditional effect size is not available since the predictor is never included in the model during the simulation process. _O: geographic region as origin location; _D: geographic region as destination location. Abbreviation: PopDens – population density; VaccRate: average vaccine rate.


Distance
Flights_O
Flights_D
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
Temperature_O
Temperature_D
Precipatation_O
Precipatation_D
Humidity_O
Humidity_D
PeakTimeRank_O
PeakTimeRank_D
EpiSizeRank_O
EpiSizeRank_D
na
na
−0.8
−0.4
0.0
0.4
Conditional effect size β
0.00
0.25
0.50
0.75
1.00
Posterior Probability E(θ)

CHAPTER 5

EVALUATE THE IMPACTS OF H3N2 LAIV ON VIRAL DIVERSITY AND DIFFUSION DYNAMICS [3]

[3] Qiu, X., Avadhanula, V., Fabrizio, T., et al. To be submitted to *Influenza And Other Respiratory Viruses*, 11/2019.

Abstract

Since the first introduction of the live attenuated influenza vaccines (LAIV) in 2003 in the U.S., safety and efficacy studies have been conducted continuously. However, studies that explore the potential impacts of LAIV on viral population are rare. This study took the advantage of data availability from the Texas Central Trial on the control of epidemic influenza via largely vaccinating school-aged children in the community during 2004-2006. Advanced phylodynamic modeling, including discrete trait analysis and structured coalescent analysis, have been conducted to quantitatively evaluate how vaccination could impact the viral genetic diversity and transmission dynamics. Results from this study indicated that the vaccinated population had higher genetic diversity, needed more external introductions to sustain the epidemics, and reduced viral dissemination resulting in lower transmissions to external local regions. Taken together, vaccination could impact viral diffusion in a beneficial way to interrupt epidemic chains, probably due to the immune landscape change in the susceptible and high-risk population.

Introduction

Seasonal influenza is a fast spreading disease that commonly affects the respiratory tract and can cause large epidemics across communities in the globe (1,2). The symptoms may vary from mild to severe respiratory diseases with potential life-threatening outcomes (4,6). The disease burden associated with seasonal influenza is very high across the globe. Every year, influenza leads to respiratory tract infections in 5–15% of the global population, severe illness in 3–5 million individuals (2), and seasonal influenza-associated respiratory deaths in 0.3–0.65 million people (10). Based on the systematic analysis of seasonal influenza in 195 countries during 1990-2016 , the incidence of influenza associated with lower respiratory infections was 5.3 cases per 1,000 persons in all age groups, 9.1 cases per 1,000 persons in children younger than 5 years, and 15.8 cases in adults older than 70 years, respectively (9). The Center for Disease Control and Prevention (CDC) has estimated that from 2010-2011 through 2017-2018 influenza seasons in the United States (U.S.): influenza has annually caused 9.3-49 million illnesses, 140,000-960,000 hospitalizations, and 12,000-79,000 deaths, where heavier outbreaks usually are associated with mismatching influenza vaccines (190).

Influenza viruses can spread rapidly due to the ease of transmission, usually via contact, fomites, droplets, or aerosols (14). Environmental and ecological factors can affect the size, stability, and inhalation of both droplets and aerosols (15–18). Such factors include temperature, humidity, population structure and density, and transportation connections, which may affect the seeding and geographic diffusion of influenza viruses (15–18). Two other important reasons for seasonal outbreaks and rapid spread of influenza viruses are the rapid viral genetic and antigenic changes, where population level immune protection is limited to prevent infections from high diverse viruses (23,24). The most common mechanisms of genetic and antigenic changes in

seasonal influenza viruses are high point mutation rates, immune selection, and viral population migration (205).

Vaccines are amongst the most cost-effective prevention measures for many infectious diseases (67). In the U.S., the development and use of influenza vaccines has started in the 30's (69) after Influenza types A and B were isolated in 1933 and 1936, respectively (68). With the discovery that influenza viruses could grow in embryonic chicken eggs, the first inactivated influenza vaccine was developed in 1938 and administered to the US soldiers during World War II, which ultimately made no difference on clinical outcomes in vaccinated and unvaccinated populations (69). Early influenza vaccines only contained inactivated type A (monovalent) but became a bivalent vaccine in 1942 with both types A and B (69). The protective efficacy of these inactivated vaccines was confirmed in the 1950s through surveillance studies and continuous efforts on vaccine development (68,69). In 1978, the first trivalent vaccine containing two type A strains and one type B strain was developed (71–73). These were split or subunit vaccines, where chemically or physically inactivated virions treated with detergent results in split vaccines and further purification of the haemagglutinin (HA) and neuraminidase (NA) of these viruses leads to subunit vaccines (71,206). The first live attenuated influenza vaccine (LAIV), called "FluMist®", was authorized by the U.S. Food and Drug Administration (FDA) in 2003, which is intranasally administered in healthy population of 5-49 years old (74). In the following years, the updated vaccines were recommended for a broader age range, including infants. In 2012, the first quadrivalent LAIV called "Fluarix®" was approved by the FDA, which contains two A subtypes (A/H1N1 and A/H3N2) and two B lineages (B/Yamagata-like and B/Victoria-like) (75).

Since the introduction of influenza vaccines into human populations, the efficacy and safety of these are one of the main concerns in the healthcare and research communities

(146,147). But evidence has shown that vaccines may impact the genetic diversity and distributions of the pathogen populations (148,149). These may be due to vaccination-induced adaptive immune responses and selection pressures on the pathogen (150). Imperfect vaccine design or vaccination procedure may also drive the evolution of pathogens (151). Additionally, the vaccine covered serotypes of the pathogen can modify the competitive ecological constraints that allow non-vaccine types to be dominant, which has been well studied in *Pneumoniae Streptococcus* with the introduction of pneumococcal conjugate vaccine (PCV) (149). Furthermore, influenza vaccines can change the host immune landscape with reducing the availability of susceptible host, which may drive viral evolution. This is especially true for seasonal influenza viruses that have high antigenic diversity, showing a ladder shape phylogeny with strong immune selection (207), (100). However, the impact of vaccines on circulating influenza viral genetic diversity and disease diffusion within and between communities have not been studied, often due to limited availability of appropriate viral genetic dataset.

The Central Texas Trial on the Control of Epidemic Influenza (referred to as "the Central Texas Trial") during 1998-2010 (208–213) has provided the opportunity to retrospectively study the impacts of the first introduction of LAIV on viral diversity and diffusion dynamics. The initial purposes of this study aimed to improve the influenza vaccination coverage in school-aged children, evaluate the direct effectiveness and herd immunity provided by vaccines, and assess the safety of different types of vaccines, especially the first introduction of LAIV in 2003. The intervention was a non-randomized and community-based study. LAIV vaccine was applied to school children (age 5-18 years) in two cities Temple and Belton (TB-v) in Texas as the experimental arm, while two cities, Waco and Bryan College Station (WBC-uv), were selected as the control arm without LAIV vaccination. This study collected influenza positive samples

and complete records of epidemiological data including population structure, population size, vaccine coverage, influenza incidence.

Recent years, phylodynamic modeling approaches have allowed both genetic data and epidemiological data to be integrated into a unified statistical framework to understand viral diffusion dynamics with discrete trait analysis (98). Another advanced approach to improve the accuracy of ancestral reconstruction is to distinguish the mutation and migration events in the coalescent procedure, which is not considered in most of the current coalescent models (112,114,214). Mutation events are naturally related to viral genetics, but migration events are related to host population structure. To incorporate the heterogeneity of host populations, a structured coalescent model can provide important insights into local scale disease dynamics including how population vaccine status or migration dynamics between populations might impact disease transmission and diffusion dynamics (114,115).

In the current study I aimed to apply these advanced tools to understand the impacts of LAIV on viral genetic diversity and diffusion dynamics with the available dataset from the Central Texas Trial. I first employed phylodynamic modeling with discrete trait analysis to understand the viral sources of Texas local dynamics and evaluate whether vaccination caused differences in local transmission dynamics and viral genetic diversity. The tested hypothesis is that increased vaccination rate will lower local persistence and lineage transmission where the epidemic will require more external introductions to be sustained. Secondly, I assessed the impacts of vaccination and local transmissions between vaccinated and unvaccinated populations with the structured coalescent framework by testing the hypothesis that pathogen migration rates from unvaccinated population to vaccinated population will be higher.

Methods

Datasets preparation

Data in this study included publicly available global H3N2 HA nucleotide sequences during 2003-2007 downloaded from the Global Initiative on Sharing All Influenza Data (GISAID, https://www.gisaid.org/), and H3N2 isolates from the Central Texas Trial during 2004-2006. Global data from two extra seasons (one season before and one season after the vaccination intervention in Texas) have been included to perform spatial diffusion analysis between the global and the local datasets. The following inclusion and exclusion criteria were used for the global data download: a) all sequences must include collection date (at least year information) and geographic location with country information; b) the minimum length of the sequences should be more than 50% of the full gene length; c) all vaccine, derivative, recombinant and laboratory sequences are excluded; d) duplicate sequences with the exact same location and 100% nucleotide similarity are excluded with the oldest dated strain retained.

Two main datasets were generated (data management process was detailed below). One was the full dataset containing 484 HA sequences of both the global and Central Texas Trial samples to study the dynamics between Texas local cities and the global regions. Another dataset was the subsampled dataset from the full dataset with 188 isolates containing Central Texas Trial samples sequences with their most closely genetic-related global sequences to further perform the structured coalescent analysis and explore the migration rates between vaccinated and unvaccinated populations.

Samples from the Central Texas Trial

The vaccination intervention in central Texas was conducted when the first LAIV was approved for use in the 5-49 year old population in 2003 in the U.S. (74). Active surveillance

was conducted during the whole flu season. Despite the trial of LAIV lasted for 3 years during 2003-2006, only the data between 2004-2006 were used because the planned 2003-2004 intervention had not been implemented due to the earlier epidemic than regular seasons. From the two-year surveillance (that is, 2004-2005 and 2005-2006 flu seasons), a total of 781 influenza A positive samples were collected and 81 H3N2 full genomes (about 10% of all positive samples) were randomly selected and sequenced to balance the sample representative and funding allocation.

NGS library preparation, sequencing and assembly

The 81 H3N2 positive samples were prepared for next-generation whole-genome sequencing (NGS). Viral RNA was extracted using the MagMAX-96 AI/ND Viral RNA Isolation Kit (Applied Biosystems AM1835) on the KingFisher Flex Magnetic Particle Processor (ThermoFisher). The RNA was then reverse transcribed using SuperScript III Reverse Transcriptase (Invitrogen 18080044) per the included protocol and using the universal influenza A primer, Uni12. The influenza genome was then amplified in a multisegment PCR protocol using Uni12/13 primers in addition to universal influenza A polymerase segment primers (primer sequences available upon request) with Phusion High-Fidelity PCR Master Mix with HF Buffer (NEB M0531S). The temperature cycle parameters were 98℃ for 30 s, with 10 cycles (98℃ for 10 s, 45℃ for 30 s, and 72℃ for 2 min), followed by 20 cycles (98℃ for 10 s, 56℃ for 30 s, and 72℃ for 2 min) and a final elongation step of 72℃ for 10 min. The PCR products were purified using the QIAquick 96 PCR Purification Kit (Qiagen 28181) and sample QC was assessed on the Agilent Bioanalyzer and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen P7589). The next generation sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) with 96 dual index barcodes according to the manufacturer's

protocol. The normalized libraries were pooled and run on the Illumina MiSeq platform using paired-end 150 base-pair cycling conditions. These paired reads were assembled with Iterative Refinement Meta-Assembler (IRMA), which was developed specifically for highly variable RNA viruses with more robust assembly and variant calling (215). IRMA v0.6.7 (https://wonder.cdc.gov/amd/flu/irma/) was used with its embedded influenza assembly module. From the 81 sequenced isolates, only 79 HA gene were generated, since two of the samples did not have enough read coverage on the HA gene.

Data alignment and inspection

Sequence alignment of the data used in this study (H3N2 HA sequences from the global and Central Texas Trial) was performed with the software MUSCLE v3.8.31(191). Manual alignment was performed to manage unreasonable indels (insertions/deletions). Primary phylogenetic analysis was conducted with the maximum-likelihood (ML) approach in Randomized Axelerated Maximum Likelihood (RAxML) v8.0 (192), which has the advantage of efficiently handling large datasets. Temporal signal of the ML tree was examined in TempEst (193) with the root-to-tip divergence regression and outliers were inspected in the distribution of residuals of the samples.

Global data spatial distribution and subsampling

After alignment and outlier removal, spatial distribution of the data was conducted. To reduce the dimension of the geographic discrete traits and have enough samples in each geographic trait, global region instead of country was used as the geographic location unit to explore the viral dynamics between the global and Texas data. The global H3N2 HA sequences during 2003-2007 were categorized into seven regions defined by the World Bank (194): East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC),

Middle East and North Africa (MENA), North America (NA, excluding the US), South Asia (SAS), and Sub-Saharan Africa (AF). The U.S. data was separately categorized to examine the potential distinctive roles of other global regions and the U.S., of which the latter was considered as a domestic location on viral diffusion into Texas. Sequences with 100% similarity in one region were removed with retaining the oldest isolate to avoid the weight of the duplicated sequences being overrepresented in the phylodynamic source-sink model, resulting in a full dataset with 4,478 sequences. A subsampling strategy was used to fully capture the viral diffusions into Texas local cities, that is, ML trees were generated with all the isolates to only obtain the global viruses that were phylogenetically clustered with the Central Texas Trial samples. The subsampled global samples with this approach (n= 423) contained 152 from EAP, 90 from ECA, 50 from LAC, 3 from MENA, 3 from SAS, and 125 from the U.S.

Central Texas Trial data subsampling

After outliers were identified by TempEst in the Central Texas Trial samples, the remained 76 isolates were designated as vaccinated population (TB-v) and unvaccinated population (WBC-uv). With removing the duplicated sequences in each population, there were 40 TB-v samples and 29 WBC-uv samples. To control the over-representation of TB-v samples in 2005-2006 season compared to WBC-uv (24 versus 11 samples), 8 sequences in TB-v group were randomly removed. Together with global data, the final dataset contained 484 HA sequences which was referred to as “full dataset”.

To further study the impacts of vaccination and reduce the dimensions of the geographic discrete traits, a smaller dataset referred to as “subsampled dataset” containing Central Texas Trial samples and their genetic closely related global sequences was prepared to perform the structured coalescent analysis to explore the migration rates between these populations. This

dataset contained 188 HA sequences, where they were categorized into three groups: 127 sequences as “Global” unvaccinated population (including 37 sequences from EAP, 26 from ECA, 4 from LAC, 1 from MENA, and 59 from the U.S.), 32 sequences as “TB-v” to represent Texas vaccinated population, and 29 sequences as “WBC-uv” to represent Texas unvaccinated population. To further quantify local dynamics, only Central Texas Trial data with two groups “TB-v” for vaccinated population and “WBC-uv” for unvaccinated population was used to conduct the structured coalescent analysis.

Bayesian phylodynamic modeling

Diffusion dynamics with source-sink model

To understand the viral sources for the epidemics in Texas local cities, source-sink phylodynamic modeling with discrete geographic locations for both full dataset and subsampled dataset was conducted in Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.10.3 (160). The well-supported and parameter-rich substitution model general time reversible (GTR) model (93,94) was applied with gamma-distributed rate variations among sites. The lognormal relaxed molecular clock (98) was used with an initial mean of 0.0033 with a uniform prior ranging from 0.0 to 1.0. Based on the prior knowledge, a smooth and time-aware Gaussian Markov random field (GMRF) process prior on the population sizes was applied in the Skyride coalescent model (161).

With dated samples and tip-associated trait characters, the continuous-time Markov chain (CTMC) model of discrete traits was applied to infer how the trait has evolved with the viral population since the sampling time (98,109). In addition, the number of state changes across the phylogeny for each state was estimated at the tree internal nodes denoting a state transition event

and referred to as Markov jump. These Markov jumps were counted along each phylogenetic tree sampled (196) to represent the overall transitions between regions during the sampling timeframe. An asymmetric CTMC matrix model was used to estimate the transition rates and absolute Markov jump counts for both directions between two trait characters. The Bayesian stochastic search for variable selection (BSSVS) approach was employed to provide the most parsimonious diffusion process and further compute the statistical support of transmissions between two trait characters (98,109).

Six independent MCMC chains of 100 million generations were simulated with sampling step every 10,000 generations to yield 10,000 trees per run. Convergence of 6 runs was diagnosed in Tracer v 1.7.1 (http://tree.bio.ed.ac.uk/software/tracer/) for all parameters to ensure a sufficient effective sample size (ESS > 200). LogCombiner v1.10.3 as part of the BEAST software package was used to combine the multiple runs to generate log and tree files after appropriate removal of the burn-in from each MCMC chain to guarantee convergence of these runs. The Maximum Clade Credibility (MCC) tree was summarized in TreeAnnotator v1.10.3 from the combined tree file. The MCC tree was visualized in FigTree v1.4.4 (https://github.com/rambaut/figtree/releases), where the posterior probability for each node (>0.50) and the 95% Bayesian credible intervals (BCI) of node age were displayed as indicators of phylogenetic estimation uncertainty. All xml files for these runs can be found here https://github.com/XuetingQiu/LAIV_impacts.

With BSSVS, Bayes factor (BF) was calculated for each transition rate with the probability of non-zero indicators and the prior probability (18,142). The value of BF reflects whether the transition rate is statistically important. The strength of statistical supports was interpreted with these criteria: $BF < 3$ indicates no significance; $3 \leq BF < 10$ indicates substantial

support, $10 \leq BF < 30$ indicates strong support, $30 \leq BF < 100$ indicates very strong support, and $BF \geq 100$ implies decisively statistical support (106). BF calculation was done by the program SpreaD3 (197) with the combined log file. Transmission rates and jump counts were extracted and calculated by python scripts with combined log file (python scripts can be found https://github.com/XuetingQiu/LAIV_impacts).

Structured coalescent

To infer the impacts of vaccination, structured coalescent was conducted to "subsampled dataset" containing three demes: Global, WBC-uv, TB-v; and also to Texas only dataset with two demes: WBC-uv and TB-v. The structured coalescent model was performed via the MultiTypeTree module (214) of BEAST v2.4.3 (216). This model can estimate migration rates between demes and effective population size for each deme. The GTR substitution model with Gamma invariant distribution was specified (93). The uncorrelated relaxed molecular clock with a log-normal distribution (98) was used to infer the evolutionary rate from dated samples. All parameters were set as the default priors, but the population size was set as a log-normal distribution and migration rate between demes was set as an exponential prior with a mean of 1.0 (115). Due to the complexity of MultiType tree parameters, twelve independent MCMC chains of 200 million generations were simulated with sampling every 20,000 steps. Convergence was assessed in Tracer v1.7.1 and proper burn-in was removed to guarantee good convergence (ESS >200). LogCombiner in BEAST v2.4.3 (216) was used to combine these multiple runs to generate the log and tree files. TreeAnnotator v2.4.3 was used to summarize all trees to report the median of tree heights and its 95% BCI (217). All xml files for these runs can be found here https://github.com/XuetingQiu/LAIV_impacts.

Ethics Statement

The majority of the genetic and epidemiological data used are publicly available, but without the host's identification. Publicly available data from GISAID do not include protected health information or personal identifiers of patients from which the samples were isolated. The newly sequenced data from 81 patients have been deidentified before these data were used. No further medical or private information was disclosed. The original intervention study and data collection was conducted with IRB approval from Baylor College of Medicine. The subsequent sequencing activity and secondary data analysis of these biological samples was also approved by Baylor's IRB.

Results

Phylogeny of Texas H3N2 samples

To examine the phylogeny of Texas samples, Bayesian phylogenetic analysis was applied to the full dataset. The phylogeny (Figure 5.1 and Supplemental Figure III-1 with taxa information) showed the typical seasonality of influenza epidemics where each season of the Central Texas Trial samples clustered together with the global and U.S. samples in the corresponding North Hemisphere season. These Texas samples were locally clustered between WBC-uv and TB-v, and some globally clustered with U.S. and ECA samples. The estimated median of the most recent common ancestor time (tMRCA) was 2004.29 with a very narrow 95% Bayesian Credible Interval (BCI) as [2004.21, 2004.37]. The estimated median of the substitution rate was 5.29E-3 substitutions/site/year with 95% BCI as [4.60E-3, 5.98E-3].

Transmission dynamics in the global scale

Discrete trait model with BSSVS was employed to further quantify the transmission dynamics between the global regions, the U.S. and Texas. Results (Figure 5.2) showed that the U.S. significantly diffused H3N2 to EAP, ECA and LAC global regions, while EAP and EAC significantly transmitted viruses to the U.S. Among the world regions, EAP and ECA were the main sources to disseminate viruses to other regions. Since Texas-clustered subsampling scheme was used to capture the transmissions into and out of Texas, the role of SAS and MENA regions on viral diffusion may not be fully depicted by this dataset. Further to explore Texas samples with two discrete traits as the vaccinated cities TB-v and the unvaccinated cities WBC-uv, the transmissions were mostly restrained between these locals or with the U.S., while global regions rarely interacted with Texas local. Only one significant transmission between ECA and WBC-uv was observed.

In detail (Table 5.1), the U.S. had 15.75 discrete state transitions with a rate of 1.52 (95% BCI: [0.56, 2.90]) events per lineage per year to transmit to WBC-uv and 24.49 discrete state transitions with a higher rate of 2.35 (95% BCI: [0.94, 4.08]) events per lineage per year to transmit to TB-v during 2004-2006, respectively. But these Texas local cities had very low transmissions back to the U.S., where the transmission between WBC-uv and the U.S. was even not statistically supported. Besides higher transmissions into TB-v from the U.S., Texas local transmissions reported 6.44 discrete state transitions from WBC-uv to TB-v, while the vice versa was 2.07, both of which were statistically supported. The exact transmission rates and their 95% BCI visualized in Figure 5.2 can be found in Table 5.1.

Transmission dynamics with both discrete trait analysis and structured coalescent

To further explore the transmission dynamics between Texas local and other regions, discrete trait model and structured coalescent model were both applied to the subsampled dataset, including 127 "Global" sequences, 32 "TB-v" and 29 "WBC-uv". The structured coalescent was used to quantify the migration rate, since it has the advantage to differentiate mutation events and migration events and improve the accuracy of phylogenetic reconstruction. The estimated medians of tMRCA and substitution rate were very similar from these two approaches (Supplemental Table III-1). These results showed a slightly wider 95% BCI compared with the estimates from the full dataset, which was probably because a smaller sample size can increase the uncertainty of the estimates.

With the discrete trait model (Supplemental Table III-2), the transmissions between the Global, WBC-uv and TB-v showed very similar results with the full dataset: higher transmissions from the Global and WBC-uv into TB-v. But the transmissions from Texas local to the global were not statistically supported. The Global had 18.3 jump counts into WBC-uv and 25.6 jump counts into TB-v with decisively statistical support (BF≥100). WBC-uv had higher jump counts into TB-v (n=5.3) with a rate of 1.06 (95% BCI: [0.03, 2.76]) events per lineage per year compared to the vice versa (n=2.3) with a rate of 0.61 (95% BCI: [0.0005, 2.08]) events per lineage per year. These results suggested that compared to the unvaccinated population, vaccination intervention in a population may interrupt the transmission chains inside one population and require more introductions from external sources. Tree structure (Supplemental Figure III-2 with taxa information) also showed that the samples from Texas vaccinated population had less clustering and higher diversity.

Comparably, results with structured coalescent model showed that the median of migration rate from TB-v to WBC-uv was 0.45 (95% BCI: [5.48E-4, 1.30]) events per lineage per year, while the vice versa was 3.65 (95% BCI: [1.18, 7.04]) events per lineage per year. This result also emphasized that for the local transmission, the vaccinated population needed more external introductions to sustain its epidemics. Furthermore, the coalescence rates reported as coalescent time by the model showed that the Texas vaccinated population had a longer coalescent time (6.50 years, 95% BCI: [0.35, 35.54]), compared to the Texas unvaccinated population (0.08 years, 95% BCI: [0.005, 0.31]). The average 6.5 years to find a coalescent event in the vaccinated population indicated that the vaccinated population cannot form community transmissions but need various viral introductions from external sources.

Structured coalescent model for Texas samples only

To further narrow down and examine the transmission dynamics between Texas local, a separate structured coalescent model was conducted to the 61 isolates in Texas during 2004-2005 and 2005-2006 influenza seasons. The phylogenetic tree (Figure 5.3a and supplemental Figure III-3 with taxa information) demonstrated a typical seasonal cluster within each flu season. The median of the estimated tMRCA was 2004.12 with 95% BCI as [2003.41, 2004.65]. The estimated median of substitution rate was 6.91E-3 with 95% BCI as [3.99E-3, 1.07E-2], which was slightly higher than the estimates from the full mater dataset and also had higher uncertainty with wider 95% BCI due to small sample size. The structured coalescent model indicated that WBC-uv was more likely to be the “source” transmitting virus to TB-v. The transmission counts from WBC-uv to TB-v had an estimated median of 57 (95% BCI: [28, 144]). The median of transmission counts from TB-v to WBC-uv was 27 (95% BCI: [0, 106]), where the 95% BCI

included 0, which indicated the transmission was not statistically supported. The median of the migration rate (Figure 5.3b) from TB-v to WBC-uv was 0.32 (95% BCI: [3.24E-4, 1.20]), and the vice versa was 2.40 (95% BCI: [0.70, 5.00]), which was statistically supported with a decisive BF= 690.78. Texas vaccinated population had an average coalescent time of 9.83 years (95% BCI: [0.06, 60.89]), while Texas unvaccinated population had the average coalescent time of 0.69 years (95% BCI: [0.02, 4.81]). This evidence echoed the results in the previous structured coalescent model conducted with the subsampled dataset, that is, vaccinated population had higher viral diversity with various introductions from external sources. The large coalescent time indicated that the vaccinated population cannot form community "clonal" transmissions.

Discussion

Global patterns of seasonal influenza H3N2 viruses on their genetic and antigenic variations have been well characterized (218). Frequently updated vaccines are necessary to match the rapidly changing antigenicity of circulating strains. Since the first introduction of the live attenuated influenza vaccines (LAIV) in 2003 in the U.S., the H3N2 vaccine candidates recommended for the North Hemisphere have been updated annually for continuous three seasons (A/Moscow/10/99-like virus for 2003-2004; A/Fujian/411/2002-like virus for 2004-2005; A/California/7/2004-like virus for 2005-2006). Safety and efficacy studies on the vaccines have been conducted continuously, however, studies that explore the potential impacts of LAIV on viral population are rare. Therefore, this study took the advantage of data availability from the Central Texas Trial on the Control of Epidemic Influenza during 2004-2006 to quantitatively evaluate how vaccination could impact the viral genetic diversity and transmission dynamics for the first time. Results from this study indicate that the vaccinated population has higher genetic

diversity, needs more external introductions to sustain the epidemics in the community, and interrupts viral dissemination resulting in lower transmissions to external populations.

The vaccination intervention in Central Texas during 2003-2006 covered the first introduction of LAIV in 2003 but before the universal recommendation of LAIV seasonal influenza vaccines in the U.S. and the globe, so that data outside the Texas represented the general population without any LAIV usage. This provided a unique opportunity to compare the population with planned usage of LAIV population to LAIV-naïve populations. To gradually narrow down the evaluation scope of the impacts of LAIV on genetic diversity and disease transmissions, analysis on different geographic scales were included: from a full global scale to a global subset scale and lastly to Texas local scale. The global transmission dynamics in this study echoed other studies regarding the important role of East Asia to disseminate H3N2 seasonal virus to other regions (20), but the U.S. itself in this study was also identified as a major source. This result may be an artifact of the phylogenetic clustering subsampling strategy used in this study, which intended to capture all isolates that were closely clustered with the Texas samples but resulted in a "biased" sample (overrepresented U.S. sample) for the global transmission. Regardless, analyses from these different geographic scales consistently supported that the Texas vaccinated population received more introductions from Texas unvaccinated population but spread less to the Texas unvaccinated population.

Furthermore, the structured coalescent model with only Texas data inferred that compared to the unvaccinated population, the vaccinated population had a very large averaged coalescent time to trace back and find a coalescent event for two viral samples. This indicated that the vaccination may interrupt the "clonal" transmissions inside one community but require more external introductions to sustain the epidemic. It supports the hypotheses I intended to test.

Though the biological mechanisms related to the host immune landscape in different populations cannot be inferred with the data, one potential explanation could be that the vaccination decreases the proportion of susceptible people in the population or/and reduces the amount of shedding virions in the infected people, which could interrupt the transmission chain within the population (219). With fewer susceptible hosts, the transmissions inside the population may be unlikely to establish but continuously introductions are needed to sustain the seasonal epidemics. Less virion shedding may not only break the transmission chain inside the vaccinated population but also reduce the possibility to transmit to the unvaccinated population. Taken together, though the epidemics in the vaccinated population have higher viral genetic diversity, which is probably a phenomenon created by "gathering" viruses from various external introductions, a lower dissemination from the vaccinated population could ultimately interrupt the transmissions during the epidemic season. This may have explained the herd immunity of vaccination from viral phylogenetic perspective.

This study quantitatively evaluated the impacts of LAIV on viral genetic diversity and transmission dynamics, and provided phylogenetic evidence to recommend the usage of vaccines. But several limitations exist. These Texas samples are secondary data from an intervention study of which the primary goals were to improve the vaccination coverage in school-aged children and assess the safety and efficacy of the vaccines, where only very small number of positive samples (about 10% randomly selected) were able to be sequenced. This may result in an unrepresentative sample of the viral population. Furthermore, the original intervention study was inconsistent for each influenza season due to changes in epidemic start time or vaccination policy. For example, the epidemic started early during 2003-2004 with the H3N2 Fujian-like virus, which resulted in vaccination intervention unable to be implemented in a

timely manner. Another limitation is that the vaccination status of these samples was categorized based on the population status rather than individual level. Original studies (208,212) to explore the vaccine efficacy showed that the vaccination provided both direct protection and herd immunity in the intervention cities, which supported the study to categorize the vaccination status on the population level. But it restrains that results in this study only speak on the population-level. Finally, the study assumed that the epidemic sizes in vaccinated and unvaccinated population were the same based on the comparable total population size and age structure during each season within each location. A different conclusion may be drawn from other populations when they have a very different epidemic size each season.

Nevertheless, this study showed consistent and statistically supported results from different geographic scales and different models, emphasizing that vaccination in a population could prevent disease transmissions to other populations. Phylogenetically, the vaccinated population has higher viral genetic diversity, needs more external introductions, but disseminates less to other populations. Further studies on a larger sample size with vaccination status available at the individual level are needed to verify these findings and identify the involved biological mechanisms. Taken together, this study has found the phylogenetic evidence for the benefits of vaccination.

Tables and Figures

Table 5.1. Estimated introductions, transmission rates and their 95% BCI between different locations. The items in the cell in order are count of introductions, transmission rate and its 95% BCI. The cell color indicates Bayes Factor levels, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

Geographic region abbreviation: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), South Asia (SAS), Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).

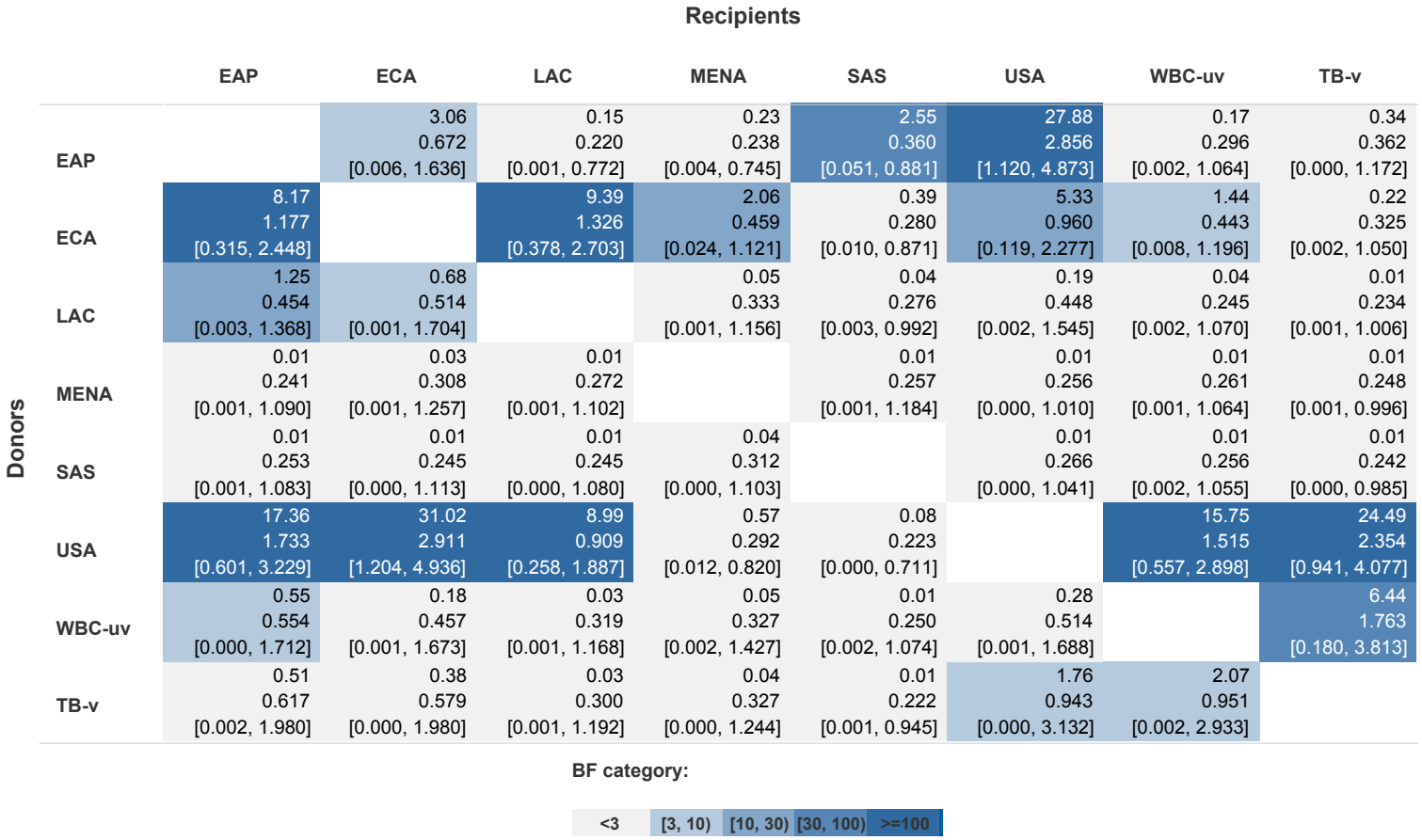

Recipients
EAP
ECA
LAC
MENA
SAS
USA
WBC-uv
TB-v
Donors
EAP
3.06
0.672
[0.006, 1.636]
0.15
0.220
[0.001, 0.772]
0.23
0.238
[0.004, 0.745]
2.55
0.360
[0.051, 0.881]
27.88
2.856
[1.120, 4.873]
0.17
0.296
[0.002, 1.064]
0.34
0.362
[0.000, 1.172]
ECA
8.17
1.177
[0.315, 2.448]
9.39
1.326
[0.378, 2.703]
2.06
0.459
[0.024, 1.121]
0.39
0.280
[0.010, 0.871]
5.33
0.960
[0.119, 2.277]
1.44
0.443
[0.008, 1.196]
0.22
0.325
[0.002, 1.050]
LAC
1.25
0.454
[0.003, 1.368]
0.68
0.514
[0.001, 1.704]
0.05
0.333
[0.001, 1.156]
0.04
0.276
[0.003, 0.992]
0.19
0.448
[0.002, 1.545]
0.04
0.245
[0.002, 1.070]
0.01
0.234
[0.001, 1.006]
MENA
0.01
0.241
[0.001, 1.090]
0.03
0.308
[0.001, 1.257]
0.01
0.272
[0.001, 1.102]
0.01
0.257
[0.001, 1.184]
0.01
0.256
[0.000, 1.010]
0.01
0.261
[0.001, 1.064]
0.01
0.248
[0.001, 0.996]
SAS
0.01
0.253
[0.001, 1.083]
0.01
0.245
[0.000, 1.113]
0.01
0.245
[0.000, 1.080]
0.04
0.312
[0.000, 1.103]
0.01
0.266
[0.000, 1.041]
0.01
0.256
[0.002, 1.055]
0.01
0.242
[0.000, 0.985]
USA
17.36
1.733
[0.601, 3.229]
31.02
2.911
[1.204, 4.936]
8.99
0.909
[0.258, 1.887]
0.57
0.292
[0.012, 0.820]
0.08
0.223
[0.000, 0.711]
15.75
1.515
[0.557, 2.898]
24.49
2.354
[0.941, 4.077]
WBC-uv
0.55
0.554
[0.000, 1.712]
0.18
0.457
[0.001, 1.673]
0.03
0.319
[0.001, 1.168]
0.05
0.327
[0.002, 1.427]
0.01
0.250
[0.002, 1.074]
0.28
0.514
[0.001, 1.688]
6.44
1.763
[0.180, 3.813]
TB-v
0.51
0.617
[0.002, 1.980]
0.38
0.579
[0.000, 1.980]
0.03
0.300
[0.001, 1.192]
0.04
0.327
[0.000, 1.244]
0.01
0.222
[0.001, 0.945]
1.76
0.943
[0.000, 3.132]
2.07
0.951
[0.002, 2.933]
BF category:
<3
[3, 10)
[10, 30)
[30, 100)
>=100

Figure 5.1. Maximum clade credibility tree of H3N2 phylogeny with global regions, the U.S., and 2004-2006 Central Texas Trial samples.

Geographic region abbreviation: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the US), South Asia (SAS), Sub-Saharan Africa (AF), Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).

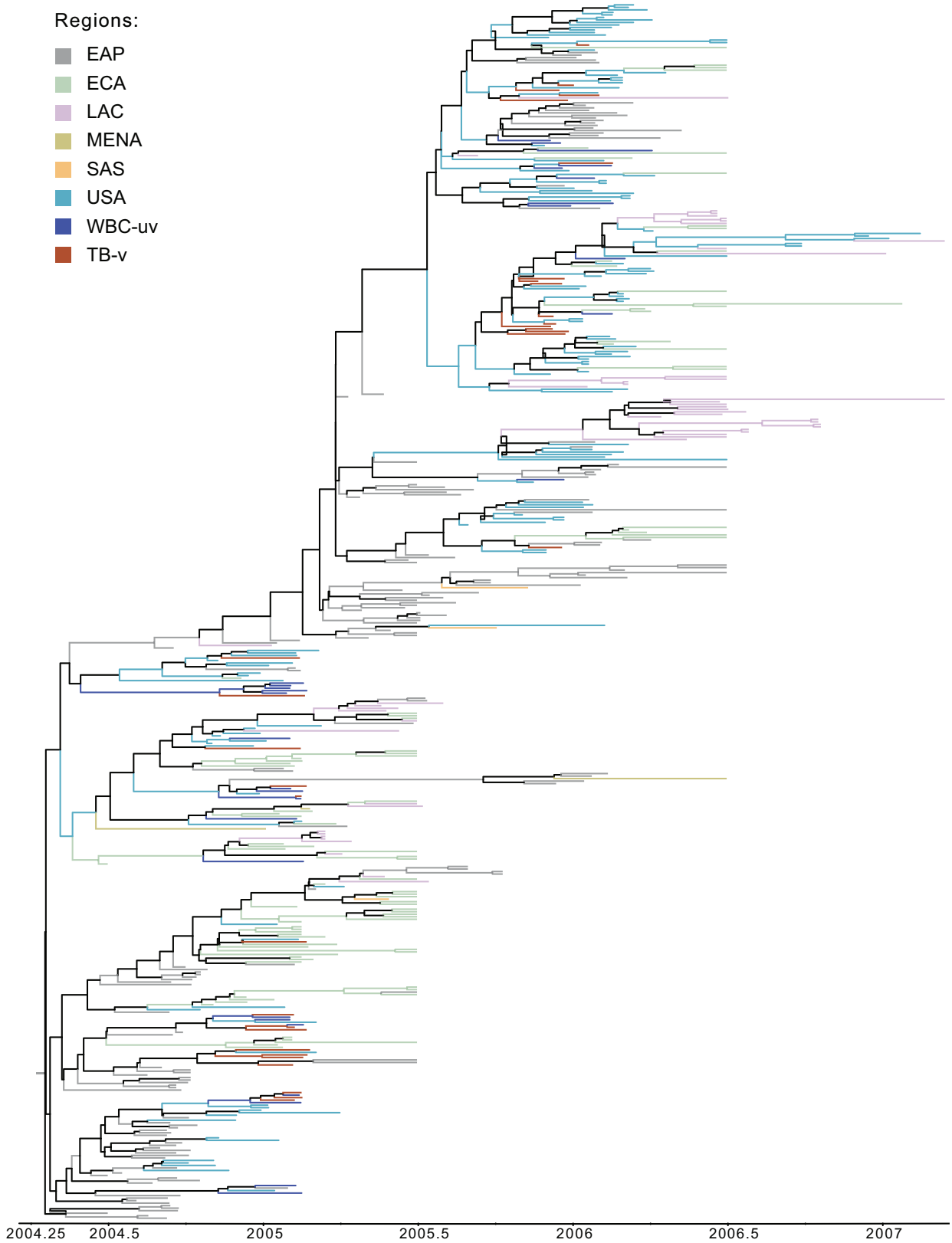

Regions:
EAP
ECA
LAC
MENA
SAS
USA
WBC-uv
TB-v
2004.25
2004.5
2005
2005.5
2006
2006.5
2007

Figure 5.2. Transmission dynamics between global regions, the U.S., Texas vaccinated cities and Texas unvaccinated cities. Only statistically supported (Bayes Factor ≥3) transmission rates are shown, with the arrow stroke weight proportional to the rate values.

Geographic region abbreviation: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the US), South Asia (SAS), Sub-Saharan Africa (AF), Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).


WHO Regions:
NA
LAC
ECA
MENA
AF
SAS
EAP
Local legend:
USA
Texas
WBC-uv
TB-v

Figure 5.3a. The typed-node maximum clade credibility tree for Texas-unvaccinated and Texas-vaccinated populations. The stroke weight of the branch is proportional to the posterior probability of the assigned type.

Geographic location abbreviation: Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).

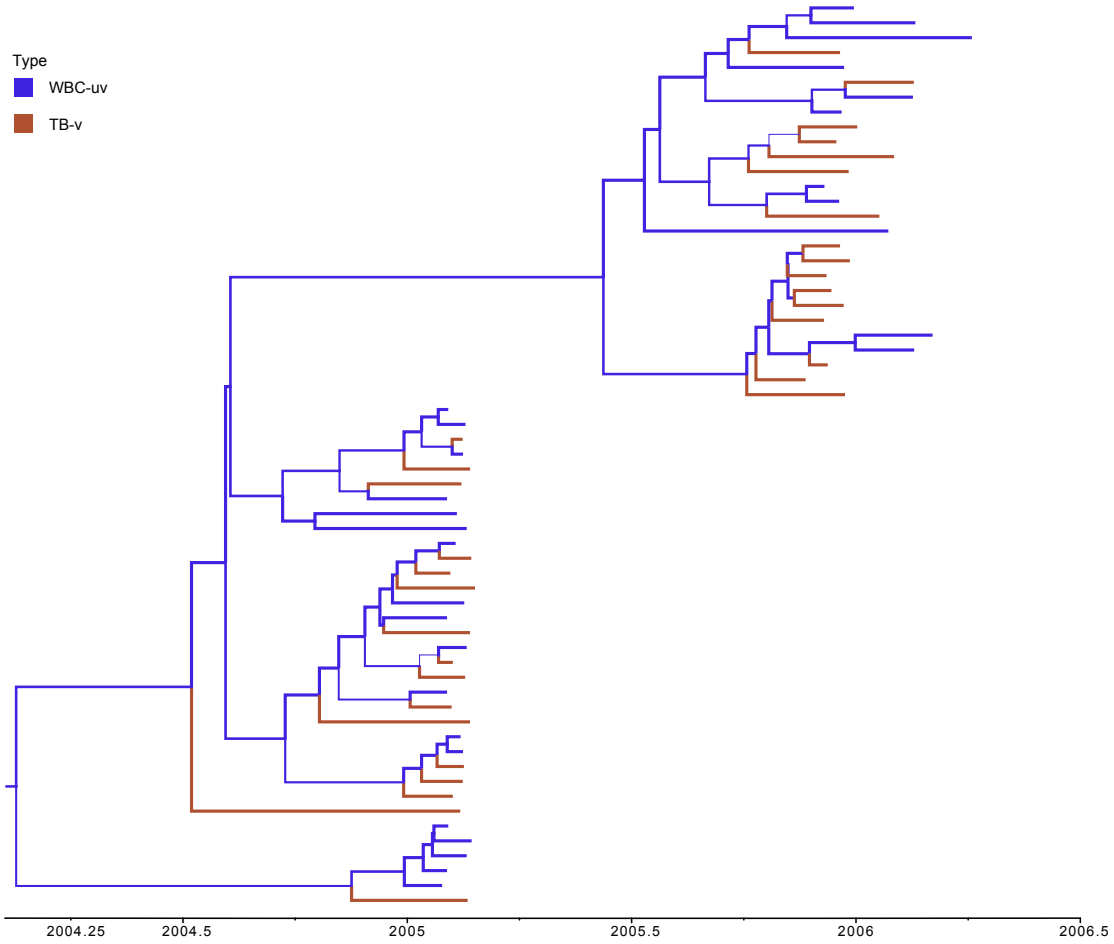
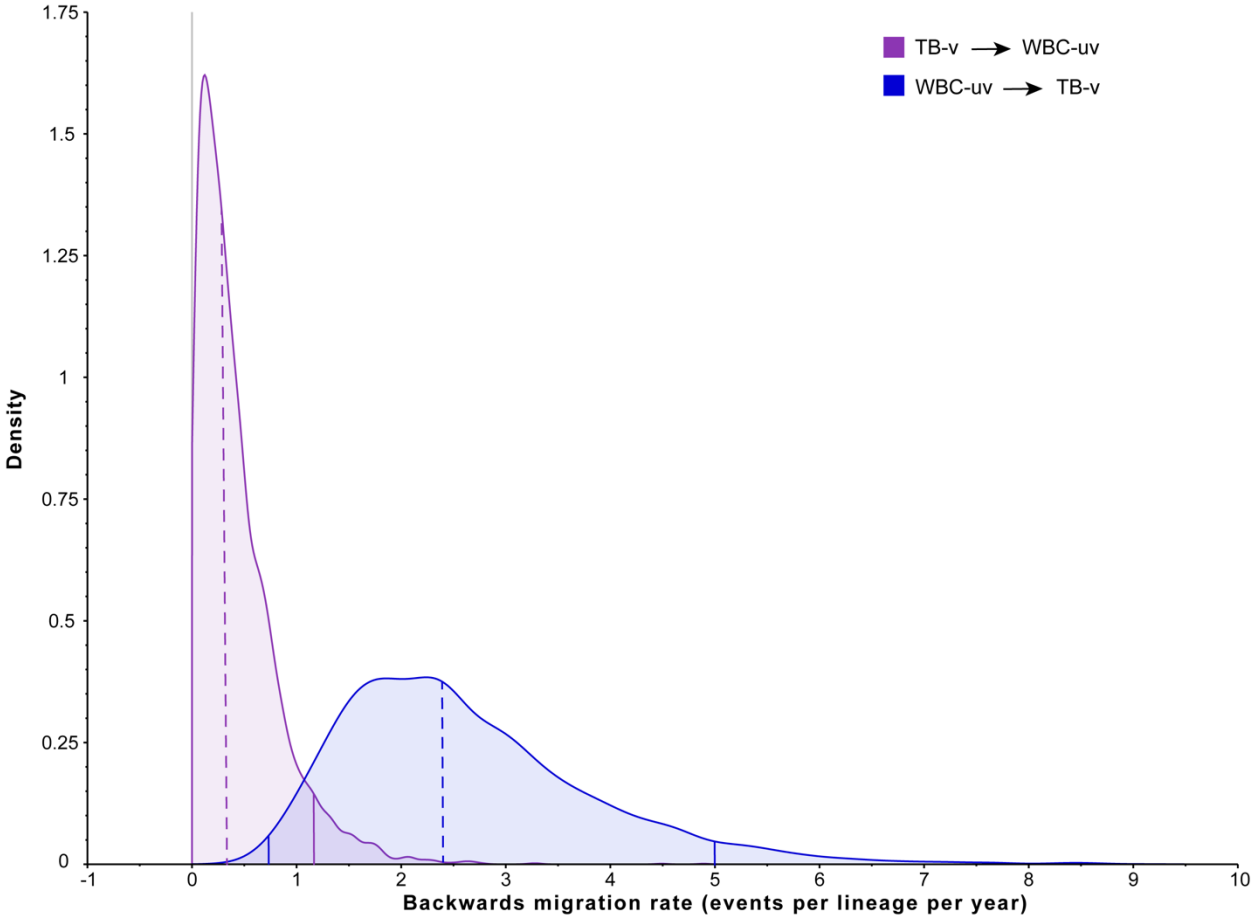

Type
WBC-uv
TB-v
2004.25
2004.5
2005
2005.5
2006
2006.5

Figure 5.3b. The posterior backwards migration rate estimated between Texas-unvaccinated and vaccinated populations. The y-axis represents density on a log scale. The left curve is the migration rate from Texas vaccinated to Texas unvaccinated population and the right curve is vice versa. The median migration rate from TB-v to WBC-uv is 0.32 (95% BCI: [3.24E-4, 1.20]) events per lineage per year, and the vice versa is 2.40 events per lineage per year (95% BCI: [0.70, 5.00]), which is statistically supported with a decisive BF= 690.78.

Geographic location abbreviation: Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).


TB-v → WBC-uv
WBC-uv → TB-v
Density
1.75
1.5
1.25
1
0.75
0.5
0.25
0
-1
0
1
2
3
4
5
6
7
8
9
10
Backwards migration rate (events per lineage per year)

CHAPTER 6

SUMMARY AND CONCLUSIONS

The major objective of this dissertation is to develop and apply advanced phylogenetic models in the Bayesian framework to enhance the understanding of viral evolution and spatiotemporal dynamics. The targeted pathogens are primarily different influenza A virus subtypes and influenza B viruses, which continue to pose global public health threats, including several pandemics in history, heavy disease burdens in humans with annual seasonal outbreaks, and high economic loss with highly pathogenic viral strains causing large outbreaks in avian hosts (220). Vaccines are the most cost-effective approach to prevent infectious diseases, however, seasonal influenza viruses require large efforts to frequently update and distribute the vaccines annually, while zoonosis of the potential pandemic strains poses another threat of no stockpile of effective vaccines.

To overcome the predicament of influenza vaccines, the National Institute of Allergy and Infectious Diseases (NIAID) has summarized the important research areas of influenza viruses that envision a transformative effort toward successful development of a universal influenza vaccine (54). A major component of research areas requires an improved understanding of influenza viral evolution and transmission to improve disease control measures. With advanced development of mathematical models, global genomic data sharing and improvement on computing power, computational modeling could be a powerful tool to understand this rapidly evolving pathogen and achieve the goal of a broadly protective or universal influenza vaccine.

Therefore, to extend and apply advanced models with integrating large genetic and epidemiological data to understand viral evolution and diffusion dynamics is critical for disease prevention and control (18).

In this dissertation, I aimed to examine the evolutionary and epidemiological dynamics of influenza virus across ecological scales in order to improve disease control. To achieve this, specific studies were conducted to provide insights on: whether integrating viral protein structure can improve phylogenetic inference and better understand the evolution of vaccine targets; how the seasonal influenza epidemic spread among U.S. regions; and whether increased vaccination rate within a community can impact viral diversity and diffusion dynamics. Specifically, aim 1 focused on developing a novel model to incorporating functional structure of hemagglutinin (HA) protein; aim 2 examined viral diffusion patterns and environmental factors that potentially affect the diffusion in the U.S. via phylodynamic modeling; aim 3 explored the impacts of H3N2 live attenuated influenza vaccine (LAIV) on viral genetic diversity and diffusion dynamics in Central Texas, U.S., with a unique dataset. In this summary chapter, I will review the major results related to each aim, explore the potential applications, and finally discuss future directions.

Highlights

The novel structurally informed model developed in aim 1 incorporates the different structure domains and immune targets of HA glycoprotein, where globular head domain is the hypervariable and immunodominant region undergoing continuous antigenic drift while the more conserved stalk domain proximal to viral envelope is the immune-subdominant region (118,221). Though it has been reported that the head domain is highly plastic and amino acid mutations on

the head domain are the main sources for immune escape (65,118,157,221), there was no computationally flexible model to specifically quantify the evolutionary characteristics on each specific immunogenic domain, especially for viruses with a large population. Previously studies have used partitioning strategy mostly for multiple viral gene segments with keeping each whole gene as a partition (136) or on the codon positions inside one single gene (137), but partitioning strategy to reflect biologically functional structures have not been applied to HA. One major concern would be over-parameterization, where partitioning scheme adds extra parameters into the model but short length in each partition may not contain sufficient information for all the parameters to provide accurate reconstruction and estimation on viral evolution (136,222). Therefore, I proposed and tested a new model to incorporate a structurally informed partitioning scheme on a single protein into phylogenetic reconstruction. I evaluated the model fit and parameter estimations of four different models – HKY base model, SRD06 codon model, HKY with a structurally informed partitioning scheme, SRD06 with a structurally informed partitioning scheme - on pandemic H1N1pdm09, seasonal H1N1postpdm, A/H3N2, B-Yamagata-like and Victoria-like lineages, and two highly pathogenic avian influenza A viruses H5Nx and H7N9. Decisive statistical support from the model selection procedure (via both path sampling and stepping-stone sampling) and accurate estimation on important evolutionary parameters validated the significance and superiority of the new model (i.e. SRD06 with a structurally informed partitioning scheme), which partitions on both codon positions and protein structural domains. The model can further provide biological insights for viral evolution, with domain-specific evolutionary rates and estimations on approximate selection pressures (i.e., $d\mathrm{C}_{1+2}/d\mathrm{C}_3$, the approximate ratio of non-synonymous changes over synonymous changes). The tree branches with domain-specific $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ can inform the approximate selection pressure on

each strain, indicating some biological explanations related to antigenic drift and emerging strains. An extended case study of the new model on RSV showed similar results and conclusions on the different domains of RSV Fusion protein which is one of the main spike-shaped surface glycoprotein on RSV viral envelope to induce immune responses in humans (127). Taken together, integrating a functionally informed partitioning scheme based on protein structures of immune targets allows for significant improvement of phylogenetic analysis and providing important biological insights for vaccine design.

Aim 2 was conducted to understand the viral diffusion patterns and environmental factors that potentially affect the spatiotemporal dynamics, which is critical for the prevention and control of influenza outbreaks (18). Standing on the knowledge learned from other global studies on seasonal influenza virus diffusion (18–20,46), aim 2 specified the global introductions into the U.S. and explored the diffusion dynamics amongst the U.S. regions. I fully took the advantage of large and complete genetic and epidemiological datasets of all four subtypes/lineages of seasonal influenza viruses to answer important questions regarding viral diffusion patterns and significant predictors. Multiple global viral sources diffused viruses into the U.S., showing that East Asia and Pacific (EAP), Europe and Central Asia (ECA), and Latin America and Caribbean (LAC) are generally the main sources of transmission to the U.S., though each flu season in the U.S. may have a different major viral source. The diffusion patterns were compared longitudinally for each seasonal influenza subtype/lineage, showing that external introductions are the only sources for A/H3N2 seasonal outbreaks but local persistence from the prior season could be a source for A/H1N1, B-Victoria and B-Yamagata (especially for B-Victoria in which each season has viral source from the prior season). This viral source information can be valuable for the prediction of circulating strains in the U.S. Dynamics

amongst different regions (based on HHS official regions) in the U.S. demonstrated complicated diffusion patterns between regions, where a major hub diffused virus to all other regions: region 5 for A/H3N2 and B-Victoria, region 4 for A/H1N1, and region 8 for B-Yamagata. With incorporated epidemiological and ecological factors for these regions into the model via generalized linear model, important predictors for disease transmission were identified, such as geographic distance, flight connections, and different population age structures. Longer geographic distance was identified as a barrier for viral diffusion. But more flight connections and lower proportion of the adults (18 to 64 years) [i.e., higher proportion of the youth (<18 years) and the elderly (>65 years)] in the population may drive viral spread. This information highlights that improving prevention hygiene in the heavy traffic settings such as airports and increasing vaccine administration to the high-risk populations could be effective measures to prevent viral spread.

Aim 3 zoomed in to check viral diffusion among local cities in Texas, with advanced phylodynamic modeling and structured coalescent model (115,223). Seasonal influenza H3N2 viruses are highly antigenic diverse and have been well characterized on its global pattern of antigenic dynamics (218). This study took the advantage of data availability from the Texas Central Trial on the Control of Epidemic Influenza during 2004-2006 to quantitatively evaluate how H3N2 LAIV could impact the viral genetic diversity and transmission dynamics for the first time. The vaccine intervention was largely implemented to the school-aged children in the community scale in the experimental cities, while the control cities had no vaccine intervention. Results from this study indicated that the vaccinated population displays higher genetic diversity, needs more external introductions to sustain the epidemics, and interrupts viral dissemination resulting in lower transmissions to external regions. Therefore, this study found the phylogenetic

evidence of vaccination, which could affect viral diffusion in a beneficial way by interrupting the epidemic chains in the population, probably due to the immune landscape change in the susceptible and high-risk population.

Applications

Findings from developing the structurally informed phylogenetic model and exploring viral diffusion dynamics carry important insights of vaccination scheme and prevention measures for influenza viruses. How to apply the new model for optimizing vaccination strategy and how to come up effective prevention measures based on these findings will be discussed in this section.

Vaccination strategy

Vaccination is amongst the most cost-effective approaches available to prevent infectious diseases (224,225). Though the efficacy of seasonal influenza vaccines could vary from 10% - 60%, the efforts on annual vaccine updates and administration have provided significant protection (84). In aim 2, results from the generalized linear model supported that higher vaccination rate in a region could lower the viral diffusion rates to other regions in the U.S. region-level. Furthermore, local dynamic study in Central Texas in aim 3 demonstrated that vaccination in school-aged children could interrupt the epidemic chain in the population with requiring more external introductions to maintain viral genetic diversity and sustain the epidemics but spread less viruses to unvaccinated populations. Therefore, the importance of vaccination to prevent influenza is valid.

Another perspective regarding vaccination reflected in the study is who should be vaccinated. Aim 2 reported that the population age structure is a significant predictor for viral

spread. Higher proportion of the youth (<18 years) and the elderly (> 65 years) is positively associated with the migration rates in some viral subtypes/lineages, and lower proportion of adults (18 to 64 years) in the population can increase the viral migration rates with decisive statistical support in the joint estimate. These populations, especially children < 5 years and the elderly > 65 years, are considered at high risk for influenza infections (226). Currently, individuals age 6 months or older are recommended to receive 1 dose inactivated influenza vaccine (IIV), recombinant influenza vaccine (RIV) or LAIV annually (227,228). With our findings highlighting the roles of young children and the elderly on spreading viruses, the typical vaccination strategy may need to be reconsidered and reformed. According to the benefits of community-based vaccination on school-aged children in Central Texas (208–210,212), mandatory vaccination via community-level intervention could be used to guarantee a high percentage of vaccination in the high risk populations, including children and the elderly in the community. Furthermore, special vaccination scheme of customized high-dose vaccine or mixed types of vaccines needs to be developed for the weaker immune responses of young children and older adults. Cowling et al (229) is currently leading a clinical trial (ClinicalTrials.gov Identifier: NCT03330132) in Hong Kong to test the immune profiles over time of older adults aged 65-82 years following different influenza vaccination strategies. This study to find the best vaccination strategy for the elderly has been undergoing since 2017. The importance of this study or future extended study for young children is supported by the findings in this dissertation.

Other Effective prevention measures for influenza

Besides vaccination, aim 2 also identified and confirmed an important congregation place where regular non-pharmaceutical prevention measures could be implemented, that is, in the

airports (or inside airplanes). Lemey et al. (18) reported the important role of flow passengers on spreading viruses in their global study. Aim 2 reported regions with busiest airports played as a primary hub for viral transmissions in the U.S. and flight connections between regions were recognized as a significant predictor for higher viral migration rates. Regular effective prevention measures include hand hygiene, wearing mask during sick, covering nose and mouth when coughing or sneezing, and disinfecting surfaces that are potentially contaminated (230,231). Especially, hand hygiene has been reported as the most effective prevention measure even in the household with close contact to sick individuals (230). A standardized disinfecting procedure for airports and airplanes, and improved hand hygiene via health education and providing disinfectant supplies to passengers during the flu season may help constrain viral spread.

Future Directions

In this dissertation, to enhance the understanding of viral evolution and diffusion dynamics and to facilitate disease prevention, novel phylogenetic model was developed and advanced phylodynamic models were applied. Findings in these specific studies can be applied to computational vaccine design and implementation of effective prevention measures for influenza infection. Future directions can be summarized into several perspectives.

Two extended applications of the structurally informed phylogenetic model

The structurally informed phylogenetic model developed in aim 1 has strong statistical support on its superiority to improve the accuracy of viral evolution reconstruction and provide valuable biological insights on protein structure domain-specific evolutionary rates and selection

pressure. To apply the novel model to computational influenza vaccine design, the reconstructed HA ancestral sequences via this model should be further tested in animal models to evaluate the magnitude and duration of cross-reactive immune responses (232). The application of this model can be extended to another influenza surface glycoprotein neuraminidase (NA) and to other viral pathogens with similar structurally distinctive immunogenic domains. The cross-reactive immunity induced by NA was largely ignored in vaccine development but has been drawing more attention in the scientific community recent years (233). NA has similar spike-shaped structure and high genetic diversity as HA, where the structurally informed model could be used to enlighten the domain-specific evolutionary history and predict how the viral population might evolve given a widely used NA vaccine. The potential to apply this model to other pathogens have been supported by the model testing on RSV in sub aim 1. It showed the model performed significantly better and provided extra information on viral protein domain-specific evolutionary characteristics. This model may provide valuable insights on vaccine design for RSV, of which currently no effective vaccine is available.

Diffusion dynamics incorporated into Nextstrain platform

Understanding and monitoring viral evolution and diffusion dynamics is important for effective prevention measures and surveillance. The diffusion dynamics from the global into the U.S. and among the U.S. regions can be incorporated into the nearly real-time tracking platform of pathogen evolution, Nextstrain (234). Nextstain is designed to integrate a database of pathogen genomes, an analysis pipeline of phylodynamic modeling and a flexibly interactive visualization (234). Though the platform is currently based on maximum likelihood approach, verifying and incorporating the dynamics generated by Bayesian framework could improve the

credibility of reconstruction on viral dynamics. The growing importance of real time tracking to prevent disease outbreak could be complemented by the accuracy and validation of viral diffusion dynamics reconstruction via Bayesian framework. In return, Nextstrain platform can provide a more informative prior to Bayesian inference via timely integrating data from new outbreak or early cases of the epidemic season, which can improve the model fitting in Bayesian framework to infer future diffusion dynamics.

Vaccine design

Traditional approaches have failed to produce stable and protective vaccines for hypervariable and rapidly-evolving viral pathogens, including influenza viruses and RSV (29,183). Currently, there is no effective vaccine for RSV (178). Challenges for influenza viruses include the rapid antigenic drifts and hypervariability of HA surface protein for different influenza subtypes, where influenza A viruses have been classified into Group 1 (H1, H2, H5, H6, H7, H8, H9, H11, H12, H13, H16, H17, and H18) and Group 2 (H3, H4, H7, H10, H14, and H15), and influenza B with two lineages B-Yamagata and B-Victoria (118). The annually WHO-selected seasonal vaccine candidates (235) have frequently mismatched with circulating strains and failed to provide broad spectrum and long-lasting protection from seasonal human strains, not to say the probability of against potential pandemic from zoonosis strains (236,237). Fortunately, the growth of databases containing genome sequences sampled throughout global epidemics (238–240), increased computational power and theoretical algorithms allow complex data sources to be integrated into a unified framework providing the opportunity for a more complete understanding of pathogen and host features. This makes computational approaches valuable to provide novel insights on vaccine selection and design.

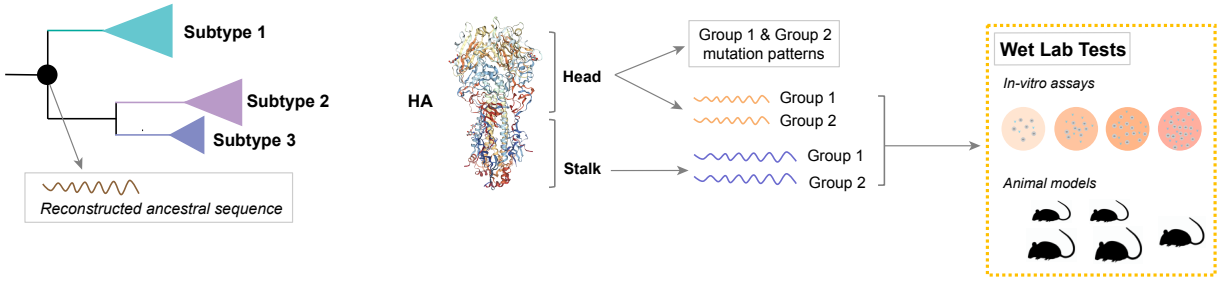
First, phylodynamic modeling to identify the viral diffusion patterns and potential predictors can improve the current approaches and be incorporated into predictive models for vaccine candidate selection (241,242). Aim 2 explored the main viral sources for the four subtypes/lineages of seasonal influenza viruses into the U.S. for each season. This information could be used to monitor and select vaccine candidates from the main viral source regions. Furthermore, epidemiological and ecological factors identified from global studies (18,20,243) and aim 2 in this dissertation could help predict the global and national diffusion patterns of seasonal influenza. Predictive models of viral evolution to forecast dominant circulating influenza viral strains in the upcoming influenza seasons through the analysis of genetic and epidemiological data from influenza surveillance system have been developed to make quantitative predictions of viral evolution and aim to improve the selection of seasonal influenza vaccine candidates (244,245). Though predictive model still relies on the traditional vaccine design pipeline requiring annual surveillance, it has demonstrated the potential to integrate multiple data sources to improve influenza vaccine design (244,245).

Secondly, computational approaches to identify candidates for influenza vaccine design have been used with a variety of novel vaccine production strategies in development, including epitope-based design (79,246,247) and multiple sequence alignment comparison to generate computationally optimized broadly reactive antigens (COBRA) (86,248–251). These approaches mainly focus on the 'unnatural immunity' (252) induced by more conserved or less immune-dominant domains in the surface proteins, internal proteins or both, to tackle with the high degree of variability in influenza viruses by boosting the immunity from the conserved or less evolvable proteins of the viruses. With the concept of unnatural immunity, ancestral sequence reconstruction based on maximum-likelihood (ML) approach (253) to generate evolutionarily

conserved sequences or HA stalk-based vaccine design (122,157,254) to use the conserved antigens in the stalk domain have been developed. The ancestral sequence reconstruction with ML approach has the disadvantage of being prone to sampling biases of viral sequences and not accountable for the variability of substitution rates among sites (253). Though currently HA stalk-based vaccine design is not based on computational approaches (232), it has demonstrated cross-reactive protection for different subtypes of influenza viruses (254). Despite the potential for HA-stalk design to elicit broadly reactive immune response, a number of challenges remain (reviewed in (255) and (65)), including a limited understanding of the full repertoire of potential epitopes on viral proteins.

The novel structurally informed model developed in aim 1 can overcome the problems in ML approach and extend the concept of stalk-based design to identify the full repertoire of conserved epitopes in all viral proteins (Figure 6.1). The structurally informed model, a Bayesian approach, generates a distribution of the possible ancestral sequences with accounting for the uncertainty of the estimation, rather than only one or a few maximum likelihood sequence (256). It also considers that the substitution rate variations in the structural domains and the codon positions of pathogen protein can be under disparate immunologic pressures and thus have impacts on the evolutionary phylogeny (118) and the accuracy of ancestral sequence reconstruction. With this approach, ancestral sequences to capture the conserved epitopes for the whole genome can be used to test the cross-reactive protection from multiple subtypes of influenza viruses, which ultimately aims for designing a universal vaccine (54). The extended review of computational approaches for influenza universal vaccine design can be found in the Appendix IV.

Figure 6.1. Ancestral sequence reconstruction from structurally informed phylogenetic modeling. This novel Bayesian approach can be used to reconstruct ancestral sequence at the ancestral node (shown as black dot on the tree). Evolutionary models that incorporate protein structural domains can separately estimate the evolutionary history on each functional partition as the HA head and stalk domains. Based on the evolutionary relationship among different subtypes of influenza A virus, common ancestral sequences of head and stalk domains can be generated within influenza A virus Group 1 (H1, H2, H5, H6, H7, H8, H9, H11, H12, H13, H16, H17, and H18) and within Group 2 (H3, H4, H7, H10, H14, and H15), respectively. It can generate common ancestral sequences or understand the mutation characteristics separately for each group. Outputs from this approach, like ancestral epitopes, peptides, or proteins will be tested at in-vitro and/or in-vivo models to evaluate their immunogenicity. The proposed concept as shown is based on HA gene sequences, but it should be used for all the gene segments of influenza viruses to generate a full profile of viral immunogenicity.



Verifying the significant predictors in a higher resolution

In aim 2, the geographic unit for testing the epidemiological and ecological predictors is the U.S. Human Health Service regions, resulting in lower resolution of these predictors with averaged measures from multiple states across several seasons. It is important to verify the conclusion in a smaller geographic unit, for example, different cities in the U.S. After verified, including these predictors in predictive models for circulating viral strains may improve the accuracy of prediction (245).

Vaccine impacts on viral evolution with long-term data

Aim 3 found that population vaccinated with live attenuated vaccines required more external viral introductions to sustain the epidemic and had less disseminations to unvaccinated population. To further confirm the impacts of vaccines on viral genetics and diffusion dynamic, long-term or multi-location datasets are needed. One potential great dataset would be from the ongoing clinical trial for the elderly in Hong Kong (229). This study could provide data with a four-year time span and using different vaccination strategies in the elderly, including standard inactivated influenza vaccine, MF59-adjuvanted inactivated influenza vaccine, high-dose inactivated influenza vaccine, and recombinant HA inactivated influenza vaccine. Data from this study will help evaluate the impacts of different types of influenza vaccine on viral diversity and diffusion dynamics when vaccinating the elderly in the community.

Conclusions

Taken together, the main contribution of this dissertation includes: 1) the integration of protein structure to improve phylogenetic model on viral evolutionary reconstruction; and 2) the application of advanced viral phylodynamic modeling to quantify viral diffusion dynamics and vaccine impacts in the U.S. The new phylogenetic evolutionary model that accounts for rate variations across a single protein and across codon positions significantly improves phylogenetic reconstruction of influenza viruses (and RSV). It is a valuable tool that is able to provide biological insights on protein structure domain specific evolutionary characteristics and approximate selection pressure on these domains. Advanced phylodynamic modeling applied across ecological scales for the U.S. national and local data delivered information on the patterns of viral diffusion, significant epidemiological and ecological predictors that affect the diffusion,

and the role of vaccination on local disease dynamics. These findings from these studies provide the standing points for future studies to apply and verify the instructive recommendations on vaccine design, surveillance and prevention measures of influenza viruses.

REFERENCES

1. Ferkol T, Schraufnagel D. The Global Burden of Respiratory Disease. Ann Am Thorac Soc [Internet]. 2014 Mar 27 [cited 2019 Feb 11];11(3):404–6. Available from: http://www.atsjournals.org/doi/abs/10.1513/AnnalsATS.201311-405PS
2. Forum of International Respiratory Societies. The Global Impact of Respiratory Disease - Second Edition [Internet]. 2012 [cited 2019 Feb 11]. Available from: https://www.fi
3. Hendaus MA, Jomha FA, Alhammadi AH. Virus-induced secondary bacterial infection: a concise review. Ther Clin Risk Manag [Internet]. 2015 [cited 2019 Feb 12];11:1265–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26345407
4. Tregoning JS, Schwarze J. Respiratory viral infections in infants: causes, clinical symptoms, virology, and immunology. Clin Microbiol Rev [Internet]. 2010 Jan [cited 2019 Feb 12];23(1):74–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20065326
5. Regamey N, Kaiser L, Roiha HL, Deffernez C, Kuehni CE, Latzin P, et al. Viral Aetiology of Acute Respiratory Infections With Cough in Infancy. Pediatr Infect Dis J [Internet]. 2008 Jan [cited 2019 Feb 12];PAP(2):100–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18174876
6. Greenberg SB. Respiratory viral infections in adults. Curr Opin Pulm Med [Internet]. 2002 May [cited 2019 Feb 12];8(3):201–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11981309
7. Sloots TP, Whiley DM, Lambert SB, Nissen MD. Emerging respiratory agents: New

viruses for old diseases? J Clin Virol [Internet]. 2008 Jul [cited 2019 Feb 12];42(3):233–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18406664

8. van der Zalm MM, van Ewijk BE, Wilbrink B, Uiterwaal CSPM, Wolfs TFW, van der Ent CK. Respiratory Pathogens in Children with and without Respiratory Symptoms. J Pediatr [Internet]. 2009 Mar [cited 2019 Feb 12];154(3):396-400.e1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18823911

9. GBD 2016 Lower Respiratory Infections Collaborators C, Blacker B, Khalil IA, Rao PC, Cao J, Zimsen SRM, et al. Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Infect Dis [Internet]. 2018 Nov 1 [cited 2019 Feb 11];18(11):1191–210. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30243584

10. Iuliano AD, Roguski KM, Chang HH, Muscatello DJ, Palekar R, Tempia S, et al. Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. Lancet (London, England) [Internet]. 2018 Mar 31 [cited 2019 Feb 12];391(10127):1285–300. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29248255

11. Shrestha S, King AA, Rohani P. Statistical Inference for Multi-Pathogen Systems. Alizon S, editor. PLoS Comput Biol [Internet]. 2011 Aug 18 [cited 2018 Dec 2];7(8):e1002135. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002135

12. Walzl G, Tafuro S, Moss P, Openshaw PJ, Hussell T. Influenza virus lung infection protects from respiratory syncytial virus-induced immunopathology. J Exp Med [Internet]. 2000 Nov 6 [cited 2019 Feb 12];192(9):1317–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11067880

13. Mallia P, Johnston SL. Influenza infection and COPD. Int J Chron Obstruct Pulmon Dis [Internet]. 2007 [cited 2019 Feb 12];2(1):55–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18044066

14. Kutter JS, Spronken MI, Fraaij PL, Fouchier RA. Transmission routes of respiratory viruses among humans. Curr Opin Virol [Internet]. 2018 Feb 1 [cited 2019 Feb 12];28:142–51. Available from: https://www.sciencedirect.com/science/article/pii/S1879625717301773

15. Fernstrom A, Goldblatt M. Aerobiology and its role in the transmission of infectious diseases. J Pathog [Internet]. 2013 [cited 2019 Feb 12];2013:493960. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23365758

16. Atkinson J, Chartier Y, Pessoa-Silva CL, Jensen P, Li Y, Seto WH editors. Natural Ventilation for Infection Control in Health-Care Settings - PubMed - NCBI [Internet]. 2009 [cited 2019 Feb 12]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23762969

17. Tatem AJ, Rogers DJ, Hay SI. Global transport networks and infectious disease spread. Adv Parasitol [Internet]. 2006 [cited 2019 Feb 12];62:293–343. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16647974

18. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. Ferguson NM, editor. PLoS Pathog [Internet]. 2014 Feb 20 [cited 2019 Feb 5];10(2):e1003932. Available from: https://dx.plos.org/10.1371/journal.ppat.1003932

19. Bahl J, Nelson MI, Chan KH, Chen R, Vijaykrishna D, Halpin RA, et al. Temporally

structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. Proc Natl Acad Sci [Internet]. 2011 Nov 29 [cited 2019 Feb 7];108(48):19359–64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22084096

20. Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature [Internet]. 2015 Jun 8 [cited 2019 Feb 7];523(7559):217. Available from: http://www.nature.com/articles/nature14460

21. Bedford T, Cobey S, Beerli P, Pascual M. Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). Ferguson NM, editor. PLoS Pathog [Internet]. 2010 May 27 [cited 2018 Dec 2];6(5):e1000918. Available from: https://dx.plos.org/10.1371/journal.ppat.1000918

22. Dalziel BD, Kissler S, Gog JR, Viboud C, Bjørnstad ON, Metcalf CJE, et al. Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. Science [Internet]. 2018 Oct 5 [cited 2019 Jan 24];362(6410):75–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30287659

23. Meng J, Stobart CC, Hotard AL, Moore ML. An Overview of Respiratory Syncytial Virus. Racaniello V, editor. PLoS Pathog [Internet]. 2014 Apr 24 [cited 2019 Feb 12];10(4):e1004016. Available from: http://dx.plos.org/10.1371/journal.ppat.1004016

24. Cobey S, Hensley SE. Immune history and influenza virus susceptibility. Curr Opin Virol [Internet]. 2017 [cited 2019 Feb 12];22:105–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28088686

25. Nectunt. Mechanisms of Evolutionary Change [Internet]. [cited 2019 Oct 3]. Available from: http://nectunt.bifi.es/to-learn-more-overview/mechanisms-of-evolutionary-change/

26. Rabadan R, Robins H. Evolution of the influenza a virus: some new advances. Evol Bioinform Online [Internet]. 2008 Jan 30 [cited 2019 Oct 3];3:299–307. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19430605

27. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci [Internet]. 2016 [cited 2019 Feb 13];73(23):4433–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27392606

28. Volz EM, Koelle K, Bedford T. Viral Phylodynamics. Wodak S, editor. PLoS Comput Biol [Internet]. 2013 Mar 21 [cited 2018 Dec 12];9(3):e1002947. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002947

29. He L, Zhu J. Computational tools for epitope vaccine design and evaluation. Curr Opin Virol [Internet]. 2015 Apr [cited 2019 Jan 21];11:103–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25837467

30. Nascimento FF, Reis M Dos, Yang Z. A biologist's guide to Bayesian phylogenetic analysis. Nat Ecol Evol [Internet]. 2017 Oct [cited 2019 Mar 8];1(10):1446–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28983516

31. Moghadami M. A Narrative Review of Influenza: A Seasonal and Pandemic Disease. Iran J Med Sci [Internet]. 2017 Jan [cited 2019 Feb 12];42(1):2–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28293045

32. Bouvier NM, Palese P. The biology of influenza viruses. Vaccine [Internet]. 2008 Sep 12 [cited 2019 Feb 12];26:D49–53. Available from: https://www.sciencedirect.com/science/article/pii/S0264410X08009377

33. Ferguson L, Olivier AK, Genova S, Epperson WB, Smith DR, Schneider L, et al. Pathogenesis of Influenza D Virus in Cattle. J Virol [Internet]. 2016 Jun 15 [cited 2019

Feb 12];90(12):5636–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27030270

34. Mosnier A, Caini S, Daviaud I, Nauleau E, Bui TT, Debost E, et al. Clinical Characteristics Are Similar across Type A and B Influenza Virus Infections. Schanzer DL, editor. PLoS One [Internet]. 2015 Sep 1 [cited 2019 Feb 12];10(9):e0136186. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26325069

35. Taubenberger JK, Morens DM. Influenza: the once and future pandemic. Public Health Rep [Internet]. 2010 Apr [cited 2019 Feb 12];125 Suppl 3(Suppl 3):16–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20568566

36. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev [Internet]. 1992 Mar 1 [cited 2019 Feb 12];56(1):152–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1579108

37. Chen X, Liu S, Goraya MU, Maarouf M, Huang S, Chen J-L. Host Immune Response to Influenza A Virus Infection. Front Immunol [Internet]. 2018 [cited 2019 Feb 12];9:320. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29556226

38. Wu Y, Wu Y, Tefsen B, Shi Y, Gao GF. Bat-derived influenza-like viruses H17N10 and H18N11. Trends Microbiol [Internet]. 2014 Apr [cited 2019 Feb 12];22(4):183–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24582528

39. Te Velthuis AJW, Fodor E. Influenza virus RNA polymerase: Insights into the mechanisms of viral RNA synthesis. Vol. 14, Nature Reviews Microbiology. Nature Publishing Group; 2016. p. 479–93.

40. Pielak RM, Chou JJ. Influenza M2 proton channels. Vol. 1808, Biochimica et Biophysica Acta - Biomembranes. 2011. p. 522–9.

41. Palese P. The genes of influenza virus. Cell [Internet]. 1977 Jan 1 [cited 2019 Feb

12];10(1):1–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/837439

42. Kanegae Y, Sugita S, Endo A, Ishida M, Senya S, Osako K, et al. Evolutionary pattern of the hemagglutinin gene of influenza B viruses isolated in Japan: cocirculating lineages in the same epidemic season. J Virol [Internet]. 1990 Jun [cited 2019 Feb 12];64(6):2860–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2335820

43. Cox NJ, Subbarao K. Global epidemiology of influenza: past and present. Annu Rev Med. 2000;

44. Kuiken T, Holmes EC, McCauley J, Rimmelzwaan GF, Williams CS, Grenfell BT. Host species barriers to influenza virus infections. Science [Internet]. 2006 Apr 21 [cited 2018 Dec 2];312(5772):394–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16627737

45. Webby RJ, Webster RG. Are we ready for pandemic influenza? Science. 2003 Nov;302(5650):1519–22.

46. Rambaut A, Pybus OG, Nelson MI, Viboud C, Taubenberger JK, Holmes EC. The genomic and epidemiological dynamics of human influenza A virus. Nature [Internet]. 2008 May 29 [cited 2019 Feb 4];453(7195):615–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18418375

47. Smith GJD, Bahl J, Vijaykrishna D, Zhang J, Poon LLM, Chen H, et al. Dating the emergence of pandemic influenza viruses. Proc Natl Acad Sci U S A. 2009;106:11709–12.

48. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature. 2009;459(7250):1122–5.

49. Fuller TL, Gilbert M, Martin V, Cappelle J, Hosseini P, Njabo KY, et al. Predicting hotspots for influenza virus reassortment. Emerg Infect Dis. 2013 Apr;19(4):581–8.

50. Guan Y, Vijaykrishna D, Bahl J, Zhu H, Wang J, Smith GJD. The emergence of pandemic influenza viruses. Protein Cell [Internet]. 2010 Jan 7 [cited 2019 Feb 12];1(1):9–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21203993

51. Zhou NN, Senne DA, Landgraf JS, Swenson SL, Erickson G, Rossow K, et al. Genetic reassortment of avian, swine, and human influenza A viruses in American pigs. J Virol [Internet]. 1999 Oct [cited 2019 Feb 13];73(10):8851–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10482643

52. Finkelman BS, Viboud C, Koelle K, Ferrari MJ, Bharti N, Grenfell BT. Global Patterns in Seasonal Activity of Influenza A/H3N2, A/H1N1, and B from 1997 to 2005: Viral Coexistence and Latitudinal Gradients. Myer L, editor. PLoS One [Internet]. 2007 Dec 12 [cited 2019 Feb 12];2(12):e1296. Available from: https://dx.plos.org/10.1371/journal.pone.0001296

53. Shi T, McAllister DA, O'Brien KL, Simoes EAF, Madhi SA, Gessner BD, et al. Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. Lancet (London, England) [Internet]. 2017 Sep 2 [cited 2019 Feb 12];390(10098):946–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28689664

54. Erbelding EJ, Post DJ, Stemmy EJ, Roberts PC, Augustine AD, Ferguson S, et al. A Universal Influenza Vaccine: The Strategic Plan for the National Institute of Allergy and Infectious Diseases. J Infect Dis [Internet]. 2018 Jul 2 [cited 2019 Jan 21];218(3):347–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29506129

55. The World Health Organization. Influenza at the human-animal interface [Internet]. 2019 [cited 2019 Feb 13]. Available from: https://www.who.int/influenza/human_animal_interface/Influenza_Summary_IRA_HA_interface_21_01_2019.pdf?ua=1

56. Cowling BJ, Jin L, Lau EH, Liao Q, Wu P, Jiang H, et al. Comparative epidemiology of human infections with avian influenza A H7N9 and H5N1 viruses in China: a population-based study of laboratory-confirmed cases. Lancet [Internet]. 2013 Jul 13 [cited 2019 Feb 13];382(9887):129–37. Available from: https://linkinghub.elsevier.com/retrieve/pii/S014067361361171X

57. Katz JM, Veguilla V, Belser JA, Maines TR, Van Hoeven N, Pappas C, et al. The public health impact of avian influenza viruses. Poult Sci [Internet]. 2009 Apr 1 [cited 2019 Feb 13];88(4):872–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19276438

58. Fitzpatrick A, Mor SK, Thurn M, Wiedenman E, Otterson T, Porter RE, et al. Outbreak of highly pathogenic avian influenza in Minnesota in 2015. J Vet Diagnostic Investig [Internet]. 2017 Mar 8 [cited 2019 Feb 13];29(2):169–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28176609

59. Zaraket H, Saito R, Sato I, Suzuki Y, Li D, Dapat C, et al. Molecular evolution of human influenza A viruses in a local area during eight influenza epidemics from 2000 to 2007. Arch Virol [Internet]. 2009 Feb 20 [cited 2019 Feb 13];154(2):285–95. Available from: http://link.springer.com/10.1007/s00705-009-0309-9

60. Margine I, Hai R, Albrecht RA, Obermoser G, Harrod AC, Banchereau J, et al. H3N2 Influenza Virus Infection Induces Broadly Reactive Hemagglutinin Stalk Antibodies in Humans and Mice. J Virol [Internet]. 2013 Apr 15 [cited 2019 Feb 17];87(8):4728–37.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/23408625

61. Shao W, Li X, Goraya MU, Wang S, Chen J-L. Evolution of Influenza A Virus by Mutation and Re-Assortment. Int J Mol Sci [Internet]. 2017 Aug 7 [cited 2019 Mar 11];18(8). Available from: http://www.ncbi.nlm.nih.gov/pubmed/28783091

62. Riwilaijaroen BNS, Uzuki YS. Review Molecular basis of the structure and function of H1 hemagglutinin of in fl uenza virus. 2012;88:226–49.

63. Wiley DC, Skehel JJ. The Structure and Function of the Hemagglutinin Membrane Glycoprotein of Influenza Virus. Annu Rev Biochem [Internet]. 1987 Jun 1 [cited 2019 Feb 13];56(1):365–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3304138

64. Gamblin SJ, Skehel JJ. Influenza Hemagglutinin and Neuraminidase Membrane Glycoproteins. J Biol Chem [Internet]. 2010 Sep 10 [cited 2019 Feb 13];285(37):28403–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20538598

65. Sautto GA, Kirchenbaum GA, Ross TM. Towards a universal influenza vaccine: different approaches for one goal. Virol J [Internet]. 2018 Dec 19 [cited 2019 Feb 13];15(1):17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29370862

66. Su YCF, Bahl J, Joseph U, Butt KM, Peck HA, Koay ESC, et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. Nat Commun [Internet]. 2015 Aug 6 [cited 2019 Feb 13];6:7952. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26245473

67. Rémy V, Zöllner Y, Heckmann U. Vaccination: the cornerstone of an efficient healthcare system. J Mark access Heal policy [Internet]. 2015 [cited 2019 Mar 11];3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27123189

68. CDC. Pinkbook | Influenza | Epidemiology of Vaccine Preventable Diseases | CDC

[Internet]. 2019 [cited 2019 Mar 11]. Available from: https://www.cdc.gov/vaccines/pubs/pinkbook/flu.html

69. Hannoun C. The evolving history of influenza viruses and influenza vaccines. Expert Rev Vaccines [Internet]. 2013 Sep 9 [cited 2019 Mar 11];12(9):1085–94. Available from: http://www.tandfonline.com/doi/full/10.1586/14760584.2013.824709

70. Hampson AW. Vaccines for pandemic influenza. The history of our current vaccines, their limitations and the requirements to deal with a pandemic threat. Ann Acad Med Singapore [Internet]. 2008 Jun [cited 2019 Mar 11];37(6):510–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18618064

71. Barberis I, Myles P, Ault SK, Bragazzi NL, Martini M. History and evolution of influenza control through vaccination: from the first monovalent vaccine to universal vaccines. J Prev Med Hyg [Internet]. 2016 [cited 2019 Mar 11];57(3):E115–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27980374

72. Keitel WA, Neuzil KM, Treanor J. Immunogenicity, efficacy of inactivated/live virus seasonal and pandemic vaccines. In: Textbook of Influenza [Internet]. Oxford, UK: John Wiley & Sons, Ltd; 2013 [cited 2019 Mar 11]. p. 311–26. Available from: http://doi.wiley.com/10.1002/9781118636817.ch20

73. Weir JP, Gruber MF. An overview of the regulation of influenza vaccines in the United States. Influenza Other Respi Viruses [Internet]. 2016 Sep [cited 2019 Mar 11];10(5):354–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27426005

74. FDA. Approved Products - June 17, 2003 Approval Letter - Influenza Virus Vaccine Live, Intranasal [Internet]. FDA. Center for Biologics Evaluation and Research; 2003 [cited 2019 Mar 11]. Available from: http://wayback.archive-

it.org/7993/20170723030839/https:/www.fda.gov/BiologicsBloodVaccines/Vaccines/ApprovedProducts/ucm123753.htm

75. Tisa V, Barberis I, Faccio V, Paganino C, Trucchi C, Martini M, et al. Quadrivalent influenza vaccine: a new opportunity to reduce the influenza burden. J Prev Med Hyg [Internet]. 2016 [cited 2019 Mar 11];57(1):E28-33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27346937

76. Centers for Disease Control and Prevention. Selecting Viruses for the Seasonal Influenza Vaccine | CDC [Internet]. 2018 [cited 2019 Feb 17]. Available from: https://www.cdc.gov/flu/about/season/vaccine-selection.htm

77. Centers for Disease Control and Prevention. Antigenic Characterization | CDC [Internet]. 2017 [cited 2019 Feb 17]. Available from: https://www.cdc.gov/flu/professionals/laboratory/antigenic.htm

78. Anderson CS, Ortega S, Chaves FA, Clark AM, Yang H, Topham DJ, et al. Natural and directed antigenic drift of the H1 influenza virus hemagglutinin stalk domain. Sci Rep [Internet]. 2017 Nov 6 [cited 2019 Feb 17];7(1):14614. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29097696

79. Berlanda Scorza F, Tsvetnitsky V, Donnelly JJ. Universal influenza vaccines: Shifting to better vaccines. Vaccine [Internet]. 2016 Jun 3 [cited 2019 Feb 17];34(26):2926–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27038130

80. Skowronski DM, Janjua NZ, De Serres G, Sabaiduc S, Eshaghi A, Dickinson JA, et al. Low 2012–13 Influenza Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3N2 Vaccine Strain Not Antigenic Drift in Circulating Viruses. Kobinger GP, editor. PLoS One [Internet]. 2014 Mar 25 [cited 2019 Feb 17];9(3):e92153. Available

from: https://dx.plos.org/10.1371/journal.pone.0092153

81. Zost SJ, Parkhouse K, Gumina ME, Kim K, Diaz Perez S, Wilson PC, et al. Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. Proc Natl Acad Sci U S A [Internet]. 2017 Nov 21 [cited 2019 Feb 17];114(47):12578–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29109276

82. Wu NC, Zost SJ, Thompson AJ, Oyen D, Nycholat CM, McBride R, et al. A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine. Palese P, editor. PLOS Pathog [Internet]. 2017 Oct 23 [cited 2019 Feb 17];13(10):e1006682. Available from: https://dx.plos.org/10.1371/journal.ppat.1006682

83. Paules CI, Sullivan SG, Subbarao K, Fauci AS. Chasing Seasonal Influenza — The Need for a Universal Influenza Vaccine. N Engl J Med [Internet]. 2018 Jan 4 [cited 2019 Feb 17];378(1):7–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29185857

84. CDC. Seasonal Influenza Vaccine Effectiveness, 2004-2018 [Internet]. 2018. Available from: https://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm

85. Neumann G, Noda T, Kawaoka Y. Emergence and pandemic potential of swine-origin H1N1 influenza virus. Nature [Internet]. 2009 Jun 18 [cited 2019 Feb 27];459(7249):931–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19525932

86. Carter DM, Darby CA, Lefoley BC, Crevar CJ, Alefantis T, Oomen R, et al. Design and Characterization of a Computationally Optimized Broadly Reactive Hemagglutinin Vaccine for H1N1 Influenza Viruses. Lyles DS, editor. J Virol [Internet]. 2016 May 1 [cited 2019 Feb 17];90(9):4720–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26912624

87. Job ER, Ysenbaert T, Smet A, Christopoulou I, Strugnell T, Oloo EO, et al. Broadened immunity against influenza by vaccination with computationally designed influenza virus N1 neuraminidase constructs. npj Vaccines [Internet]. 2018 Dec 29 [cited 2019 Feb 17];3(1):55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30510776

88. Holmes EC. Evolutionary history and phylogeography of human viruses. Annu Rev Microbiol. 2008 Jan;62:307–28.

89. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10(8):540–50.

90. Kühnert D, Wu CH, Drummond AJ. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. Infect Genet Evol. 2011;11(8):1825–41.

91. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. Salathé M, editor. PLOS Comput Biol [Internet]. 2015 Dec 30 [cited 2019 Mar 10];11(12):e1004613. Available from: https://dx.plos.org/10.1371/journal.pcbi.1004613

92. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol [Internet]. 1985 [cited 2019 Feb 20];22(2):160–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3934395

93. Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. J Theor Biol [Internet]. 1990 Feb 22 [cited 2019 Apr 18];142(4):485–501. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2338834

94. Abadi S, Azouri D, Pupko T, Mayrose I. Model selection may not be a mandatory step for phylogeny reconstruction. Nat Commun [Internet]. 2019 Dec 25 [cited 2019 Aug 5];10(1):934. Available from: http://www.nature.com/articles/s41467-019-08822-w

95. Ho SYW, Phillips MJ. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. Syst Biol [Internet]. 2009 [cited 2019 Mar 10];58(3):367–80. Available from: https://biology.columbian.gwu.edu/sites/g/files/zaxdzs1961/f/downloads/Ho_&_Phillips_2009_Calibration_Uncertainty.pdf

96. Pybus OG. Model Selection and the Molecular Clock. PLoS Biol [Internet]. 2006 May 16 [cited 2019 Mar 10];4(5):e151. Available from: https://dx.plos.org/10.1371/journal.pbio.0040151

97. Ferreira MAR, Suchard MA. Bayesian analysis of elapsed times in continuous-time Markov chains. Can J Stat [Internet]. 2008 Sep 1 [cited 2019 Apr 24];36(3):355–68. Available from: http://doi.wiley.com/10.1002/cjs.5550360302

98. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. Syst Biol [Internet]. 2017 Jun 1 [cited 2018 Dec 2];66(1):e47–65. Available from: https://academic.oup.com/sysbio/article/doi/10.1093/sysbio/syw054/2670001/Emerging-concepts-of-data-integration-in-pathogen

99. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. Penny D, editor. PLoS Biol [Internet]. 2006 Mar 14 [cited 2019 Mar 8];4(5):e88. Available from: https://dx.plos.org/10.1371/journal.pbio.0040088

100. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-) [Internet]. 2004 Jan 16 [cited 2018 Dec 12];303(5656):327–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14726583

101. Gattepaille L, Günther T, Jakobsson M. Inferring Past Effective Population Size from Distributions of Coalescent Times. Genetics [Internet]. 2016 [cited 2019 Mar 10];204(3):1191–206. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27638421

102. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V. Improving the accuracy of demographic and molecular clock model comparison while accomodating phylogenetic uncertainty. Mol Biol Evol [Internet]. 2012 Jan; Available from: http://www.ncbi.nlm.nih.gov/pubmed/17065594

103. Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika [Internet]. 1970 Apr [cited 2019 Mar 10];57(1):97. Available from: https://www.jstor.org/stable/2334940?origin=crossref

104. Faria NR, Suchard MA, Rambaut A, Lemey P. Toward a quantitative understanding of viral phylogeography. Curr Opin Virol [Internet]. 2011 Nov [cited 2019 Mar 11];1(5):423–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22440846

105. Bloomquist EW, Lemey P, Suchard MA. Three roads diverged? Routes to phylogeographic inference. Trends Ecol Evol [Internet]. 2010 Nov [cited 2019 Mar 11];25(11):626–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20863591

106. Bahl J, Pham TT, Hill NJ, Hussein ITM, Ma EJ, Easterday BC, et al. Ecosystem Interactions Underlie the Spread of Avian Influenza A Viruses with Pandemic Potential. Ferguson NM, editor. PLOS Pathog [Internet]. 2016 May 11 [cited 2018 Dec 2];12(5):e1005620. Available from: https://dx.plos.org/10.1371/journal.ppat.1005620

107. Jin Y, Yu D, Ren H, Yin Z, Huang Z, Hu M, et al. Phylogeography of Avian influenza A H9N2 in China. BMC Genomics. 2014;15(1110).

108. Trovão NS, Suchard MA, Baele G, Gilbert M, Lemey P. Bayesian inference reveals host-

specific contributions to the epidemic expansion of Influenza A H5N1. Mol Biol Evol. 2015;

109. Lemey P, Rambaut A, Drummond AJ, Suchard M a. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009;5(9):e1000520.

110. Lam TT-Y, Zhou B, Wang J, Chai Y, Shen Y, Chen X, et al. Dissemination, divergence and establishment of H7N9 influenza viruses in China. Nature. 2015 Jun;522(7554):102–5.

111. Beard R, Magee D, Suchard MA, Lemey P, Scotch M. Generalized linear models for identifying predictors of the evolutionary diffusion of viruses. AMIA Jt Summits Transl Sci proceedings AMIA Jt Summits Transl Sci [Internet]. 2014 [cited 2018 Dec 2];2014:23–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25717395

112. Mu NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its Approximations. 2017;34(11):2970–81.

113. Vaughan TG, Kü Hnert D, Popinga A, Welch D, Drummond AJ. Phylogenetics Efficient Bayesian inference under the structured coalescent. 2014 [cited 2019 Nov 4];30(16):2272–9. Available from: http://compevol.github.io/MultiTypeTree.

114. De Maio N, Wu C-H, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. Pritchard JK, editor. PLOS Genet [Internet]. 2015 Aug 12 [cited 2019 Feb 21];11(8):e1005421. Available from: http://dx.plos.org/10.1371/journal.pgen.1005421

115. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife [Internet]. 2018 Jan 16 [cited 2018 Dec 12];7. Available from: https://elifesciences.org/articles/31257

116. Möller S, du Plessis L, Stadler T. Impact of the tree prior on estimating clock rates during epidemic outbreaks. Proc Natl Acad Sci U S A. 2018 Apr 17;115(16):4200–5.

117. Kirschner M, Gerhart J. Evolvability. Proc Natl Acad Sci U S A [Internet]. 1998 Jul 21 [cited 2019 Feb 24];95(15):8420–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9671692

118. Kirkpatrick E, Qiu X, Wilson PC, Bahl J, Krammer F. The influenza virus hemagglutinin head evolves faster than the stalk domain. Sci Rep [Internet]. 2018 Dec 11 [cited 2019 Feb 23];8(1):10432. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29992986

119. Sriwilaijaroen N, Suzuki Y. Molecular basis of the structure and function of H1 hemagglutinin of influenza virus. Proc Jpn Acad Ser B Phys Biol Sci [Internet]. 2012 [cited 2019 Mar 11];88(6):226–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22728439

120. Nachbagauer R, Miller MS, Hai R, Ryder AB, Rose JK, Palese P, et al. Hemagglutinin Stalk Immunity Reduces Influenza Virus Replication and Transmission in Ferrets. 2016;90(6):3268–73.

121. Wang TT, Palese P. Catching a Moving Target. Science (80-) [Internet]. 2011 Aug 12;333(6044):834 LP – 835. Available from: http://science.sciencemag.org/content/333/6044/834.abstract

122. Krammer F, Hai R, Yondola M, Tan GS, Leyva-Grado VH, Ryder AB, et al. Assessment of influenza virus hemagglutinin stalk-based immunity in ferrets. J Virol [Internet]. 2014 Mar 15 [cited 2017 Nov 16];88(6):3432–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24403585

123. Subbarao K, Matsuoka Y. The prospects and challenges of universal vaccines for

influenza. Trends Microbiol [Internet]. 2013 Nov 15;21(7):350–8. Available from: http://dx.doi.org/10.1016/j.tim.2013.04.003

124. Kmiecik S, Gront D, Kolinski M, Wieteska L, Dawid AE, Kolinski A. Coarse-Grained Protein Models and Their Applications. Chem Rev [Internet]. 2016 Jul 27 [cited 2019 Feb 23];116(14):7898–936. Available from: http://pubs.acs.org/doi/10.1021/acs.chemrev.6b00163

125. Kleinman CL, Rodrigue N, Lartillot N, Philippe H. Statistical Potentials for Improved Structurally Constrained Evolutionary Models. 2010 [cited 2019 Feb 17]; Available from: http://mbe.oxfordjournals.org/

126. Booker TR, Keightley PD. Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. Mol Biol Evol [Internet]. 2018 Oct 8 [cited 2019 Apr 28];35(12):2971–88. Available from: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msy188/5123518

127. Swanson KA, Settembre EC, Shaw CA, Dey AK, Rappuoli R, Mandl CW, et al. Structural basis for immunization with postfusion respiratory syncytial virus fusion F glycoprotein (RSV F) to elicit high neutralizing antibody titers. Proc Natl Acad Sci U S A [Internet]. 2011 Jun 7 [cited 2019 Mar 12];108(23):9619–24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21586636

128. Zhao X, Singh M, Malashkevich VN, Kim PS. Structural characterization of the human respiratory syncytial virus fusion protein core. Proc Natl Acad Sci. 2000 Dec 19;97(26):14172–7.

129. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, et al. High-resolution mapping of protein sequence-function relationships. Nat Methods [Internet].

2010 Sep 15 [cited 2019 Feb 23];7(9):741–6. Available from: http://www.nature.com/articles/nmeth.1492

130. Araya CL, Fowler DM. Deep mutational scanning: assessing protein function on a massive scale. Trends Biotechnol [Internet]. 2011 Sep [cited 2019 Feb 23];29(9):435–42. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167779911000692

131. Traxlmayr MW, Hasenhindl C, Hackl M, Stadlmayr G, Rybka JD, Borth N, et al. Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing. J Mol Biol [Internet]. 2012 Oct 26 [cited 2019 Feb 23];423(3):397–412. Available from: https://www.sciencedirect.com/science/article/pii/S0022283612005840

132. Melamed D, Young DL, Gamble CE, Miller CR, Fields S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA [Internet]. 2013 Nov 24 [cited 2019 Feb 23];19(11):1537–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24064791

133. Roscoe BP, Thayer KM, Zeldovich KB, Fushman D, Bolon DNA. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. J Mol Biol [Internet]. 2013 Apr 26 [cited 2019 Feb 23];425(8):1363–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23376099

134. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol [Internet]. 2014 Jun [cited 2019 Feb 23];31(6):1581–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24567513

135. Bloom JDJD. An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit. Mol Biol Evol [Internet]. 2014 Aug 1 [cited 2019 Feb 23];31(8):1956–

78. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu173

136. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol Biol Evol [Internet]. 2012 Jun 1 [cited 2019 Feb 20];29(6):1695–701. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss020

137. Shapiro B, Rambaut A, Drummond AJ. Choosing Appropriate Substitution Models for the Phylogenetic Analysis of Protein-Coding Sequences. Mol Biol Evol [Internet]. 2006 Jan 1;23(1):7–9. Available from: http://dx.doi.org/10.1093/molbev/msj021

138. Influenza (Seasonal) [Internet]. [cited 2019 Feb 12]. Available from: https://www.who.int/en/news-room/fact-sheets/detail/influenza-(seasonal)

139. Nair H, Brooks WA, Katz M, Roca A, Berkley JA, Madhi SA, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. Lancet [Internet]. 2011 Dec 3 [cited 2019 Mar 12];378(9807):1917–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22078723

140. Nakapan S, Tripathi NK, Tipdecho T, Souris M. Spatial diffusion of influenza outbreak-related climate factors in Chiang Mai Province, Thailand. Int J Environ Res Public Health [Internet]. 2012 Oct 24 [cited 2019 Mar 12];9(11):3824–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23202819

141. Nsoesie EO, Brownstein JS, Ramakrishnan N, Marathe M V. A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza Other Respi Viruses [Internet]. 2014 May 1 [cited 2019 Feb 13];8(3):309–16. Available from: http://doi.wiley.com/10.1111/irv.12226

142. Magee D, Suchard MA, Scotch M. Bayesian phylogeography of influenza A/H3N2 for the 2014-15 season in the United States using three frameworks of ancestral state reconstruction. Koelle K, editor. PLOS Comput Biol [Internet]. 2017 Feb 7 [cited 2019 Aug 7];13(2):e1005389. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28170397

143. Pica N, Bouvier NM. Environmental factors affecting the transmission of respiratory viruses. Curr Opin Virol [Internet]. 2012 Feb [cited 2019 Mar 12];2(1):90–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22440971

144. Fuhrmann C. The Effects of Weather and Climate on the Seasonality of Influenza: What We Know and What We Need to Know. Geogr Compass [Internet]. 2010 [cited 2019 Mar 12];4(7):718–30. Available from: http://www.sercc.com/FuhrmannGeogCompassFlu.pdf

145. Roussel M, Pontier D, Cohen J-M, Lina B, Fouchet D. Linking influenza epidemic onsets to covariates at different scales using a dynamical model. PeerJ [Internet]. 2018 [cited 2019 Mar 12];6:e4440. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29568702

146. Moriarty LF, Omer SB. Infants and the seasonal influenza vaccine. A global perspective on safety, effectiveness, and alternate forms of protection. Hum Vaccin Immunother [Internet]. 2014 [cited 2019 Mar 12];10(9):2721–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25483664

147. Bouvier NM. The Future of Influenza Vaccines: A Historical and Clinical Perspective. Vaccines [Internet]. 2018 Aug 30 [cited 2019 Mar 12];6(3). Available from: http://www.ncbi.nlm.nih.gov/pubmed/30200179

148. Chong YL, Padhi A, Hudson PJ, Poss M. The effect of vaccination on the evolution and population dynamics of avian paramyxovirus-1. PLoS Pathog [Internet]. 2010 Apr 22 [cited 2019 Mar 12];6(4):e1000872. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/20421950

149. Azarian T, Grant LR, Arnold BJ, Hammitt LL, Reid R, Santosham M, et al. The impact of serotype-specific vaccination on phylodynamic parameters of Streptococcus pneumoniae and the pneumococcal pan-genome. Tang C, editor. PLOS Pathog [Internet]. 2018 Apr 4 [cited 2018 Dec 18];14(4):e1006966. Available from: https://dx.plos.org/10.1371/journal.ppat.1006966

150. Cobey S. Pathogen evolution and the immunological niche. Ann N Y Acad Sci [Internet]. 2014 Jul [cited 2019 Mar 12];1320(1):1–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25040161

151. Read AF, Baigent SJ, Powers C, Kgosana LB, Blackwell L, Smith LP, et al. Imperfect Vaccination Can Enhance the Transmission of Highly Virulent Pathogens. PLoS Biol [Internet]. 2015 [cited 2019 Mar 12];13(7):e1002198. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26214839

152. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? Parrish C, editor. PLOS Pathog [Internet]. 2018 Feb 8 [cited 2019 Jun 29];14(2):e1006885. Available from: https://dx.plos.org/10.1371/journal.ppat.1006885

153. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends Ecol Evol [Internet]. 2015 Jun [cited 2019 Jan 9];30(6):306–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25887947

154. Lanave C, Preparata G, Sacone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol [Internet]. 1984;20(1):86–93. Available from: https://doi.org/10.1007/BF02101990

155. Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Am Math Soc Lect Math Life Sci [Internet]. 1986;17:57–86. Available from: citeulike-article-id:6732633

156. Koonin E, Galperin M. Evolutionary Concept in Genetics and Genomics. In: Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003.

157. Krammer F, Palese P. Influenza virus hemagglutinin stalk-based antibodies and vaccines. Curr Opin Virol [Internet]. 2013;3(5):521–30. Available from: http://www.sciencedirect.com/science/article/pii/S187962571300134X

158. Qiu X, Duvvuri VR, Gubbay JB, Webby RJ, Kayali G, Bahl J. specific epitope profiles for HPAI H5 pre- - pandemic vaccine selection and evaluation. 2017;(July):445–56.

159. Vijaykrishna D, Holmes EC, Joseph U, Fourment M, Su YCF, Halpin R, et al. The contrasting phylodynamics of human influenza B viruses. Elife. 2015;4:1–23.

160. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol [Internet]. 2012 Aug [cited 2019 Feb 5];29(8):1969–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22367748

161. Minin VN, Bloomquist EW, Suchard MA. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. Mol Biol Evol [Internet]. 2008 Apr 3 [cited 2019 Feb 20];25(7):1459–71. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msn090

162. Lartillot N, Philippe H, Lewis P. Computing Bayes Factors Using Thermodynamic Integration. Syst Biol [Internet]. 2006 Apr 1;55(2):195–207. Available from: http://dx.doi.org/10.1080/10635150500433722

163. Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. Improving Marginal Likelihood Estimation for Bayesian Phylogenetic Model Selection. Syst Biol [Internet]. 2011 Mar 1 [cited 2019 May 5];60(2):150–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21187451

164. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc [Internet]. 1995 Jun 1;90(430):773–95. Available from: http://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

165. Mintaev RR, Alexeevski A V, Kordyukova L V. Co-evolution analysis to predict protein–protein interactions within influenza virus envelope. J Bioinform Comput Biol [Internet]. 2014 Mar 6;12(02):1441008. Available from: https://doi.org/10.1142/S021972001441008X

166. Li C, Lu G, Ortí G. Optimal Data Partitioning and a Test Case for Ray-Finned Fishes (Actinopterygii) Based on Ten Nuclear Loci. Buckley T, editor. Syst Biol [Internet]. 2008 Aug 1 [cited 2019 Apr 18];57(4):519–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18622808

167. Brandley MC, Schmitz A, Reeder TW. Partitioned Bayesian Analyses, Partition Choice, and the Phylogenetic Relationships of Scincid Lizards. Anderson F, editor. Syst Biol [Internet]. 2005 Jun 1 [cited 2019 Apr 18];54(3):373–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16012105

168. McGuire JA, Witt CC, Altshuler DL, Remsen J V. Phylogenetic Systematics and Biogeography of Hummingbirds: Bayesian and Maximum Likelihood Analyses of Partitioned Data and Selection of an Appropriate Partitioning Strategy. Zamudio K, Sullivan J, editors. Syst Biol [Internet]. 2007 Oct 1 [cited 2019 Apr 18];56(5):837–56. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17934998

169. Kordyukova L. Structural and functional specificity of Influenza virus haemagglutinin and paramyxovirus fusion protein anchoring peptides. Virus Res [Internet]. 2017;227(Supplement C):183–99. Available from: http://www.sciencedirect.com/science/article/pii/S0168170216304063

170. Lu X, Shi Y, Gao F, Xiao H, Wang M, Qi J, et al. Insights into Avian Influenza Virus Pathogenicity: the Hemagglutinin Precursor HA0 of Subtype H16 Has an Alpha-Helix Structure in Its Cleavage Site with Inefficient HA1/HA2 Cleavage. J Virol [Internet]. 2012 Dec 25;86(23):12861–70. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3497694/

171. Skehel JJ, Wiley DC. Receptor Binding and Membrane Fusion in Virus Entry: The Influenza Hemagglutinin. Annu Rev Biochem [Internet]. 2000 Jun 1;69(1):531–69. Available from: https://doi.org/10.1146/annurev.biochem.69.1.531

172. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty Research article. 2012;29(9):2157–67.

173. Raftery A, Newton M, Satagopan J, Krivitsky P. Estimating the Integrated Likelihood via Posterior Simulation Using the Harmonic Mean Identity. Meml Sloan-Kettering Cancer Center, Dept Epidemiol Biostat Work Pap Ser [Internet]. 2006 Apr 14 [cited 2017 Nov 16]; Available from: http://biostats.bepress.com/mskccbiostat/paper6

174. Baele G, Lok W, Li S, Drummond AJ, Suchard MA, Lemey P. Accurate Model Selection of Relaxed Molecular Clocks in Bayesian Phylogenetics Letter Fast Track. 2012;30(2):239–43.

175. Kandeil A, Sabir JSM, Abdelaal A, Mattar EH, El-Taweel AN, Sabir MJ, et al. Efficacy of

commercial vaccines against newly emerging avian influenza H5N8 virus in Egypt. Sci Rep. 2018 Dec 1;8(1).

176. Hamid Samaha MP. Avian influenza vaccination in Egypt: Limitations of the current strategy. J Mol Genet Med. 2009;03(02).

177. Spielman SJ, Wilke CO. The relationship between dN/dS and scaled selection coefficients. Mol Biol Evol [Internet]. 2015 Apr [cited 2019 Jul 19];32(4):1097–108. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25576365

178. Borchers AT, Chang C, Gershwin ME, Gershwin LJ. Respiratory Syncytial Virus—A Comprehensive Review. Clin Rev Allerg Immunol [Internet]. 2013;45(3):331–79. Available from: https://doi.org/10.1007/s12016-013-8368-9

179. Pale M, Nacoto A, Tivane A, Nguenha N, Machalele L, Gundane F, et al. Respiratory syncytial and influenza viruses in children under 2 years old with severe acute respiratory infection (SARI) in Maputo, 2015. Cormier SA, editor. PLoS One [Internet]. 2017 Nov 30 [cited 2019 Feb 11];12(11):e0186735. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29190684

180. Karron RA, Black RE. Determining the burden of respiratory syncytial virus disease: the known and the unknown. Lancet (London, England) [Internet]. 2017 Sep 2 [cited 2019 Feb 12];390(10098):917–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28689665

181. Griffiths C, Drews SJ, Marchant DJ. Respiratory Syncytial Virus: Infection, Detection, and New Options for Prevention and Treatment. Clin Microbiol Rev [Internet]. 2017 Jan 1 [cited 2019 Aug 10];30(1):277–319. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27903593

182. Collins PL, Fearns R, Graham BS. Respiratory syncytial virus: virology, reverse genetics, and pathogenesis of disease. Curr Top Microbiol Immunol [Internet]. 2013 [cited 2019 Aug 10];372:3–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24362682

183. Hurwitz JL. Respiratory syncytial virus vaccine development. Expert Rev Vaccines [Internet]. 2011 Oct [cited 2019 Feb 17];10(10):1415–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21988307

184. Goto-Sugai K, Tsukagoshi H, Mizuta K, Matsuda S, Noda M, Sugai T, et al. Genotyping and phylogenetic analysis of the major genes in respiratory syncytial virus isolated from infants with bronchiolitis. Jpn J Infect Dis. 2010;63(6):393–400.

185. Martinelli M, Frati ER, Zappa A, Ebranati E, Bianchi S, Pariani E, et al. Phylogeny and population dynamics of respiratory syncytial virus (Rsv) A and B. Virus Res [Internet]. 2014;189:293–302. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0168170214002494

186. McLellan JS, Ray WC, Peeples ME. Structure and function of respiratory syncytial virus surface glycoproteins. Curr Top Microbiol Immunol [Internet]. 2013 [cited 2019 Jun 9];372:83–104. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24362685

187. Hause AM, Henke DM, Avadhanula V, Shaw CA, Tapia LI, Piedra PA. Sequence variability of the respiratory syncytial virus (RSV) fusion gene among contemporary and historical genotypes of RSV/A and RSV/B. Tregoning JS, editor. PLoS One [Internet]. 2017 Apr 17 [cited 2019 Aug 10];12(4):e0175792. Available from: https://dx.plos.org/10.1371/journal.pone.0175792

188. Zhao X, Singh M, Malashkevich VN, Kim PS. Structural characterization of the human respiratory syncytial virus fusion protein core. Proc Natl Acad Sci [Internet]. 2000 Dec 19

[cited 2019 Aug 10];97(26):14172–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11106388

189. WHO | WHO Meeting to Launch Phase-2 of the RSV Surveillance Pilot Based on the Global Influenza Surveillance and Response System. WHO [Internet]. 2019 [cited 2019 Aug 10]; Available from: https://www.who.int/influenza/rsv/who_rsv_surveillance_2nd_phase/en/

190. CDC. Disease Burden of Influneza [Internet]. [cited 2019 Apr 18]. Available from: https://www.cdc.gov/flu/about/burden/index.html

191. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res [Internet]. 2004 [cited 2019 Feb 5];32(5):1792–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15034147

192. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics [Internet]. 2014 May 1 [cited 2019 Feb 5];30(9):1312–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24451623

193. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol [Internet]. 2016 Jan [cited 2019 Feb 5];2(1):vew007. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27774300

194. WHO. Definition of regional groupings [Internet]. WHO. World Health Organization; 2017 [cited 2019 Apr 10]. Available from: https://www.who.int/healthinfo/global_burden_disease/definition_regions/en/

195. Yang J, Müller NF, Bouckaert R, Xu B, Drummond AJ. Bayesian phylodynamics of avian influenza A virus H9N2 in Asia with time-dependent predictors of migration. Cobey S,

editor. PLOS Comput Biol [Internet]. 2019 Aug 6 [cited 2019 Aug 9];15(8):e1007189. Available from: http://dx.plos.org/10.1371/journal.pcbi.1007189

196. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol [Internet]. 2007 Nov 30 [cited 2019 Feb 20];56(3):391–412. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17874105

197. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. Mol Biol Evol [Internet]. 2016 Aug 1 [cited 2019 Feb 21];33(8):2167–9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw082

198. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol [Internet]. 2018 Jan 1 [cited 2019 Feb 21];4(1). Available from: https://academic.oup.com/ve/article/doi/10.1093/ve/vey016/5035211

199. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. Integrating influenza antigenic dynamics with molecular evolution. Elife [Internet]. 2014 Feb 4 [cited 2019 Feb 24];3. Available from: https://elifesciences.org/articles/01914

200. Volz EM. Complex population dynamics and the coalescent under neutrality. Genetics [Internet]. 2012 Jan 1 [cited 2019 Aug 9];190(1):187–201. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22042576

201. WHO. WHO | WHO recommendations on the composition of influenza virus vaccines. 2019 [cited 2019 Aug 9]; Available from: https://www.who.int/influenza/vaccines/virus/recommendations/en/

202. Airport Planning F, Division APP- E. Airport Capacity Profiles, July 2014 [Internet]. 2001

[cited 2019 Aug 2]. Available from: www.faa.gov/airports/planning_capacity/

203. Tamerius JD, Shaman J, Alonso WJ, Bloom-Feshbach K, Uejio CK, Comrie A, et al. Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. PLoS Pathog. 2013;9(3).

204. Hsieh YH. Age groups and spread of influenza: Implications for vaccination strategy. BMC Infect Dis. 2010 Apr 30;10.

205. Petrova VN, Russell CA. The evolution of seasonal influenza viruses. Nat Rev Microbiol [Internet]. 2017 Oct 30 [cited 2019 Apr 18];16(1):47–60. Available from: http://www.nature.com/doifinder/10.1038/nrmicro.2017.118

206. Krammer F, Palese P. Advances in the development of influenza virus vaccines. Nat Rev Drug Discov [Internet]. 2015 Mar 27 [cited 2019 Aug 5];14(3):167–82. Available from: http://www.nature.com/articles/nrd4529

207. Mumford JA. Vaccines and viral antigenic diversity. Rev Sci Tech [Internet]. 2007 Apr [cited 2019 May 15];26(1):69–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17633294

208. Glezen WP, Gaglani MJ, Kozinetz CA, Piedra PA. Direct and indirect effectiveness of influenza vaccination delivered to children at school preceding an epidemic caused by 3 new influenza virus variants. J Infect Dis [Internet]. 2010 Dec 1 [cited 2019 Mar 12];202(11):1626–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21028955

209. Gaglani MJ, Piedra PA, Riggs M, Herschler G, Fewlass C, Glezen WP. Safety of the Intranasal, Trivalent, Live Attenuated Influenza Vaccine (LAIV) in Children With Intermittent Wheezing in an Open-Label Field Trial. Pediatr Infect Dis J [Internet]. 2008 May [cited 2019 Mar 12];27(5):444–52. Available from:

https://insights.ovid.com/crossref?an=00006454-200805000-00012

210. Piedra PA, Gaglani MJ, Kozinetz CA, Herschler GB, Fewlass C, Harvey D, et al. Trivalent live attenuated intranasal influenza vaccine administered during the 2003-2004 influenza type A (H3N2) outbreak provided immediate, direct, and indirect protection in children. Pediatrics [Internet]. 2007 Sep 1 [cited 2019 Mar 12];120(3):e553-64. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17698577

211. Piedra PA, Gaglani MJ, Riggs M, Herschler G, Fewlass C, Watts M, et al. Live Attenuated Influenza Vaccine, Trivalent, Is Safe in Healthy Children 18 Months to 4 Years, 5 to 9 Years, and 10 to 18 Years of Age in a Community-Based, Nonrandomized, Open-Label Trial. Pediatrics [Internet]. 2005 Sep 1 [cited 2019 Mar 12];116(3):e397–407. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16140685

212. Piedra PA, Gaglani MJ, Kozinetz CA, Herschler G, Riggs M, Griffith M, et al. Herd immunity in adults against influenza-related illnesses with use of the trivalent-live attenuated influenza vaccine (CAIV-T) in children. Vaccine [Internet]. 2005 Feb 18 [cited 2019 Mar 12];23(13):1540–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15694506

213. Piedra PA. Safety of the trivalent, cold-adapted influenza vaccine (CAIV-T) in children. Semin Pediatr Infect Dis [Internet]. 2002 Apr [cited 2019 Mar 12];13(2):90–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12122958

214. Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ, Kü Hnert D, et al. No Title. Bioinformatics [Internet]. 2014 Aug 15 [cited 2018 Dec 12];30(16). Available from: http://compevol.github.io/MultiTypeTree.

215. Shepard SS, Meno S, Bahl J, Wilson MM, Barnes J, Neuhaus E. Viral deep sequencing

needs an adaptive approach: IRMA, the iterative refinement meta-assembler. BMC Genomics [Internet]. 2016 Dec 5 [cited 2019 Apr 18];17(1):708. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3030-6

216. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. Prlic A, editor. PLoS Comput Biol [Internet]. 2014 Apr 10 [cited 2019 Feb 5];10(4):e1003537. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003537

217. Heled J, Bouckaert RR. Looking for trees in the forest: summary tree from posterior samples. BMC Evol Biol [Internet]. 2013 Oct 4 [cited 2019 Mar 8];13(1):221. Available from: http://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-13-221

218. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, et al. Integrating influenza antigenic dynamics with molecular evolution. Elife [Internet]. 2014 Feb 4 [cited 2019 Feb 4];3. Available from: https://elifesciences.org/articles/01914

219. Rodpothong P, Auewarakul P. Viral evolution and transmission effectiveness. World J Virol [Internet]. 2012 Oct 12 [cited 2019 May 29];1(5):131–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24175217

220. Taubenberger JK, Kash JC. Influenza virus evolution, host adaptation, and pandemic formation. Cell Host Microbe [Internet]. 2010 Jun 25 [cited 2019 Aug 6];7(6):440–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20542248

221. Neu KE, Henry Dunand CJ, Wilson PC. Heads, stalks and everything else: how can antibodies eradicate influenza as a human disease? Curr Opin Immunol [Internet]. 2016 [cited 2019 Aug 6];42:48–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27268395

222. Anisimova M, Kosiolà C. Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. [cited 2019 Aug 6]; Available from: https://pdfs.semanticscholar.org/d914/921fc95967c3028d3bf84524bb02a76fc30d.pdf

223. Mu NF, Rasmussen DA, Stadler T, Müller NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its Approximations. Mol Biol Evol [Internet]. 2017 Nov 1 [cited 2018 Dec 2];34(11):2970–81. Available from: https://academic.oup.com/mbe/article/34/11/2970/3896419

224. WHO. WHO | Vaccination greatly reduces disease, disability, death and inequity worldwide. WHO [Internet]. 2011 [cited 2019 Aug 7]; Available from: https://www.who.int/bulletin/volumes/86/2/07-040089/en/

225. Tsang TK, Fang VJ, Ip DKM, Perera RAPM, So HC, Leung GM, et al. Indirect protection from vaccinating children against influenza in households. Nat Commun [Internet]. 2019 Dec 10 [cited 2019 Aug 7];10(1):106. Available from: http://www.nature.com/articles/s41467-018-08036-6

226. CDC. Key Facts About Influenza (Flu) | CDC [Internet]. [cited 2019 Aug 7]. Available from: https://www.cdc.gov/flu/about/keyfacts.htm

227. CDC. Routine Vaccines | Disease Directory | Travelers' Health | CDC [Internet]. 2018 [cited 2019 Aug 7]. Available from: https://wwwnc.cdc.gov/travel/diseases/routine

228. CDC. Adult Immunization Schedule by Vaccine and Age Group | CDC [Internet]. 2019 [cited 2019 Aug 7]. Available from: https://www.cdc.gov/vaccines/schedules/hcp/imz/adult.html

229. Cowling BJ. Immunogenicity of Alternative Annual Influenza Vaccination Strategies in Older Adults in Hong Kong - Full Text View - ClinicalTrials.gov [Internet].

ClinicalTrial.gov. 2017 [cited 2019 Aug 7]. Available from: https://clinicaltrials.gov/ct2/show/NCT03330132

230. Lau MSY, Cowling BJ, Cook AR, Riley S. Inferring influenza dynamics and control in households. Proc Natl Acad Sci U S A [Internet]. 2015 Jul 21 [cited 2019 Feb 7];112(29):9094–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26150502

231. CDC. CDC Says "Take 3" Actions to Fight the Flu | CDC [Internet]. [cited 2019 Aug 7]. Available from: https://www.cdc.gov/flu/prevent/preventing.htm

232. Qiu X, Duvvuri VR, Bahl J. Computational Approaches and Challenges to Developing Universal Influenza Vaccines. Vaccines [Internet]. 2019 May 28 [cited 2019 Aug 6];7(2):45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31141933

233. Krammer F, Fouchier RAM, Eichelberger MC, Webby RJ, Shaw-Saliba K, Wan H, et al. NAction! How Can Neuraminidase-Based Immunity Contribute to Better Influenza Virus Vaccines? MBio [Internet]. 2018 May 3 [cited 2019 Feb 23];9(2):e02332-17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29615508

234. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Kelso J, editor. Bioinformatics [Internet]. 2018 Dec 1 [cited 2019 Aug 7];34(23):4121–3. Available from: https://academic.oup.com/bioinformatics/article/34/23/4121/5001388

235. Centers for Disease Control and Prevention. How Influenza (Flu) Vaccines Are Made | CDC [Internet]. 2018 [cited 2019 Feb 17]. Available from: https://www.cdc.gov/flu/protect/vaccine/how-fluvaccine-made.htm

236. Kreijtz JHCM, Fouchier RAM, Rimmelzwaan GF. Immune responses to influenza virus infection. Virus Res [Internet]. 2011 Dec 1 [cited 2019 Feb 23];162(1–2):19–30.

Available from: http://www.ncbi.nlm.nih.gov/pubmed/21963677

237. Wong S-S, Webby RJ. Traditional and new influenza vaccines. Clin Microbiol Rev [Internet]. 2013 Jul [cited 2019 Feb 17];26(3):476–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23824369

238. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance [Internet]. 2017 Mar 30 [cited 2019 Apr 28];22(13):30494. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28382917

239. NCBI. Influenza virus database - NCBI [Internet]. [cited 2019 Apr 28]. Available from: https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database

240. WHO. WHO | FluID - a global influenza epidemiological data sharing platform. WHO [Internet]. 2017 [cited 2019 Apr 28]; Available from: https://www.who.int/influenza/surveillance_monitoring/fluid/en/

241. Alkhamis MA, Arruda AG, Morrison RB, Perez AM. Novel approaches for Spatial and Molecular Surveillance of Porcine Reproductive and Respiratory Syndrome Virus (PRRSv) in the United States. Sci Rep [Internet]. 2017 Dec 28 [cited 2019 Aug 7];7(1):4343. Available from: http://www.nature.com/articles/s41598-017-04628-2

242. Kühnert D, Wu C-H, Drummond AJ. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. Infect Genet Evol [Internet]. 2011 Dec 1 [cited 2019 Aug 7];11(8):1825–41. Available from: https://www.sciencedirect.com/science/article/pii/S156713481100284X

243. Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, et al. The global circulation of seasonal influenza A (H3N2) viruses. Science [Internet]. 2008 Apr 18 [cited 2019 Feb 12];320(5874):340–6. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/18420927

244. Klingen TR, Reimering S, Guzmán CA, McHardy AC. In Silico Vaccine Strain Prediction for Human Influenza Viruses. Trends Microbiol [Internet]. 2018 Feb [cited 2019 Apr 27];26(2):119–31. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29032900

245. Morris DH, Gostic KM, Pompei S, Bedford T, Łuksza M, Neher RA, et al. Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. Trends Microbiol [Internet]. 2018 Feb [cited 2019 Apr 27];26(2):102–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29097090

246. Bui H-H, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics [Internet]. 2006 Mar 17 [cited 2019 May 12];7(1):153. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16545123

247. Sette A, Rappuoli R. Reverse Vaccinology: Developing Vaccines in the Era of Genomics. Immunity [Internet]. 2010 Oct 29 [cited 2019 Feb 23];33(4):530–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21029963

248. Wong TM, Allen JD, Bebin-Blackwell A-G, Carter DM, Alefantis T, DiNapoli J, et al. Computationally Optimized Broadly Reactive Hemagglutinin Elicits Hemagglutination Inhibition Antibodies against a Panel of H3N2 Influenza Virus Cocirculating Variants. J Virol [Internet]. 2017 Dec 15 [cited 2019 Feb 17];91(24):e01581-17. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28978710

249. Giles BM, Crevar CJ, Carter DM, Bissel SJ, Schultz-Cherry S, Wiley CA, et al. A Computationally Optimized Hemagglutinin Virus-Like Particle Vaccine Elicits Broadly Reactive Antibodies that Protect Nonhuman Primates from H5N1 Infection. J Infect Dis

[Internet]. 2012 May 15 [cited 2019 Feb 17];205(10):1562–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22448011

250. Giles BM, Ross TM. A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. Vaccine [Internet]. 2011 Apr 5 [cited 2019 Feb 17];29(16):3043–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21320540

251. Crevar CJ, Carter DM, Lee KYJ, Ross TM. Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades. Hum Vaccin Immunother [Internet]. 2015 Mar 4 [cited 2019 Feb 17];11(3):572–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25671661

252. Nabel GJ, Fauci AS. Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine. Nat Med [Internet]. 2010 Dec 1 [cited 2019 Feb 17];16(12):1389–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21135852

253. Ducatez MF, Bahl J, Griffin Y, Stigger-Rosser E, Franks J, Barman S, et al. Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development. Proc Natl Acad Sci U S A [Internet]. 2011 Jan 4 [cited 2019 Feb 17];108(1):349–54. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21173241

254. Jang YH, Seong BL. Options and obstacles for designing a universal influenza vaccine. Viruses [Internet]. 2014 Aug 18 [cited 2019 May 4];6(8):3159–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25196381

255. Zhang H, Wang L, Compans RW, Wang B-Z. Universal Influenza Vaccines, a Dream to Be Realized Soon. Viruses [Internet]. 2014 Apr 29 [cited 2019 May 4];6(5):1974. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24784572

256. Baum DA, Smith SD. Tree Thinking: An Introduction to Phylogenetic Biology. Roberts and Co., Greenwood Village, CO; 2012.

APPENDICES

I. SUPPMENTAL MATERIAL FOR CHAPTER 2

STRUCTUALLY INFORMED EVOLUTIONARY MODELS IMPROVE PHYLOGENETIC RECONSTRUCTION FOR EMERGING, SEASONAL, AND PANDEMIC INFLUENZA VIRUSES

Supplemental Table I-1. The nucleotide sites of functional partitions for each influenza subtype. The nucleotide sites are corresponding to the aligned HA open reading frame of nucleotide sequences for each influenza subtype. STC represents the combination of signal peptide, transmembrane domain and cytoplasmic tail. HPAI means highly pathogenic avian influenza.

Influenza subtypes	STC domain	Stalk domain	Head domain
A/H1N1pdm09	1-51, 1591-1701	52-216, 874-1590	217-873
A/H1N1postpdm	1-51, 1591-1701	52-216, 874-1590	217-873
A/H3N2	1-48, 1588-1701	49-204, 877-1587	205-876
B-Victoria	1-45, 1645-1758	46-171, 922-1644	172-921
B-Yamagata	1-45, 1642-1755	46-171, 919-1641	172-918
HPAI H7N9	1-54, 1573-1683	55-216, 856-1572	217-855
HPAI H5Nx	1-48, 1597-1713	49-174, 868-1596	175-867

Supplemental Table I-2. The domain-specific rates from p model compared to cp model for each influenza subtype. The results showed that p-model slightly underestimates the domain-specific rate compared to cp model. HPAI means highly pathogenic avian influenza.

Models: The **P model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **CP model** combines both c and p models, which estimates codon positions in protein structural partitions. Both models use HKY substitution model.

Datasets	Substitution rates (subs/site/year)			
	P model		CP model	
	Head	Stalk	Head	Stalk
A/H1N1pdm09	9.18E-03	6.46E-03	1.05E-02	7.81E-03
A/H1N1postpdm	3.37E-03	2.98E-03	4.12E-03	3.81E-03
A/H3N2	3.87E-03	3.02E-03	4.54E-03	3.83E-03
B-Victoria	2.50E-03	1.68E-03	3.07E-03	2.29E-03
B-Yamagata	1.80E-03	7.73E-04	3.47E-03	2.72E-03
HPAI H7N9	4.45E-03	3.52E-03	5.43E-03	4.48E-03
HPAI H5Nx-full	5.26E-03	3.80E-03	6.85E-03	5.26E-03

Supplemental Table I-3. Model selection and parameter estimations from two subsets of H3N2. The full H3N2 dataset is randomly sampled down to 40% to generate two subsets, which results in each containing 365 isolates from different geographical and temporal distributions. The results show that the new model performs stably for different subsets and are not sensitive the different distributions of the dataset by year and by country (Distributions of the two subsets are shown in Table I-3 in appendix figures a and b following the table).

Models: HKY model is a substitution model that considers different base frequencies and assigns different rates for transitions v.s. transversions. The **c model** represents SRD06 codon position model. The partitioning strategy is to analyze codon positions 1 + 2 and codon position 3 separately. The **p model** takes the protein structure partitions into account based on the amino acid positions for each domain on the linear diagram. The **cp model** combines both c and p models, which estimates codon positions in protein structural partitions. Both models use HKY substitution model.

Parameters	H3N2 subset 1	H3N2 subset 2
Bayes Factors by path-sampling for model selection		
HKY model	-	-
c model	146.3	130.9
p model	5.2	6.0
cp model	167.3	144.5
Bayes Factors by stepping-stone for model selection		
HKY model	-	-
c model	146.2	130.8
p model	5.2	5.7
cp model	167.8	144.0
Substitution rates for each model		
HKY model	4.39E-03 [3.9596E-3,4.8401E-3]	4.33E-03 [3.9024E-3,4.7817E-3]
c model	4.51E-03 [4.0855E-3,4.9475E-3]	4.43E-03 [3.9946E-3,4.8712E-3]
p model	4.37E-03 [3.9456E-3,4.8208E-3]	4.31E-03 [3.8896E-3,4.7522E-3]
cp model	4.50E-03 [4.0775E-3,4.9391E-3]	4.36E-03 [3.9300E-3,4.8035E-3]
Root Height for each model		
HKY model	13.851 [13.2406,14.7147]	13.866 [13.2242,14.7571]
c model	13.855 [13.2584,14.7190]	13.881 [13.1729,14.7161]
p model	13.858 [13.2862,14.7688]	13.878 [13.2330,14.7818]
cp model	13.859 [13.2574,14.7566]	13.874 [13.1963,14.7553]
Partition-Specific substitution rates		
Stalk domain	5.08E-03	4.93E-03
Head domain	6.01E-03	5.71E-03

Table I-3 appendix figure a. Distribution of H3N2 subset 1 by year and by country.

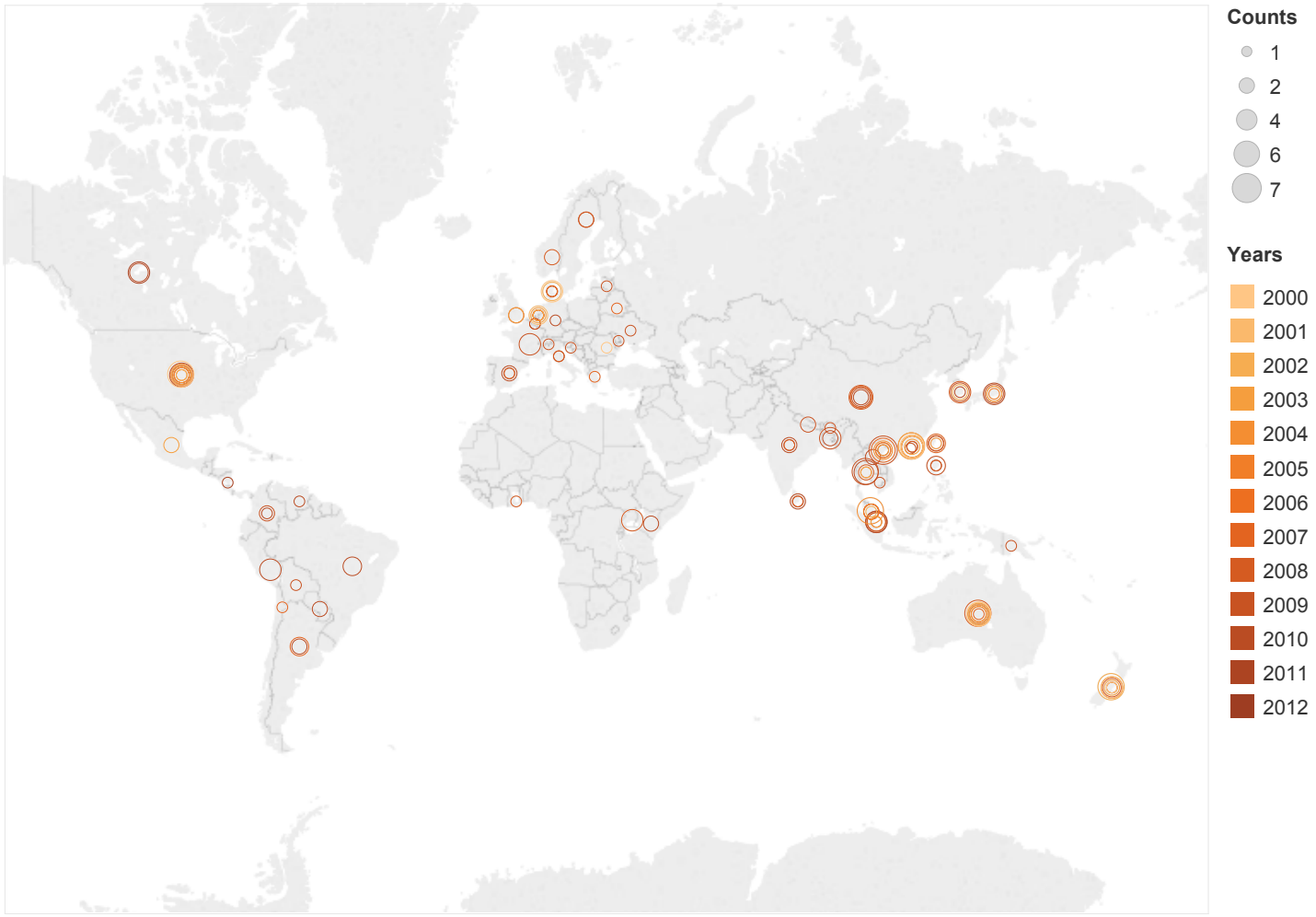
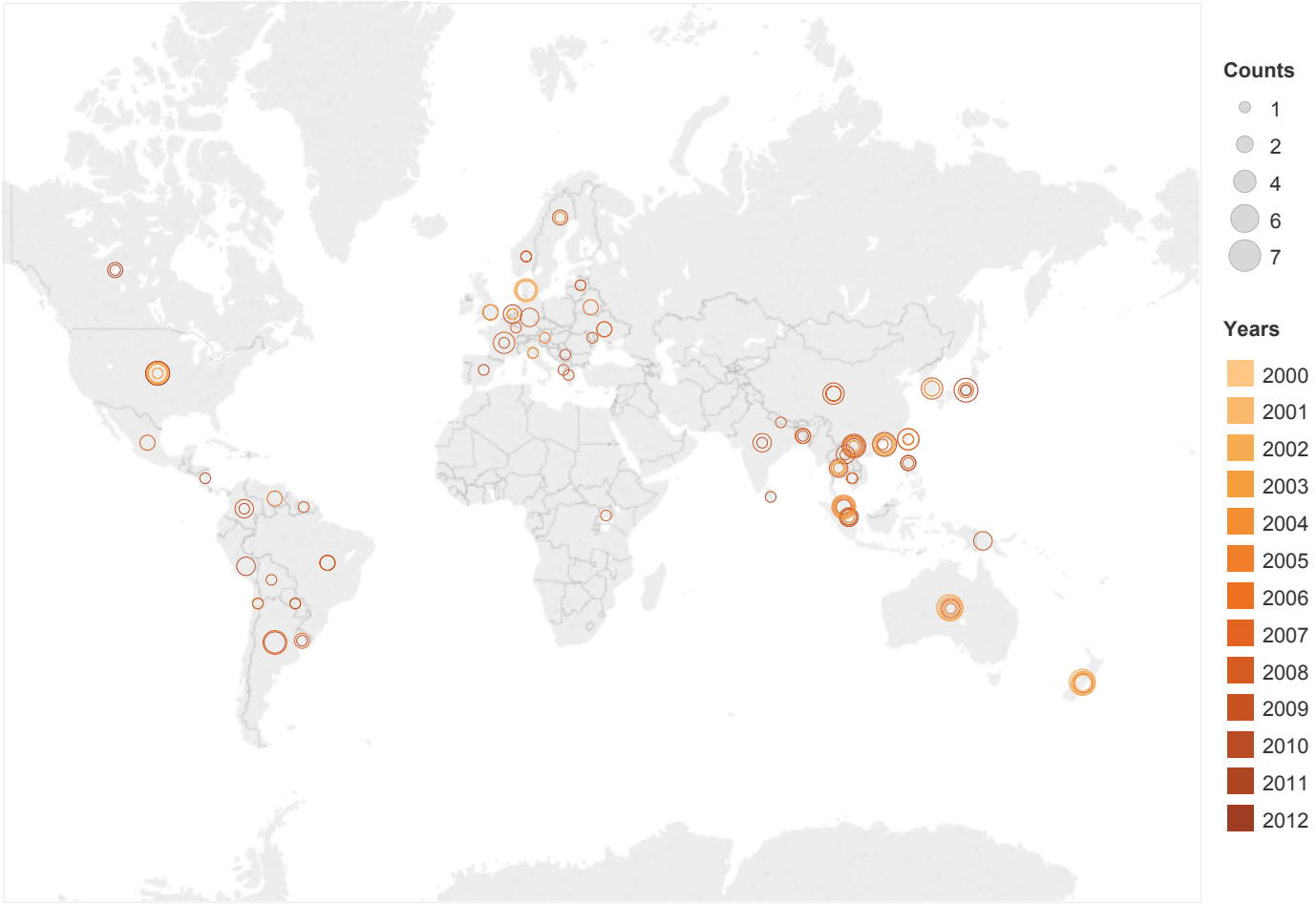


Table I-3 appendix figure b. Distribution of H3N2 subset 2 by year and by country.



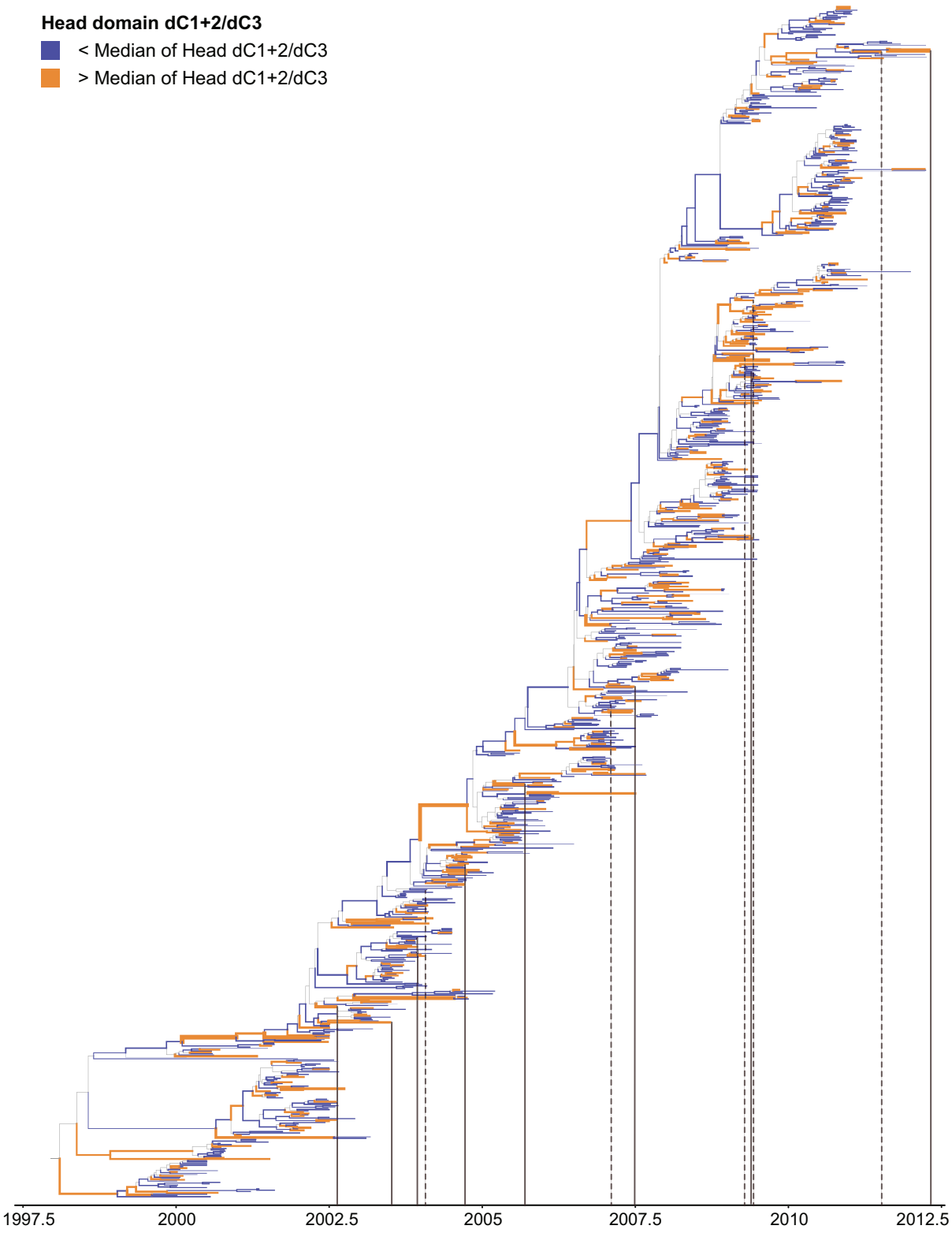
Supplemental Table I-4. Model selection and parameter estimation from the Early and Later epidemic stage of H1N1pdm09. The early epidemic stage of H1N1pdm09 is defined as 04/01/2019-07/15/2019 with 274 isolates, and the Later epidemic stage of H1N1pdm09 is defined as 07/16/2019-12/31/2019 with 231 isolates. The results show that the superiority of the new model holds (BF>5 compared to HKY model) and performs stably (overall substitution rates, root height and domain-specific substitution rates are similar from each model) for dataset from different stage of epidemics. The substitution rates and overall $d\mathrm{C}_{1+2}/d\mathrm{C}_3$ are lower at the later stage of the pandemic than the early stage of the pandemic.

Parameters	H1N1pdm09-Early	H1N1pdm09-Later
Bayes Factors by path-sampling for model selection		
HKY model	-	-
c model	34.0	28.2
p model	5.3	7.5
cp model	40.9	35.7
Bayes Factors by stepping-stone for model selection		
HKY model	-	-
c model	33.6	27.5
p model	5.7	7.0
cp model	42.5	34.9
Substitution rates for each model		
HKY model	9.66E-03 [7.54E-03, 1.20E-02]	8.03E-03 [6.57E-03, 9.54E-03]
c model	9.84E-03 [7.71E-03, 1.21E-02]	8.06E-03 [6.65E-03, 9.54E-03]
p model	9.66E-03 [7.49E-03, 1.19E-02]	8.02E-03 [6.50E-03, 9.52E-03]
cp model	9.75E-03 [7.51E-03, 1.20E-02]	8.09E-03 [6.66E-03, 9.59E-03]
Root Height for each model		
HKY model	0.551 [0.4311, 0.7001]	0.8801 [0.7322, 1.0806]
c model	0.546 [0.4313, 0.7009]	0.8821 [0.7353, 1.0773]
p model	0.547 [0.4305, 0.6950]	0.8819 [0.7379, 1.0863]
cp model	0.551 [0.4334, 0.7066]	0.8808 [0.7313, 1.0704]
Partition-Specific substitution rates		

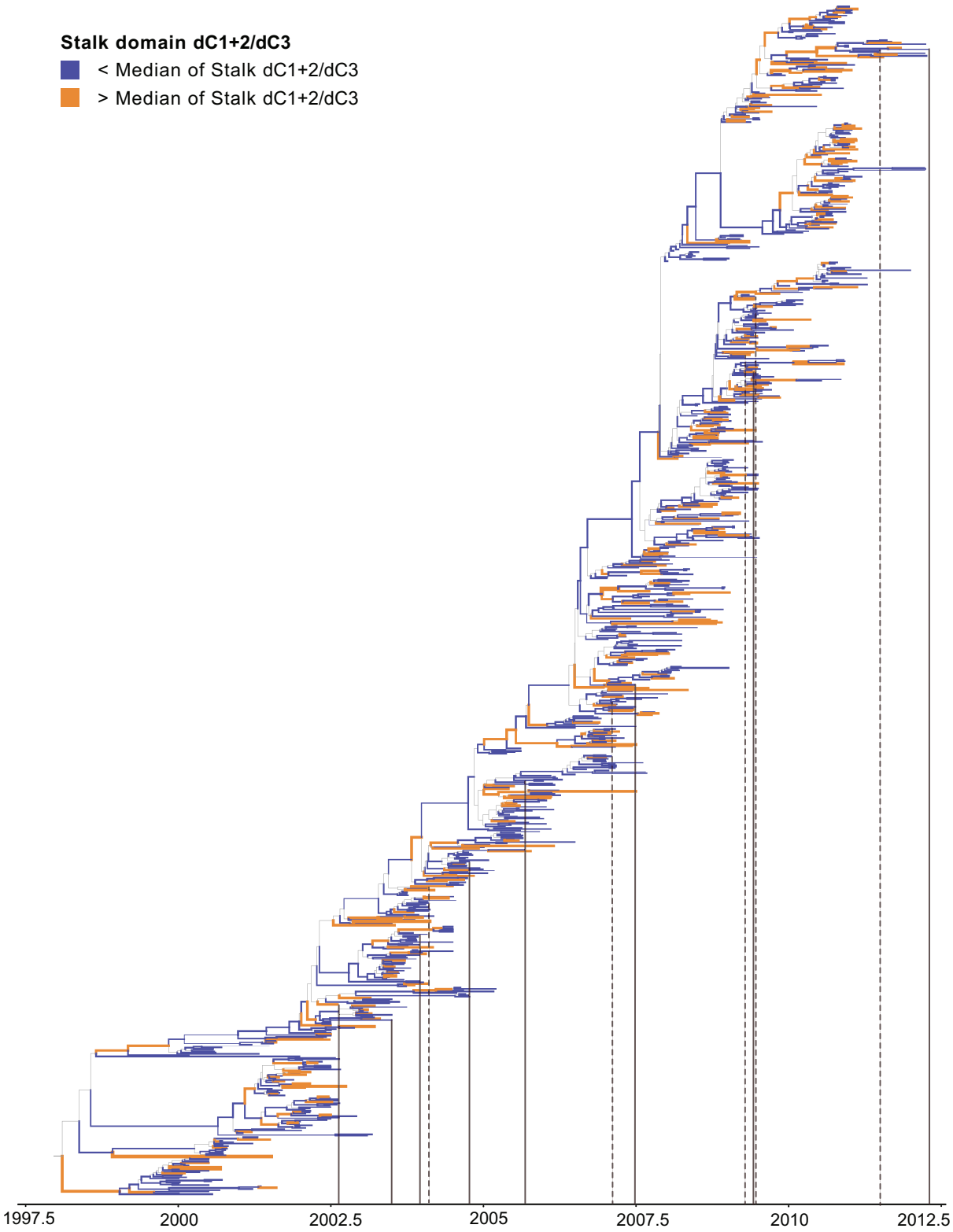
Stalk domain	9.36E-03	8.59E-03
Head domain	1.32E-02	1.10E-02
Overall dC_{1+2}/dC_3	0.45	0.39

Supplemental Figure I-1. H3N2 Branch Specific dC$_{1+2}$/dC$_3$ for Head and Stalk domain separately. dC$_{1+2}$/dC$_3$ means the ratio of substitution rate of codon positions 1 and 2 over the rate of codon position 3. Figure a is for Head domain and b is for Stalk domain. These branch specific dC$_{1+2}$/dC$_3$ is the median value for each branch from the MCMC steps. The branch stroke weight is proportional to the value of dC$_{1+2}$/dC$_3$ for each tree. The blue branch color represents the branch specific dC$_{1+2}$/dC$_3$ is lower than the overall median of dC$_{1+2}$/dC$_3$ for all isolates, while the orange branch color means the branch specific value is higher than the overall median. Vertical lines indicate the introduction time of vaccine strains selected by WHO. Solid vertical line represents North Hemisphere vaccine strain and dashed line represents South Hemisphere vaccine strain. There were 13 vaccine strains introduced during a 10-year period.
Head domain has much higher dC$_{1+2}$/dC$_3$ (0.63-0.75) compared to stalk domain (0.21-0.23) but all dC$_{1+2}$/dC$_3$ are less than 1, which means the stalk domain is under higher purifying selection to maintain its conserved functionality. Some branches have higher dC$_{1+2}$/dC$_3$, where the color is orange and branch weight is thicker. Taken together, this model could provide potential biological and quantitative information to understand viral evolution.

a. Head domain



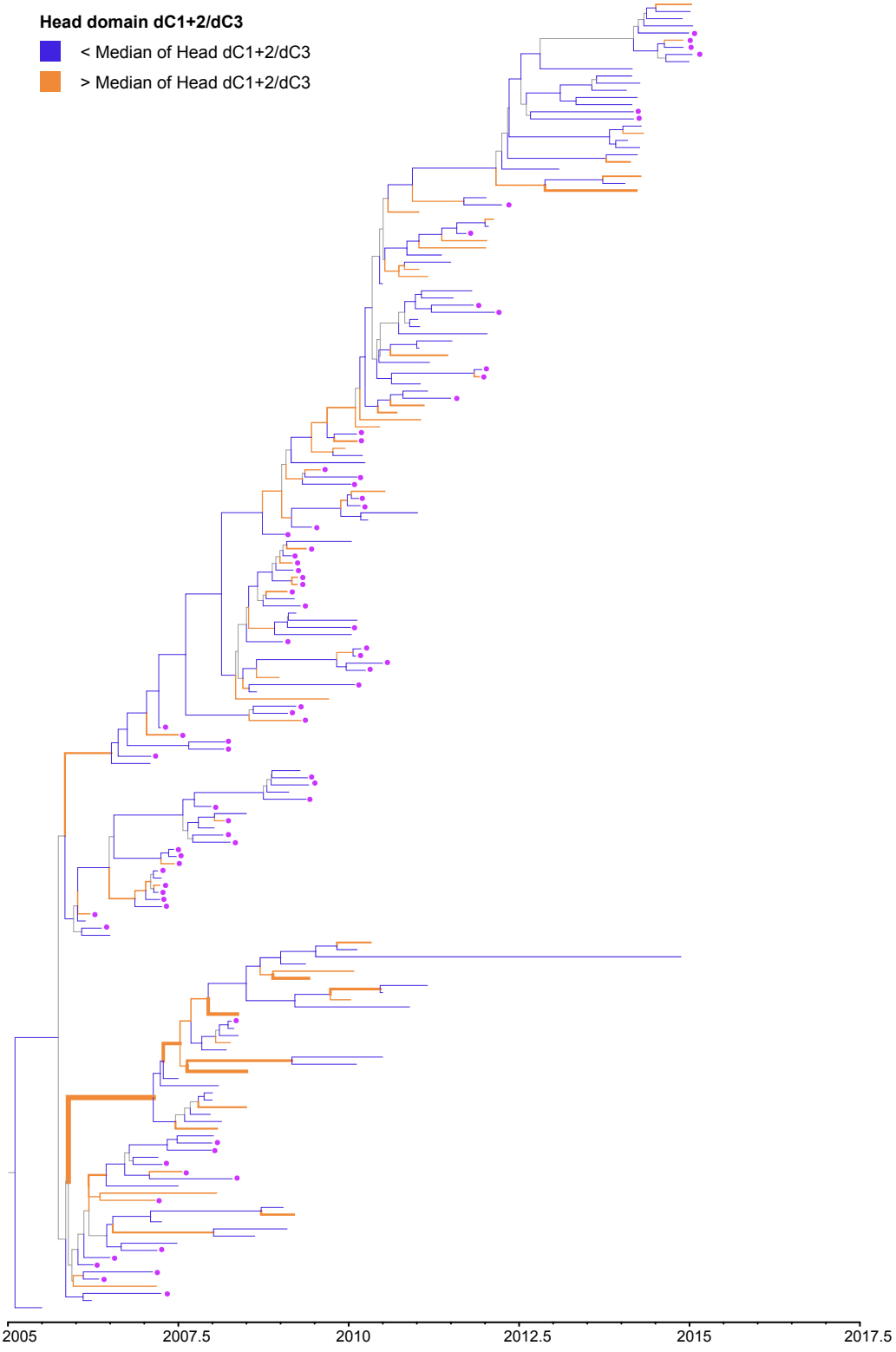
b. Stalk domain



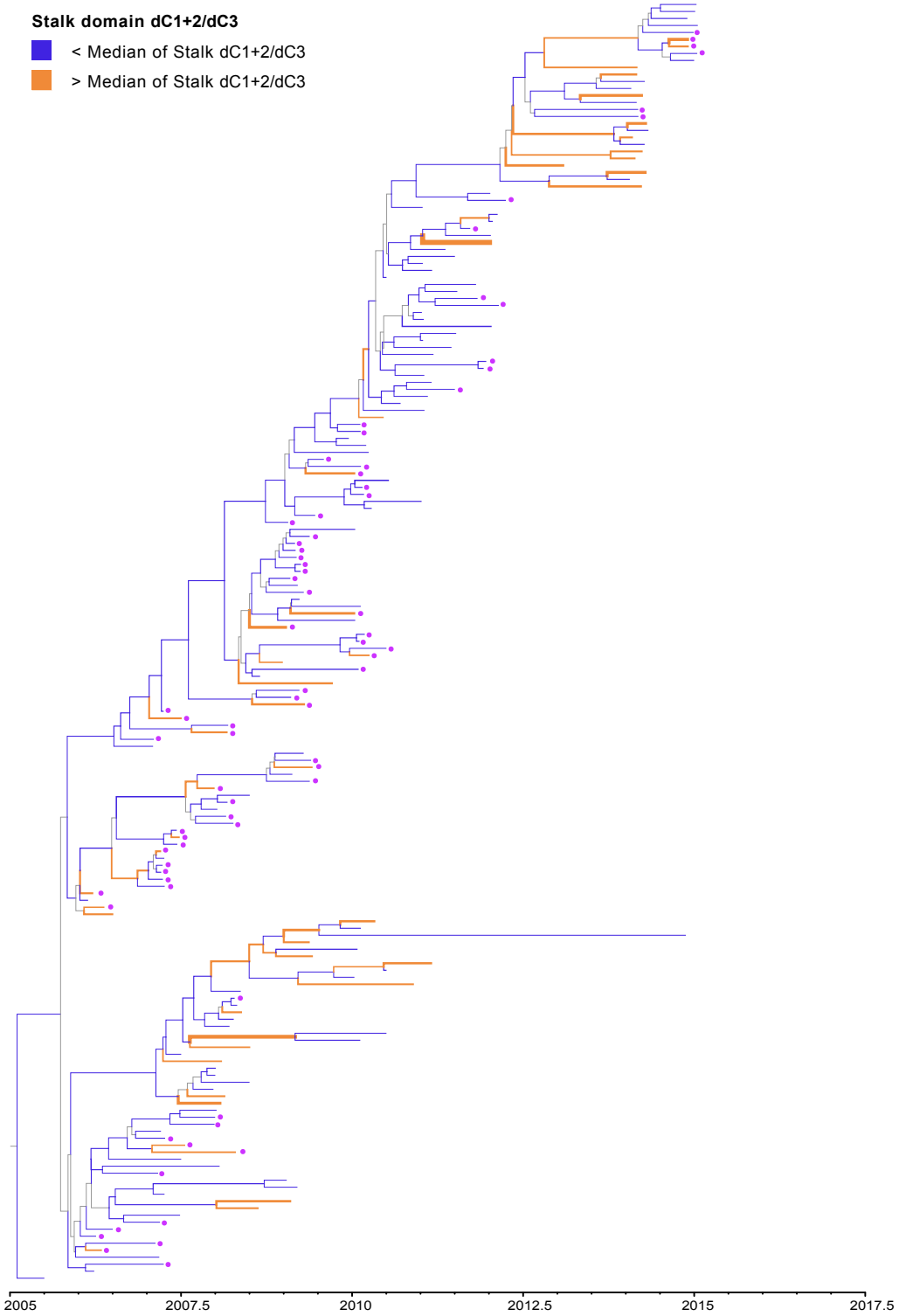
Supplemental Figure I-2. Egypt H5Nx Branch Specific dC_{1+2}/dC_3 for Head and Stalk domain separately. dC_{1+2}/dC_3 means the ratio of substitution rate of codon positions 1 and 2 over the rate of codon position 3. Figure a is for Head domain and b is for Stalk domain. These branch specific dC_{1+2}/dC_3 is the median value for each branch from the MCMC steps. The branch stroke weight is proportional to the value of dC_{1+2}/dC_3 for each tree. The blue branch color represents the branch specific dC_{1+2}/dC_3 is lower than the overall median of dC_{1+2}/dC_3 for all isolates, while the orange branch color means the branch specific value is higher than the overall median.

Head domain has much higher dC_{1+2}/dC_3 (0.12 – 1.87) compared to stalk domain (0.09 – 0.32), which means the stalk domain is under higher purifying selection to maintain its conservation and head domain occasionally experiences diversifying selection with $dC_{1+2}/dC_3 >1$ in avian population. Furthermore, higher dC_{1+2}/dC_3 are observed in avian viral branches, which means that viruses experience lesser purifying selection pressure to generate more diverse on non-synonymous changes in avian populations. Taken together, this model could provide potential biological explanations of host factors on viral evolution.

a. Head domain



b. Stalk domain



II. SUPPMENTAL MATERIAL FOR CHAPTER 4

DIFFUSION DYNAMICS OF SEASONAL INFLUENZA VIRUSES IN THE U.S. DRIVEN BY FLIGHT CONNECTIONS AND HIGH-RISK POPULATIONS

Supplemental Text II-1. Subsampling strategy

After coded into the WHO World Bank regions (https://www.who.int/healthinfo/global_burden_disease/definition_regions/en/; U.S. was kept separately from North America region for purpose of analysis) or U.S. Health and Human Services regions (HHS, https://www.hhs.gov/about/agencies/iea/regional-offices/index.html; Hawaii was kept separately from Region 9 and Alaska was kept separately from Region 10 because of distinctive geographic location and climate conditions), identical sequences collected from the same region were removed. After conducting descriptive statistics of metadata distributions by region and by year, random sampling was used to generate a representative subset in each region.

For the global regions, if the data records of a region in a given year had less than 20 isolates, all of these isolates were kept; if the records had more than 20 but less than 200 isolates, 20 isolates were randomly selected; if the records had more than 200 but less than 1000 isolates, 30 isolates were randomly selected; if the records contained more than 1000 isolates, 50 isolates were randomly retained. For the U.S. data in the global dataset, since the study aimed to capture the picture of introductions into the U.S., so slightly higher weights were put on the U.S. samples. 50 isolates were randomly selected if the records for the year were less than 1000 and 60 isolates were randomly selected if the records had more than 1000 isolates.

When it came to the U.S. HHS region data, after examining the data distributions by each region, all isolates were kept for the region that had less than 80 isolates, 80 isolates were randomly selected if the records for the region had more than 80 but less than 200 isolates, 90 isolates for the region that had more than 200 but less than 300 isolates, 100 isolates for the region that had more than 300 but less than 500 isolates, 110 isolates for the region that had more than 500 but less than 1000 isolates, and 120 isolates for the region that had more than 1000 isolates.

The brief summary of subsampling strategy was below:

Global dataset subsampling criteria (by region and year)			U.S. region dataset subsampling criteria (by region)	
	Total records	Subsampled	Total records	Subsampled
For seven WHO defined global regions	<20	Keep all	N < 80	Keep all
	20 ≤ N < 200	20	80 ≤ N < 200	80
	200 ≤ N < 1000	30	200 ≤ N < 300	90
	N ≥ 1000	50	300 ≤ N < 500	100
For the U.S. itself	N < 1000	50	500 ≤ N < 1000	110
	N ≥ 1000	60	N ≥ 1000	120

Multiple sets of subsampled data were generated for global dataset and for the U.S. region dataset. Each subset was conducted with a preliminary analysis by checking the estimated root height, evolutionary rates and potential outliers to confirm that the subsampling strategy was proper and consistent results from different subsets were generated.

Supplemental Table II-1. Distribution of the global dataset by type, year and region before and after the subsampling procedure.

Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

Region	Year	A/H1N1		A/H3N2		B-Victoria		B-Yamagata	
		Record	Subset	Record	Subset	Record	Subset	Record	Subset
AF	2011	120	20	79	20	47	20	12	12
AF	2012	45	20	100	20	50	20	26	20
AF	2013	92	20	78	20	26	20	42	20
AF	2014	35	20	153	20	33	20	28	20
AF	2015	166	20	190	20	4	4	90	20
AF	2016	137	20	161	20	117	20	54	20
AF	2017	250	30	219	30	89	20	27	20
AF	2018	206	30	48	20	41	20	98	20
EAP	2011	866	30	695	30	292	30	103	20
EAP	2012	170	20	600	30	230	30	180	20
EAP	2013	566	30	580	30	133	20	239	30
EAP	2014	536	30	882	30	89	20	264	30
EAP	2015	440	30	1,299	50	157	20	434	30
EAP	2016	947	30	1,261	50	538	30	298	30
EAP	2017	723	30	1,832	50	342	30	635	30
EAP	2018	742	30	661	30	155	20	487	30
ECA	2011	854	30	236	30	136	20	45	20
ECA	2012	180	20	833	30	114	20	200	20
ECA	2013	907	30	441	30	111	20	441	30
ECA	2014	690	30	979	30	28	20	201	30
ECA	2015	869	30	1,335	50	78	20	527	30
ECA	2016	2,009	50	1,806	50	784	40	118	20
ECA	2017	494	30	2,271	50	327	30	768	30
ECA	2018	1,023	50	1,369	50	97	20	1,328	50
LAC	2011	132	20	228	30	20	20	1	1
LAC	2012	261	30	158	20	55	20	22	20
LAC	2013	242	30	113	20	41	20	23	20
LAC	2014	126	20	301	30	42	20	92	20
LAC	2015	219	30	368	30	63	20	141	20
LAC	2016	713	30	161	20	172	20	98	20
LAC	2017	118	20	716	30	159	20	294	30

LAC	2018	693	30	341	30	129	20	247	30
MENA	2011	86	20	79	20	18	18	2	2
MENA	2012	46	20	53	20	12	12	8	8
MENA	2013	50	20	38	20	6	6	24	20
MENA	2014	66	20	81	20	0	0	43	20
MENA	2015	215	30	80	20	10	10	38	20
MENA	2016	193	20	216	30	74	20	17	17
MENA	2017	177	20	153	20	55	20	78	20
MENA	2018	99	20	213	30	17	17	92	20
NA	2011	5	5	93	20	6	6	2	2
NA	2012	33	20	121	20	1	1	1	1
NA	2013	93	20	92	20	4	4	9	9
NA	2014	29	20	246	30	2	2	13	13
NA	2015	13	13	205	30	7	7	10	10
NA	2016	262	30	531	30	21	20	14	14
NA	2017	45	20	1,502	30	79	20	87	20
NA	2018	65	20	794	30	38	20	74	20
SAS	2011	49	20	37	20	22	20	3	3
SAS	2012	175	20	36	20	21	20	19	19
SAS	2013	124	20	118	20	9	9	12	12
SAS	2014	58	20	47	20	10	10	18	18
SAS	2015	316	30	63	20	4	4	30	20
SAS	2016	61	20	49	20	64	20	57	20
SAS	2017	178	20	121	20	41	20	33	20
SAS	2018	135	20	64	20	14	14	54	20
USA	11-12	214	50	371	50	47	47	51	50
USA	12-13	133	50	866	50	62	50	108	50
USA	13-14	532	50	294	50	51	50	72	50
USA	14-15	49	30	1,256	60	128	50	297	50
USA	15-16	1,107	60	572	50	316	50	399	50
USA	16-17	253	50	1,895	60	417	50	352	50
USA	17-18	654	50	1,142	60	163	50	584	50
Total		**21,086**	**1,718**	**31,922**	**1,930**	**6,418**	**1,341**	**10,164**	**1,461**

Supplemental Table II-2. Distribution of U.S. dataset by type and region before and after the subsampling procedure.

Region	A/H1N1		A/H3N2		B-Victoria		B-Yamagata	
	Record	Subset	Record	Subset	Record	Subset	Record	Subset
Region 1	275	90	733	110	134	80	210	90
Region 2	339	100	599	110	76	76	113	80
Region 3	420	100	832	120	164	80	254	90
Region 4	638	110	1,121	120	251	90	383	100
Region 5	505	110	1,289	140	200	90	326	100
Region 6	423	100	940	120	159	80	262	90
Region 7	195	80	417	100	65	65	125	80
Region 8	378	100	666	110	135	80	294	90
Region 9	449	100	862	120	157	80	242	90
Region 10	226	90	634	110	88	80	202	90
Hawaii	104	80	228	90	70	70	104	80
Alaska	68	68	215	90	20	20	65	65
Total	**4,020**	**1,128**	**8,536**	**1,340**	**1,519**	**891**	**2,580**	**1,045**

Supplemental Table II-3a. The estimated median and 95% BCI of the most recent common ancestor time from phylogenetic modeling for the global datasets.

Dataset	The oldest isolate date	Median of tMRCA*	95% BCI+
A/H1N1	2011.00	2010.14	(2010.01, 2010.26)
A/H3N2	2011.00	2010.24	(2010.10, 2010.36)
B-Victoria	2011.00	2009.69	(2009.50, 2009.87)
B-Yamagata	2011.01	2010.26	(2010.09, 2010.42)

tMRCA*: The most common ancestor time.
95% BCI+: 95% Bayesian Credible Interval.

Supplemental Table II-3b. The estimated median and 95% BCI of the most recent common ancestor time from phylogenetic modeling for U.S. datasets.

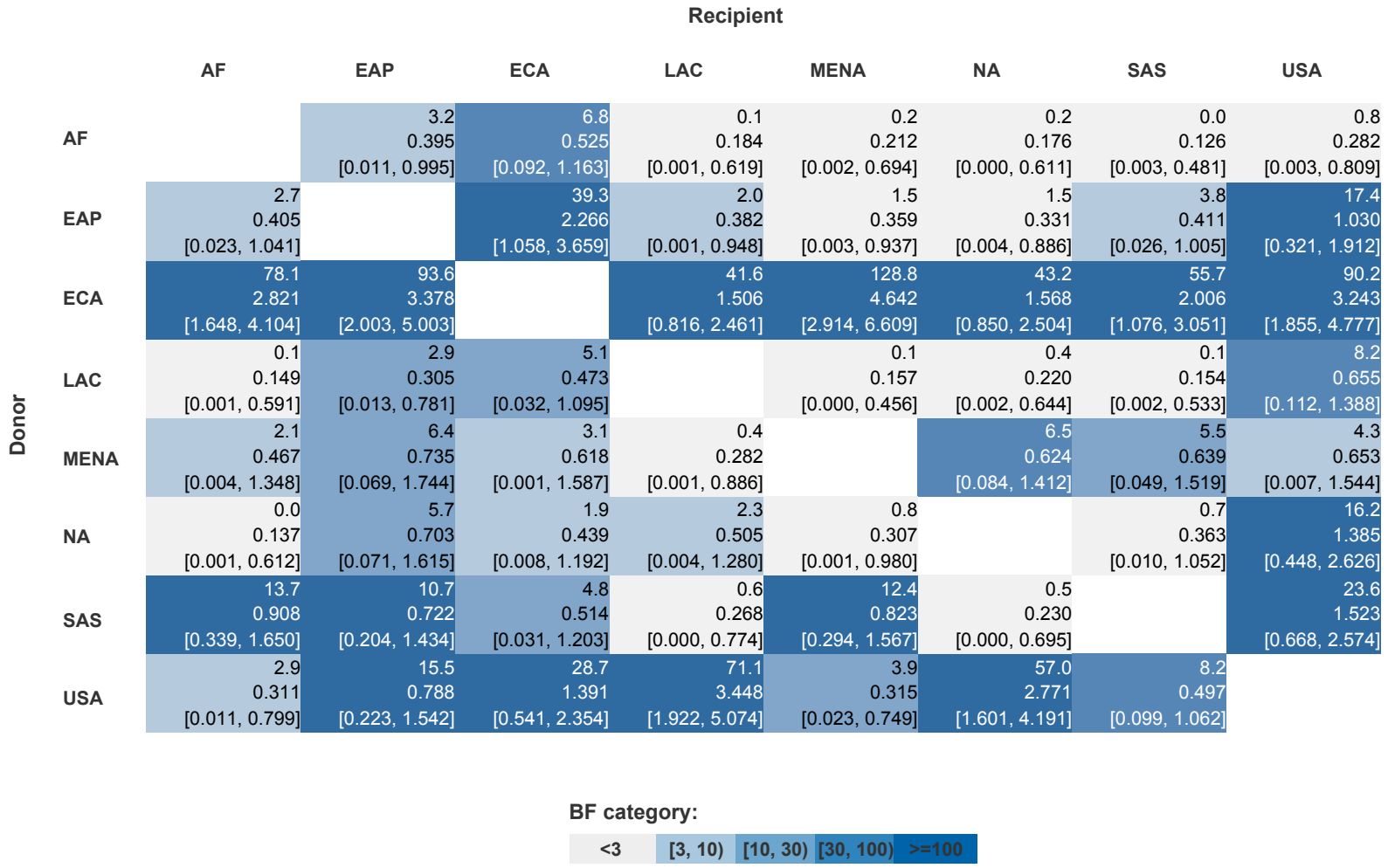
Dataset	The oldest isolate date	Median of tMRCA*	95% BCI+
A/H1N1	2011.95	2011.34	(2011.18, 2011.48)
A/H3N2	2011.77	2011.14	(2010.98, 2011.30)
B-Victoria	2011.76	2010.93	(2010.70, 2011.17)
B-Yamagata	2011.78	2011.04	(2010.81, 2011.24)

tMRCA*: The most common ancestor time.
95% BCI+: 95% Bayesian Credible Interval.

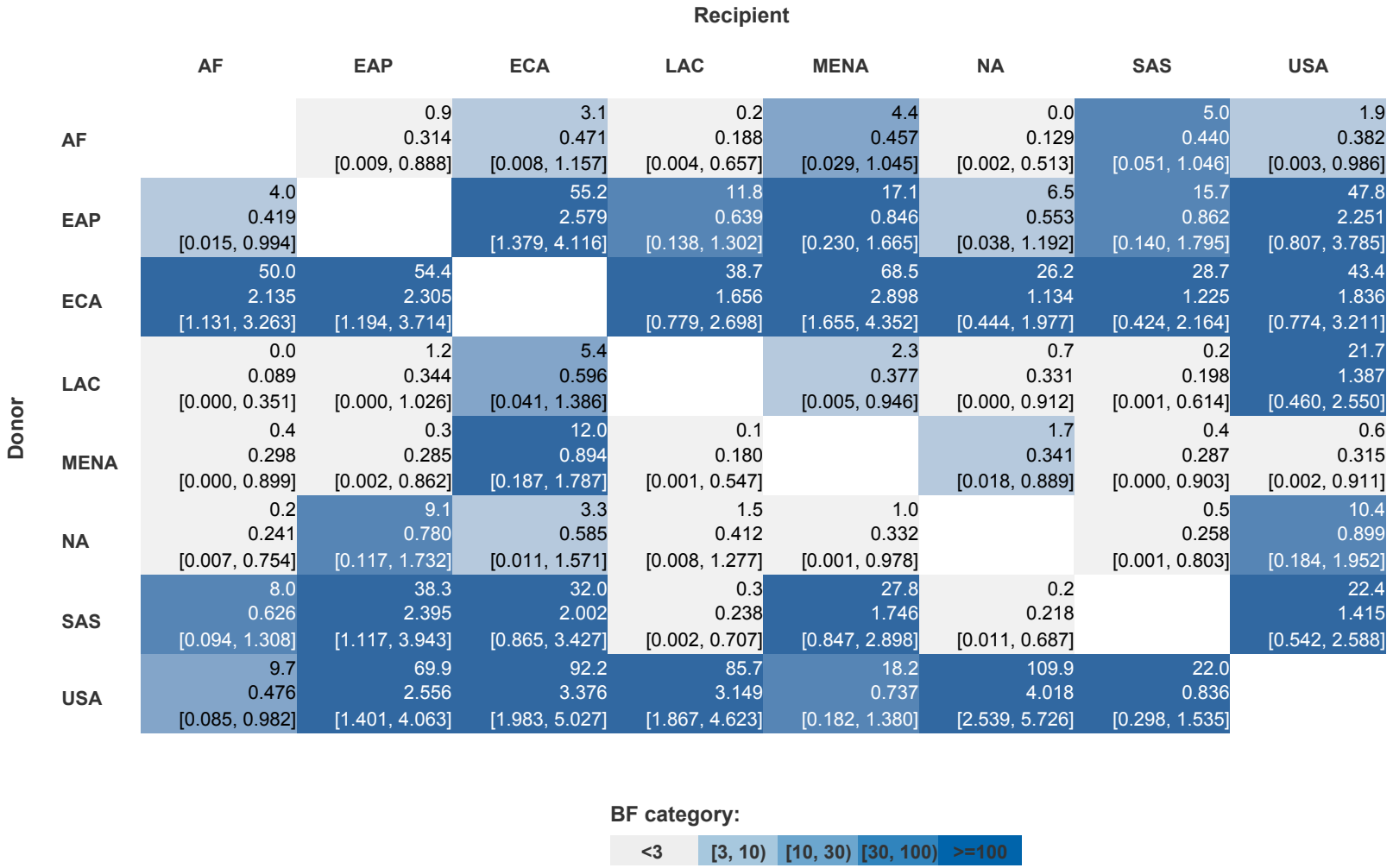
Supplemental Table II-4. The full global transmission matrix for each influenza type. The three items in each matrix cell in order are count of introductions, migration rate, and its 95% Bayesian Credible Interval (95% BCI). The rows of the table are the donor region and the columns are the recipient region. The cell color indicates BF levels, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

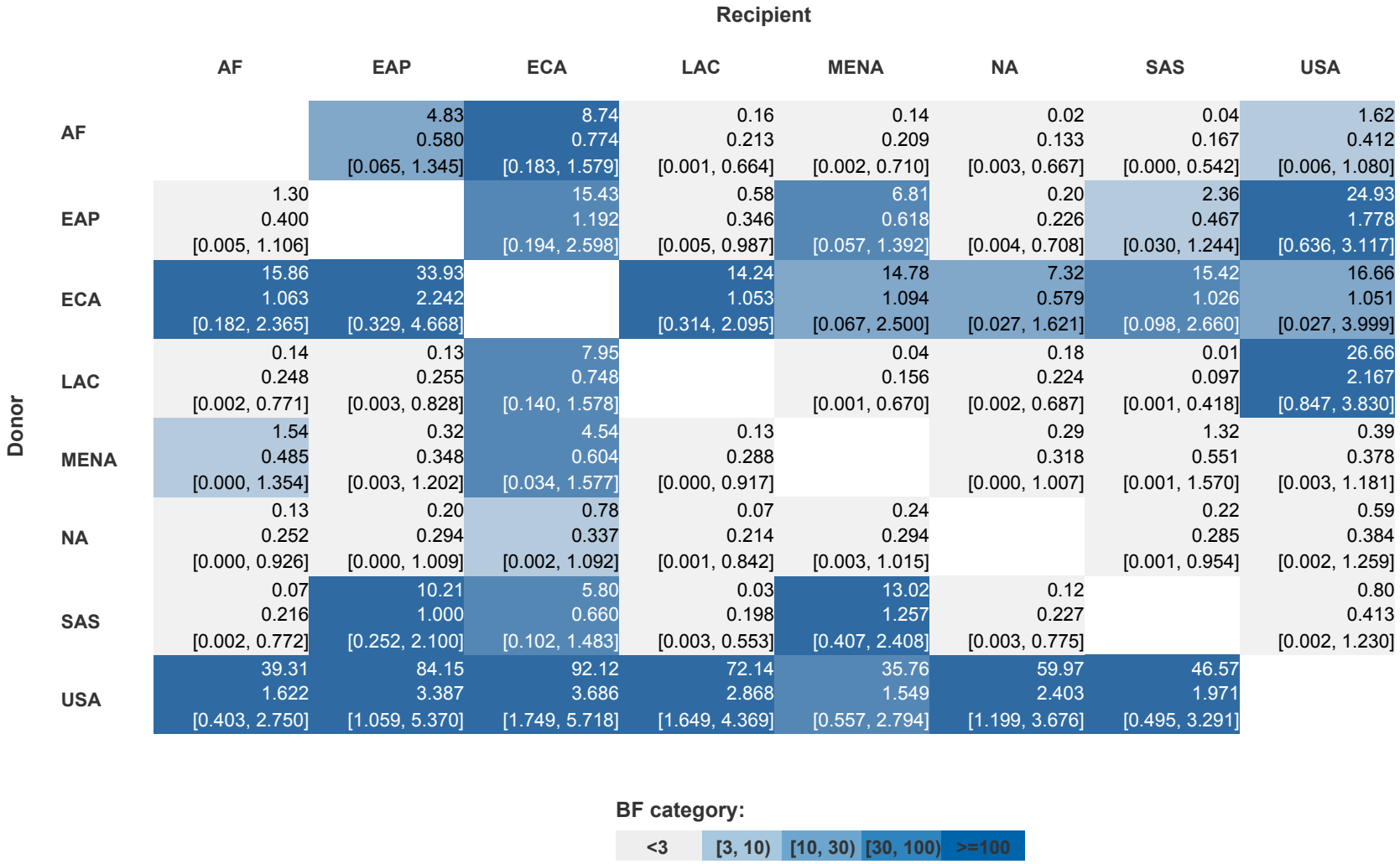
II-4a. A/H1N1


Recipient
AF
EAP
ECA
LAC
MENA
NA
SAS
USA
Donor
AF
3.2
0.395
[0.011, 0.995]
6.8
0.525
[0.092, 1.163]
0.1
0.184
[0.001, 0.619]
0.2
0.212
[0.002, 0.694]
0.2
0.176
[0.000, 0.611]
0.0
0.126
[0.003, 0.481]
0.8
0.282
[0.003, 0.809]
EAP
2.7
0.405
[0.023, 1.041]
39.3
2.266
[1.058, 3.659]
2.0
0.382
[0.001, 0.948]
1.5
0.359
[0.003, 0.937]
1.5
0.331
[0.004, 0.886]
3.8
0.411
[0.026, 1.005]
17.4
1.030
[0.321, 1.912]
ECA
78.1
2.821
[1.648, 4.104]
93.6
3.378
[2.003, 5.003]
41.6
1.506
[0.816, 2.461]
128.8
4.642
[2.914, 6.609]
43.2
1.568
[0.850, 2.504]
55.7
2.006
[1.076, 3.051]
90.2
3.243
[1.855, 4.777]
LAC
0.1
0.149
[0.001, 0.591]
2.9
0.305
[0.013, 0.781]
5.1
0.473
[0.032, 1.095]
0.1
0.157
[0.000, 0.456]
0.4
0.220
[0.002, 0.644]
0.1
0.154
[0.002, 0.533]
8.2
0.655
[0.112, 1.388]
MENA
2.1
0.467
[0.004, 1.348]
6.4
0.735
[0.069, 1.744]
3.1
0.618
[0.001, 1.587]
0.4
0.282
[0.001, 0.886]
6.5
0.624
[0.084, 1.412]
5.5
0.639
[0.049, 1.519]
4.3
0.653
[0.007, 1.544]
NA
0.0
0.137
[0.001, 0.612]
5.7
0.703
[0.071, 1.615]
1.9
0.439
[0.008, 1.192]
2.3
0.505
[0.004, 1.280]
0.8
0.307
[0.001, 0.980]
0.7
0.363
[0.010, 1.052]
16.2
1.385
[0.448, 2.626]
SAS
13.7
0.908
[0.339, 1.650]
10.7
0.722
[0.204, 1.434]
4.8
0.514
[0.031, 1.203]
0.6
0.268
[0.000, 0.774]
12.4
0.823
[0.294, 1.567]
0.5
0.230
[0.000, 0.695]
23.6
1.523
[0.668, 2.574]
USA
2.9
0.311
[0.011, 0.799]
15.5
0.788
[0.223, 1.542]
28.7
1.391
[0.541, 2.354]
71.1
3.448
[1.922, 5.074]
3.9
0.315
[0.023, 0.749]
57.0
2.771
[1.601, 4.191]
8.2
0.497
[0.099, 1.062]
BF category:
<3
[3, 10)
[10, 30)
[30, 100)
>=100

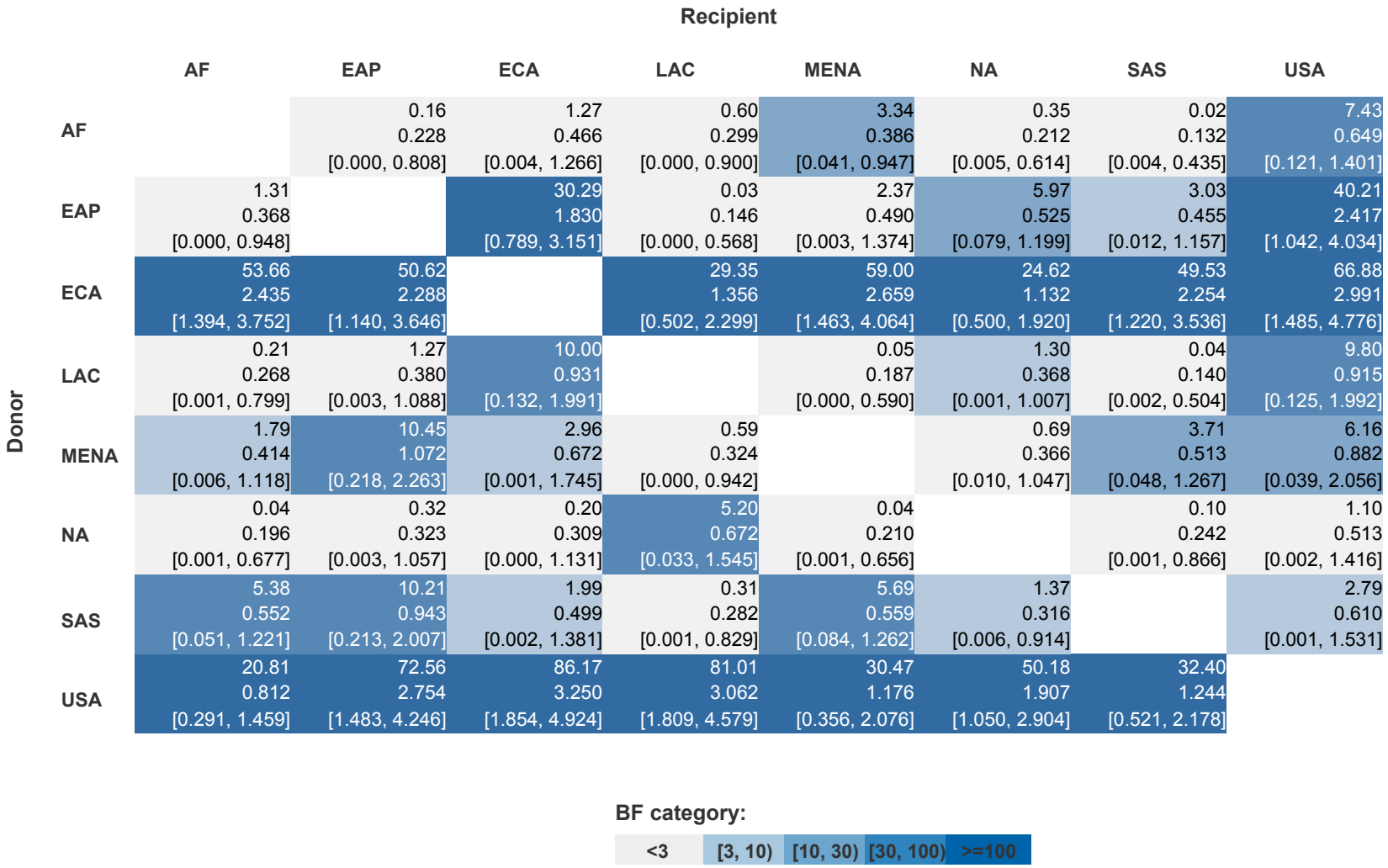
II-4b. A/H3N2


Recipient
AF
EAP
ECA
LAC
MENA
NA
SAS
USA
Donor
AF
0.9
0.314
[0.009, 0.888]
3.1
0.471
[0.008, 1.157]
0.2
0.188
[0.004, 0.657]
4.4
0.457
[0.029, 1.045]
0.0
0.129
[0.002, 0.513]
5.0
0.440
[0.051, 1.046]
1.9
0.382
[0.003, 0.986]
EAP
4.0
0.419
[0.015, 0.994]
55.2
2.579
[1.379, 4.116]
11.8
0.639
[0.138, 1.302]
17.1
0.846
[0.230, 1.665]
6.5
0.553
[0.038, 1.192]
15.7
0.862
[0.140, 1.795]
47.8
2.251
[0.807, 3.785]
ECA
50.0
2.135
[1.131, 3.263]
54.4
2.305
[1.194, 3.714]
38.7
1.656
[0.779, 2.698]
68.5
2.898
[1.655, 4.352]
26.2
1.134
[0.444, 1.977]
28.7
1.225
[0.424, 2.164]
43.4
1.836
[0.774, 3.211]
LAC
0.0
0.089
[0.000, 0.351]
1.2
0.344
[0.000, 1.026]
5.4
0.596
[0.041, 1.386]
2.3
0.377
[0.005, 0.946]
0.7
0.331
[0.000, 0.912]
0.2
0.198
[0.001, 0.614]
21.7
1.387
[0.460, 2.550]
MENA
0.4
0.298
[0.000, 0.899]
0.3
0.285
[0.002, 0.862]
12.0
0.894
[0.187, 1.787]
0.1
0.180
[0.001, 0.547]
1.7
0.341
[0.018, 0.889]
0.4
0.287
[0.000, 0.903]
0.6
0.315
[0.002, 0.911]
NA
0.2
0.241
[0.007, 0.754]
9.1
0.780
[0.117, 1.732]
3.3
0.585
[0.011, 1.571]
1.5
0.412
[0.008, 1.277]
1.0
0.332
[0.001, 0.978]
0.5
0.258
[0.001, 0.803]
10.4
0.899
[0.184, 1.952]
SAS
8.0
0.626
[0.094, 1.308]
38.3
2.395
[1.117, 3.943]
32.0
2.002
[0.865, 3.427]
0.3
0.238
[0.002, 0.707]
27.8
1.746
[0.847, 2.898]
0.2
0.218
[0.011, 0.687]
22.4
1.415
[0.542, 2.588]
USA
9.7
0.476
[0.085, 0.982]
69.9
2.556
[1.401, 4.063]
92.2
3.376
[1.983, 5.027]
85.7
3.149
[1.867, 4.623]
18.2
0.737
[0.182, 1.380]
109.9
4.018
[2.539, 5.726]
22.0
0.836
[0.298, 1.535]
BF category:
<3
[3, 10)
[10, 30)
[30, 100)
>=100

II-4c. B-Victoria



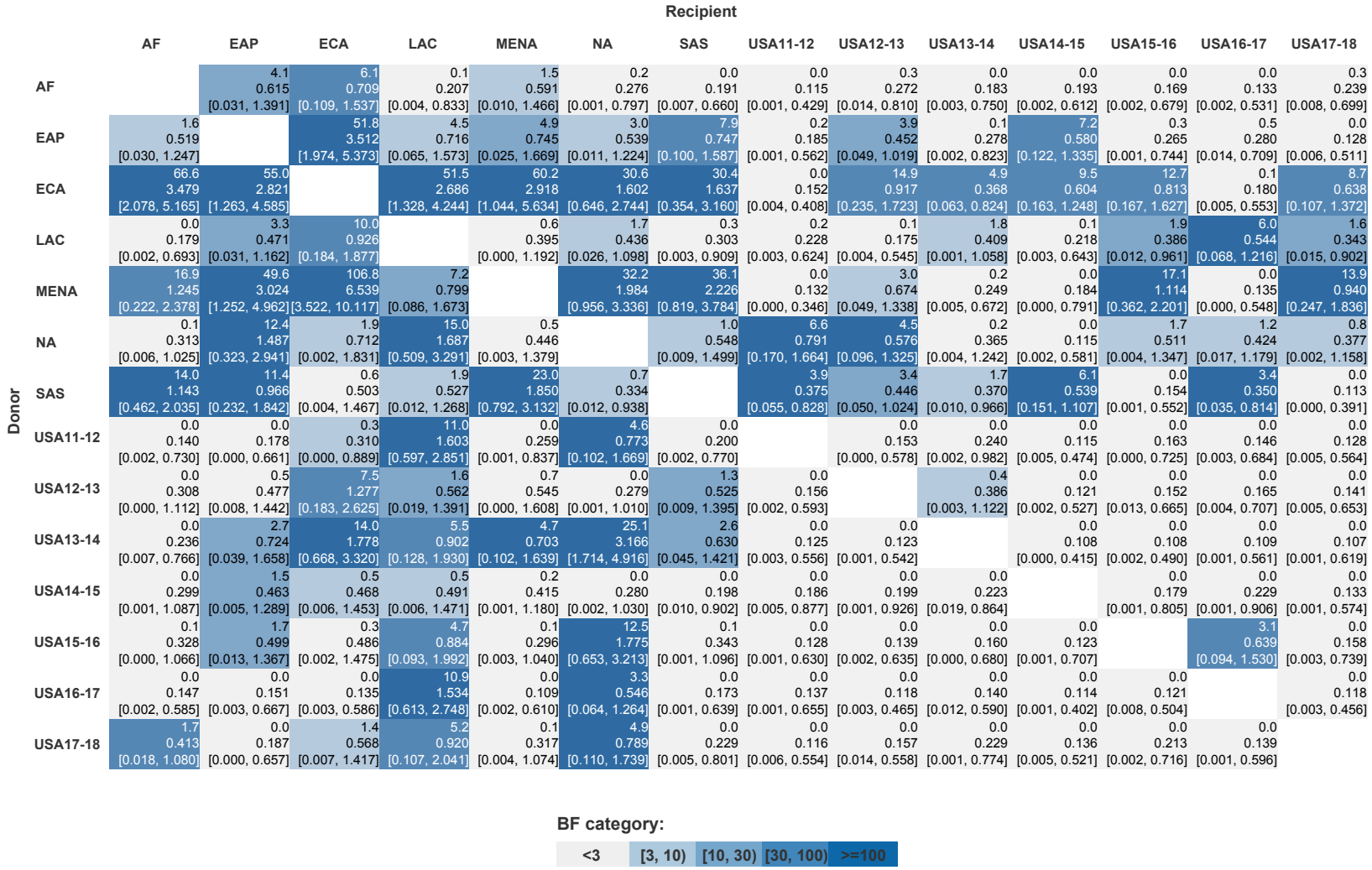
II-4d. B-Yamagata



Supplemental Table II-5. The full global transmission matrix and U.S. season-specific transmissions for each influenza type. The three items in each matrix cell in order are count of introductions, migration rate, and its 95% Bayesian Credible Interval (95%BCI). The rows of the table are the donor region and the columns are the recipient region. The cell color indicates BF levels, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

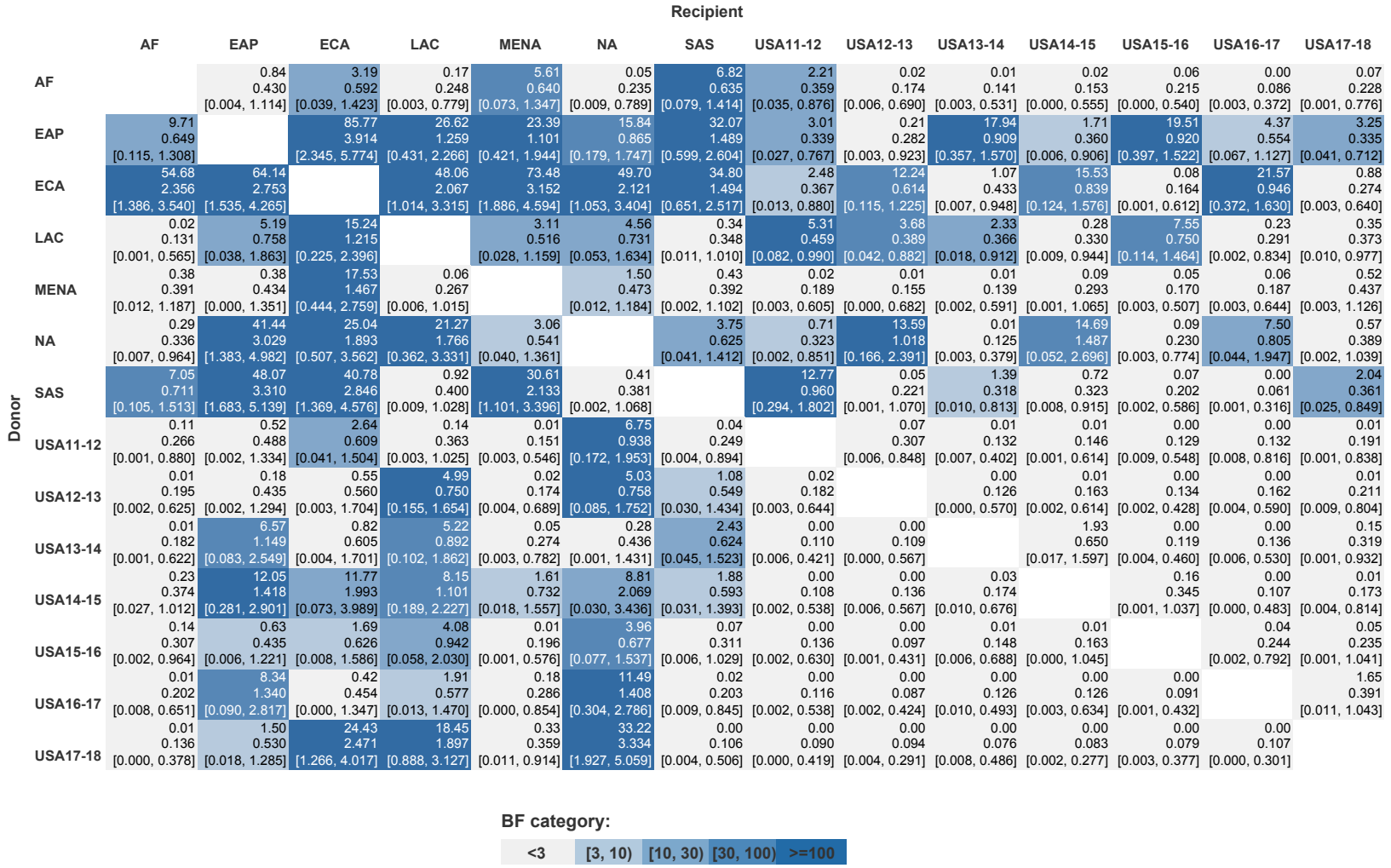
Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

II-5a. A/H1N1

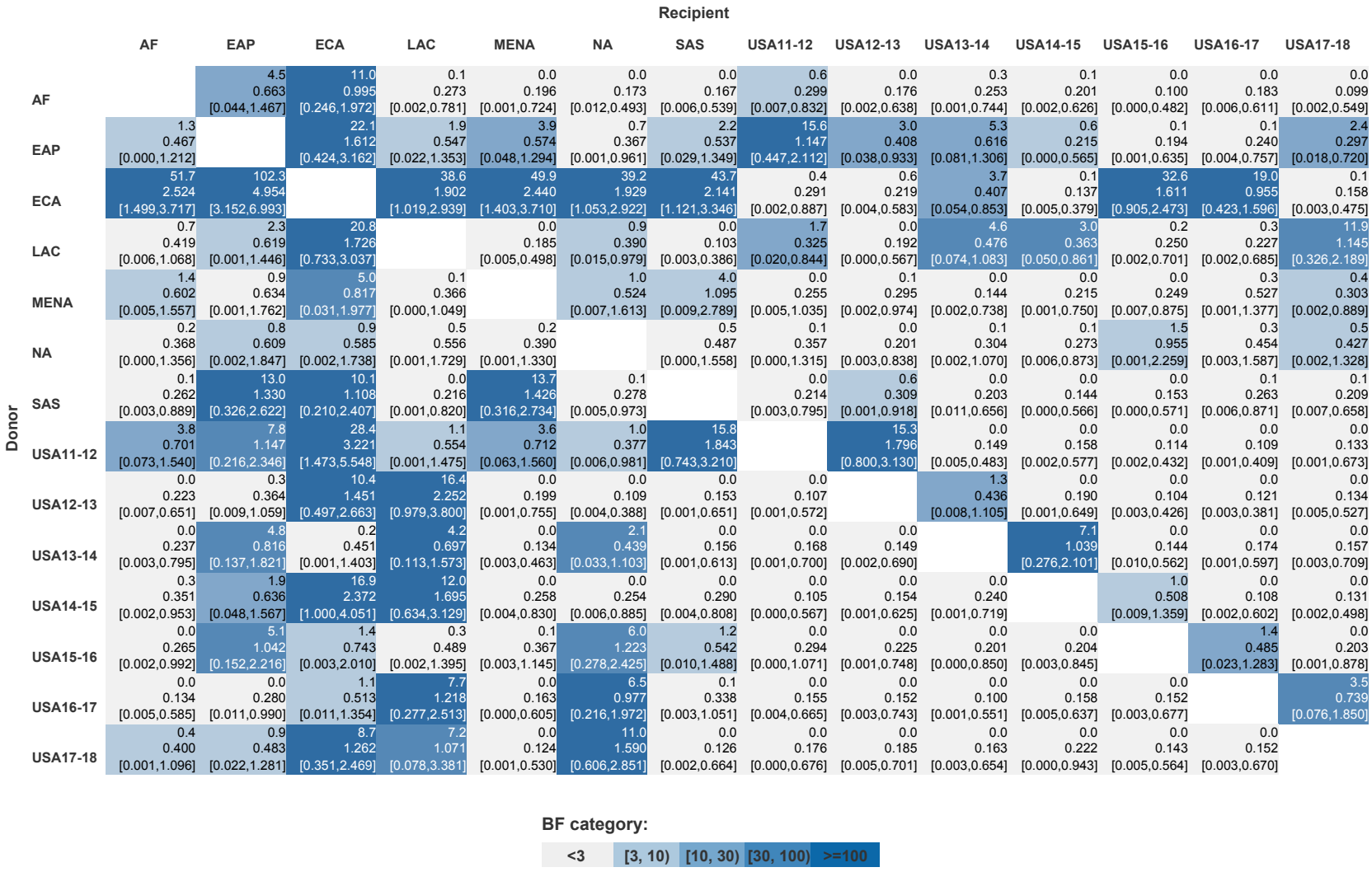


Donor \ Recipient	AF	EAP	ECA	LAC	MENA	NA	SAS	USA11-12	USA12-13	USA13-14	USA14-15	USA15-16	USA16-17	USA17-18
AF		4.1 0.615 [0.031, 1.391]	6.1 0.709 [0.109, 1.537]	0.1 0.207 [0.004, 0.833]	1.5 0.591 [0.010, 1.466]	0.2 0.276 [0.001, 0.797]	0.0 0.191 [0.007, 0.660]	0.0 0.115 [0.001, 0.429]	0.3 0.272 [0.014, 0.810]	0.0 0.183 [0.003, 0.750]	0.0 0.193 [0.002, 0.612]	0.0 0.169 [0.002, 0.679]	0.0 0.133 [0.002, 0.531]	0.3 0.239 [0.008, 0.699]
EAP	1.6 0.519 [0.030, 1.247]		51.8 3.512 [1.974, 5.373]	4.5 0.716 [0.065, 1.573]	4.9 0.745 [0.025, 1.669]	3.0 0.539 [0.011, 1.224]	7.9 0.747 [0.100, 1.587]	0.2 0.185 [0.001, 0.562]	3.9 0.452 [0.049, 1.019]	0.1 0.278 [0.002, 0.823]	7.2 0.580 [0.122, 1.335]	0.3 0.265 [0.001, 0.744]	0.5 0.280 [0.014, 0.709]	0.0 0.128 [0.006, 0.511]
ECA	66.6 3.479 [2.078, 5.165]	55.0 2.821 [1.263, 4.585]		51.5 2.686 [1.328, 4.244]	60.2 2.918 [1.044, 5.634]	30.6 1.602 [0.646, 2.744]	30.4 1.637 [0.354, 3.160]	0.0 0.152 [0.004, 0.408]	14.9 0.917 [0.235, 1.723]	4.9 0.368 [0.063, 0.824]	9.5 0.604 [0.163, 1.248]	12.7 0.813 [0.167, 1.627]	0.1 0.180 [0.005, 0.553]	8.7 0.638 [0.107, 1.372]
LAC	0.0 0.179 [0.002, 0.693]	3.3 0.471 [0.031, 1.162]	10.0 0.926 [0.184, 1.877]		0.6 0.395 [0.000, 1.192]	1.7 0.436 [0.026, 1.098]	0.3 0.303 [0.003, 0.909]	0.2 0.228 [0.003, 0.624]	0.1 0.175 [0.004, 0.545]	1.8 0.409 [0.001, 1.058]	0.1 0.218 [0.003, 0.643]	1.9 0.386 [0.012, 0.961]	6.0 0.544 [0.068, 1.216]	1.6 0.343 [0.015, 0.902]
MENA	16.9 1.245 [0.222, 2.378]	49.6 3.024 [1.252, 4.962]	106.8 6.539 [3.522, 10.117]	7.2 0.799 [0.086, 1.673]		32.2 1.984 [0.956, 3.336]	36.1 2.226 [0.819, 3.784]	0.0 0.132 [0.000, 0.346]	3.0 0.674 [0.049, 1.338]	0.2 0.249 [0.005, 0.672]	0.0 0.184 [0.000, 0.791]	17.1 1.114 [0.362, 2.201]	0.0 0.135 [0.000, 0.548]	13.9 0.940 [0.247, 1.836]
NA	0.1 0.313 [0.006, 1.025]	12.4 1.487 [0.323, 2.941]	1.9 0.712 [0.002, 1.831]	15.0 1.687 [0.509, 3.291]	0.5 0.446 [0.003, 1.379]		1.0 0.548 [0.009, 1.499]	6.6 0.791 [0.170, 1.664]	4.5 0.576 [0.096, 1.325]	0.2 0.365 [0.004, 1.242]	0.0 0.115 [0.002, 0.581]	1.7 0.511 [0.004, 1.347]	1.2 0.424 [0.017, 1.179]	0.8 0.377 [0.002, 1.158]
SAS	14.0 1.143 [0.462, 2.035]	11.4 0.966 [0.232, 1.842]	0.6 0.503 [0.004, 1.467]	1.9 0.527 [0.012, 1.268]	23.0 1.850 [0.792, 3.132]	0.7 0.334 [0.012, 0.938]		3.9 0.375 [0.055, 0.828]	3.4 0.446 [0.050, 1.024]	1.7 0.370 [0.010, 0.966]	6.1 0.539 [0.151, 1.107]	0.0 0.154 [0.001, 0.552]	3.4 0.350 [0.035, 0.814]	0.0 0.113 [0.000, 0.391]
USA11-12	0.0 0.140 [0.002, 0.730]	0.0 0.178 [0.000, 0.661]	0.3 0.310 [0.000, 0.889]	11.0 1.603 [0.597, 2.851]	0.0 0.259 [0.001, 0.837]	4.6 0.773 [0.102, 1.669]	0.0 0.200 [0.002, 0.770]		0.0 0.153 [0.000, 0.578]	0.0 0.240 [0.002, 0.982]	0.0 0.115 [0.005, 0.474]	0.0 0.163 [0.000, 0.725]	0.0 0.146 [0.003, 0.684]	0.0 0.128 [0.005, 0.564]
USA12-13	0.0 0.308 [0.000, 1.112]	0.5 0.477 [0.008, 1.442]	7.5 1.277 [0.183, 2.625]	1.6 0.562 [0.019, 1.391]	0.7 0.545 [0.000, 1.608]	0.0 0.279 [0.001, 1.010]	1.3 0.525 [0.009, 1.395]	0.0 0.156 [0.002, 0.593]		0.4 0.386 [0.003, 1.122]	0.0 0.121 [0.002, 0.527]	0.0 0.152 [0.013, 0.665]	0.0 0.165 [0.004, 0.707]	0.0 0.141 [0.005, 0.653]
USA13-14	0.0 0.236 [0.007, 0.766]	2.7 0.724 [0.039, 1.658]	14.0 1.778 [0.668, 3.320]	5.5 0.902 [0.128, 1.930]	4.7 0.703 [0.102, 1.639]	25.1 3.166 [1.714, 4.916]	2.6 0.630 [0.045, 1.421]	0.0 0.125 [0.003, 0.556]	0.0 0.123 [0.001, 0.542]		0.0 0.108 [0.000, 0.415]	0.0 0.108 [0.002, 0.490]	0.0 0.109 [0.001, 0.561]	0.0 0.107 [0.001, 0.619]
USA14-15	0.0 0.299 [0.001, 1.087]	1.5 0.463 [0.005, 1.289]	0.5 0.468 [0.006, 1.453]	0.5 0.491 [0.006, 1.471]	0.2 0.415 [0.001, 1.180]	0.0 0.280 [0.002, 1.030]	0.0 0.198 [0.010, 0.902]	0.0 0.186 [0.005, 0.877]	0.0 0.199 [0.001, 0.926]	0.0 0.223 [0.019, 0.864]		0.0 0.179 [0.001, 0.805]	0.0 0.229 [0.001, 0.906]	0.0 0.133 [0.001, 0.574]
USA15-16	0.1 0.328 [0.000, 1.066]	1.7 0.499 [0.013, 1.367]	0.3 0.486 [0.002, 1.475]	4.7 0.884 [0.093, 1.992]	0.1 0.296 [0.003, 1.040]	12.5 1.775 [0.653, 3.213]	0.1 0.343 [0.001, 1.096]	0.0 0.128 [0.001, 0.630]	0.0 0.139 [0.002, 0.635]	0.0 0.160 [0.000, 0.680]	0.0 0.123 [0.001, 0.707]		3.1 0.639 [0.094, 1.530]	0.0 0.158 [0.003, 0.739]
USA16-17	0.0 0.147 [0.002, 0.585]	0.0 0.151 [0.003, 0.667]	0.0 0.135 [0.003, 0.586]	10.9 1.534 [0.613, 2.748]	0.0 0.109 [0.002, 0.610]	3.3 0.546 [0.064, 1.264]	0.0 0.173 [0.001, 0.639]	0.0 0.137 [0.001, 0.655]	0.0 0.118 [0.003, 0.465]	0.0 0.140 [0.012, 0.590]	0.0 0.114 [0.001, 0.402]	0.0 0.121 [0.008, 0.504]		0.0 0.118 [0.003, 0.456]
USA17-18	1.7 0.413 [0.018, 1.080]	0.0 0.187 [0.000, 0.657]	1.4 0.568 [0.007, 1.417]	5.2 0.920 [0.107, 2.041]	0.1 0.317 [0.004, 1.074]	4.9 0.789 [0.110, 1.739]	0.0 0.229 [0.005, 0.801]	0.0 0.116 [0.006, 0.554]	0.0 0.157 [0.014, 0.558]	0.0 0.229 [0.001, 0.774]	0.0 0.136 [0.005, 0.521]	0.0 0.213 [0.002, 0.716]	0.0 0.139 [0.001, 0.596]	

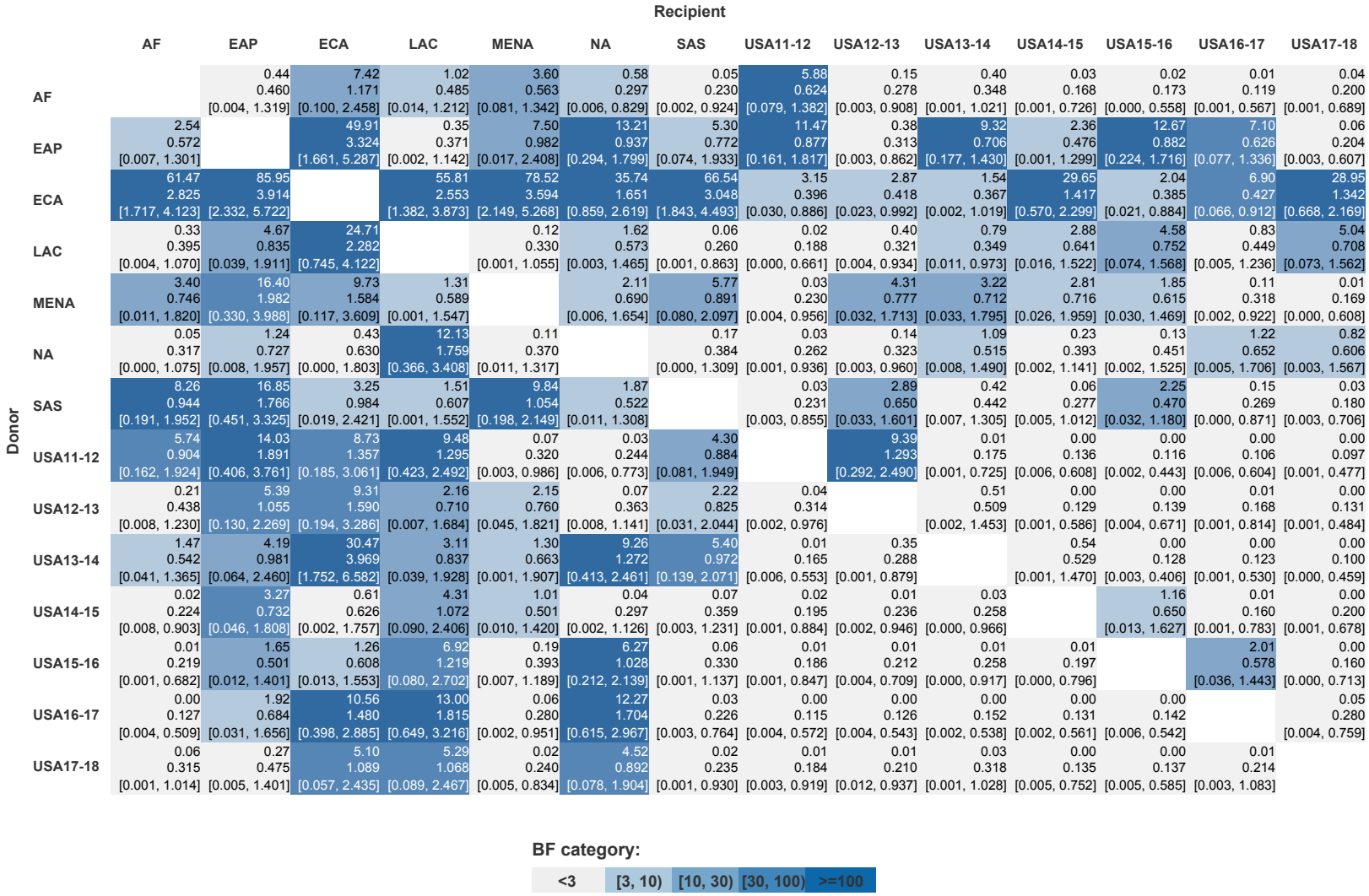
II-5b. A/H3N2


Recipient
AF EAP ECA LAC MENA NA SAS USA11-12 USA12-13 USA13-14 USA14-15 USA15-16 USA16-17 USA17-18
Donor
AF 0.84 0.430 [0.004, 1.114] 3.19 0.592 [0.039, 1.423] 0.17 0.248 [0.003, 0.779] 5.61 0.640 [0.073, 1.347] 0.05 0.235 [0.009, 0.789] 6.82 0.635 [0.079, 1.414] 2.21 0.359 [0.035, 0.876] 0.02 0.174 [0.006, 0.690] 0.01 0.141 [0.003, 0.531] 0.02 0.153 [0.000, 0.555] 0.06 0.215 [0.000, 0.540] 0.00 0.086 [0.003, 0.372] 0.07 0.228 [0.001, 0.776]
EAP 9.71 0.649 [0.115, 1.308] 85.77 3.914 [2.345, 5.774] 26.62 1.259 [0.431, 2.266] 23.39 1.101 [0.421, 1.944] 15.84 0.865 [0.179, 1.747] 32.07 1.489 [0.599, 2.604] 3.01 0.339 [0.027, 0.767] 0.21 0.282 [0.003, 0.923] 17.94 0.909 [0.357, 1.570] 1.71 0.360 [0.006, 0.906] 19.51 0.920 [0.397, 1.522] 4.37 0.554 [0.067, 1.127] 3.25 0.335 [0.041, 0.712]
ECA 54.68 2.356 [1.386, 3.540] 64.14 2.753 [1.535, 4.265] 48.06 2.067 [1.014, 3.315] 73.48 3.152 [1.886, 4.594] 49.70 2.121 [1.053, 3.404] 34.80 1.494 [0.651, 2.517] 2.48 0.367 [0.013, 0.880] 12.24 0.614 [0.115, 1.225] 1.07 0.433 [0.007, 0.948] 15.53 0.839 [0.124, 1.576] 0.08 0.164 [0.001, 0.612] 21.57 0.946 [0.372, 1.630] 0.88 0.274 [0.003, 0.640]
LAC 0.02 0.131 [0.001, 0.565] 5.19 0.758 [0.038, 1.863] 15.24 1.215 [0.225, 2.396] 3.11 0.516 [0.028, 1.159] 4.56 0.731 [0.053, 1.634] 0.34 0.348 [0.011, 1.010] 5.31 0.459 [0.082, 0.990] 3.68 0.389 [0.042, 0.882] 2.33 0.366 [0.018, 0.912] 0.28 0.330 [0.009, 0.944] 7.55 0.750 [0.114, 1.464] 0.23 0.291 [0.002, 0.834] 0.35 0.373 [0.010, 0.977]
MENA 0.38 0.391 [0.012, 1.187] 0.38 0.434 [0.000, 1.351] 17.53 1.467 [0.444, 2.759] 0.06 0.267 [0.006, 1.015] 1.50 0.473 [0.012, 1.184] 0.43 0.392 [0.002, 1.102] 0.02 0.189 [0.003, 0.605] 0.01 0.155 [0.000, 0.682] 0.01 0.139 [0.002, 0.591] 0.09 0.293 [0.001, 1.065] 0.05 0.170 [0.003, 0.507] 0.06 0.187 [0.003, 0.644] 0.52 0.437 [0.003, 1.126]
NA 0.29 0.336 [0.007, 0.964] 41.44 3.029 [1.383, 4.982] 25.04 1.893 [0.507, 3.562] 21.27 1.766 [0.362, 3.331] 3.06 0.541 [0.040, 1.361] 3.75 0.625 [0.041, 1.412] 0.71 0.323 [0.002, 0.851] 13.59 1.018 [0.166, 2.391] 0.01 0.125 [0.003, 0.379] 14.69 1.487 [0.052, 2.696] 0.09 0.230 [0.003, 0.774] 7.50 0.805 [0.044, 1.947] 0.57 0.389 [0.002, 1.039]
SAS 7.05 0.711 [0.105, 1.513] 48.07 3.310 [1.683, 5.139] 40.78 2.846 [1.369, 4.576] 0.92 0.400 [0.009, 1.028] 30.61 2.133 [1.101, 3.396] 0.41 0.381 [0.002, 1.068] 12.77 0.960 [0.294, 1.802] 0.05 0.221 [0.001, 1.070] 1.39 0.318 [0.010, 0.813] 0.72 0.323 [0.008, 0.915] 0.07 0.202 [0.002, 0.586] 0.00 0.061 [0.001, 0.316] 2.04 0.361 [0.025, 0.849]
USA11-12 0.11 0.266 [0.001, 0.880] 0.52 0.488 [0.002, 1.334] 2.64 0.609 [0.041, 1.504] 0.14 0.363 [0.003, 1.025] 0.01 0.151 [0.003, 0.546] 6.75 0.938 [0.172, 1.953] 0.04 0.249 [0.004, 0.894] 0.07 0.307 [0.006, 0.848] 0.01 0.132 [0.007, 0.402] 0.01 0.146 [0.001, 0.614] 0.00 0.129 [0.009, 0.548] 0.00 0.132 [0.008, 0.816] 0.01 0.191 [0.001, 0.838]
USA12-13 0.01 0.195 [0.002, 0.625] 0.18 0.435 [0.002, 1.294] 0.55 0.560 [0.003, 1.704] 4.99 0.750 [0.155, 1.654] 0.02 0.174 [0.004, 0.689] 5.03 0.758 [0.085, 1.752] 1.08 0.549 [0.030, 1.434] 0.02 0.182 [0.003, 0.644] 0.00 0.126 [0.000, 0.570] 0.01 0.163 [0.002, 0.614] 0.00 0.134 [0.002, 0.428] 0.00 0.162 [0.004, 0.590] 0.01 0.211 [0.009, 0.804]
USA13-14 0.01 0.182 [0.001, 0.622] 6.57 1.149 [0.083, 2.549] 0.82 0.605 [0.004, 1.701] 5.22 0.892 [0.102, 1.862] 0.05 0.274 [0.003, 0.782] 0.28 0.436 [0.001, 1.431] 2.43 0.624 [0.045, 1.523] 0.00 0.110 [0.006, 0.421] 0.00 0.109 [0.000, 0.567] 1.93 0.650 [0.017, 1.597] 0.00 0.119 [0.004, 0.460] 0.00 0.136 [0.006, 0.530] 0.15 0.319 [0.001, 0.932]
USA14-15 0.23 0.374 [0.027, 1.012] 12.05 1.418 [0.281, 2.901] 11.77 1.993 [0.073, 3.989] 8.15 1.101 [0.189, 2.227] 1.61 0.732 [0.018, 1.557] 8.81 2.069 [0.030, 3.436] 1.88 0.593 [0.031, 1.393] 0.00 0.108 [0.002, 0.538] 0.00 0.136 [0.006, 0.567] 0.03 0.174 [0.010, 0.676] 0.16 0.345 [0.001, 1.037] 0.00 0.107 [0.000, 0.483] 0.01 0.173 [0.004, 0.814]
USA15-16 0.14 0.307 [0.002, 0.964] 0.63 0.435 [0.006, 1.221] 1.69 0.626 [0.008, 1.586] 4.08 0.942 [0.058, 2.030] 0.01 0.196 [0.001, 0.576] 3.96 0.677 [0.077, 1.537] 0.07 0.311 [0.006, 1.029] 0.00 0.136 [0.002, 0.630] 0.00 0.097 [0.001, 0.431] 0.01 0.148 [0.006, 0.688] 0.01 0.163 [0.000, 1.045] 0.04 0.244 [0.002, 0.792] 0.05 0.235 [0.001, 1.041]
USA16-17 0.01 0.202 [0.008, 0.651] 8.34 1.340 [0.090, 2.817] 0.42 0.454 [0.000, 1.347] 1.91 0.577 [0.013, 1.470] 0.18 0.286 [0.000, 0.854] 11.49 1.408 [0.304, 2.786] 0.02 0.203 [0.009, 0.845] 0.00 0.116 [0.002, 0.538] 0.00 0.087 [0.002, 0.424] 0.00 0.126 [0.010, 0.493] 0.00 0.126 [0.003, 0.634] 0.00 0.091 [0.001, 0.432] 1.65 0.391 [0.011, 1.043]
USA17-18 0.01 0.136 [0.000, 0.378] 1.50 0.530 [0.018, 1.285] 24.43 2.471 [1.266, 4.017] 18.45 1.897 [0.888, 3.127] 0.33 0.359 [0.011, 0.914] 33.22 3.334 [1.927, 5.059] 0.00 0.106 [0.004, 0.506] 0.00 0.090 [0.000, 0.419] 0.00 0.094 [0.004, 0.291] 0.00 0.076 [0.008, 0.486] 0.00 0.083 [0.002, 0.277] 0.00 0.079 [0.003, 0.377] 0.00 0.107 [0.000, 0.301]
BF category:
<3 [3, 10) [10, 30) [30, 100) >=100

II-5c. B-Victoria

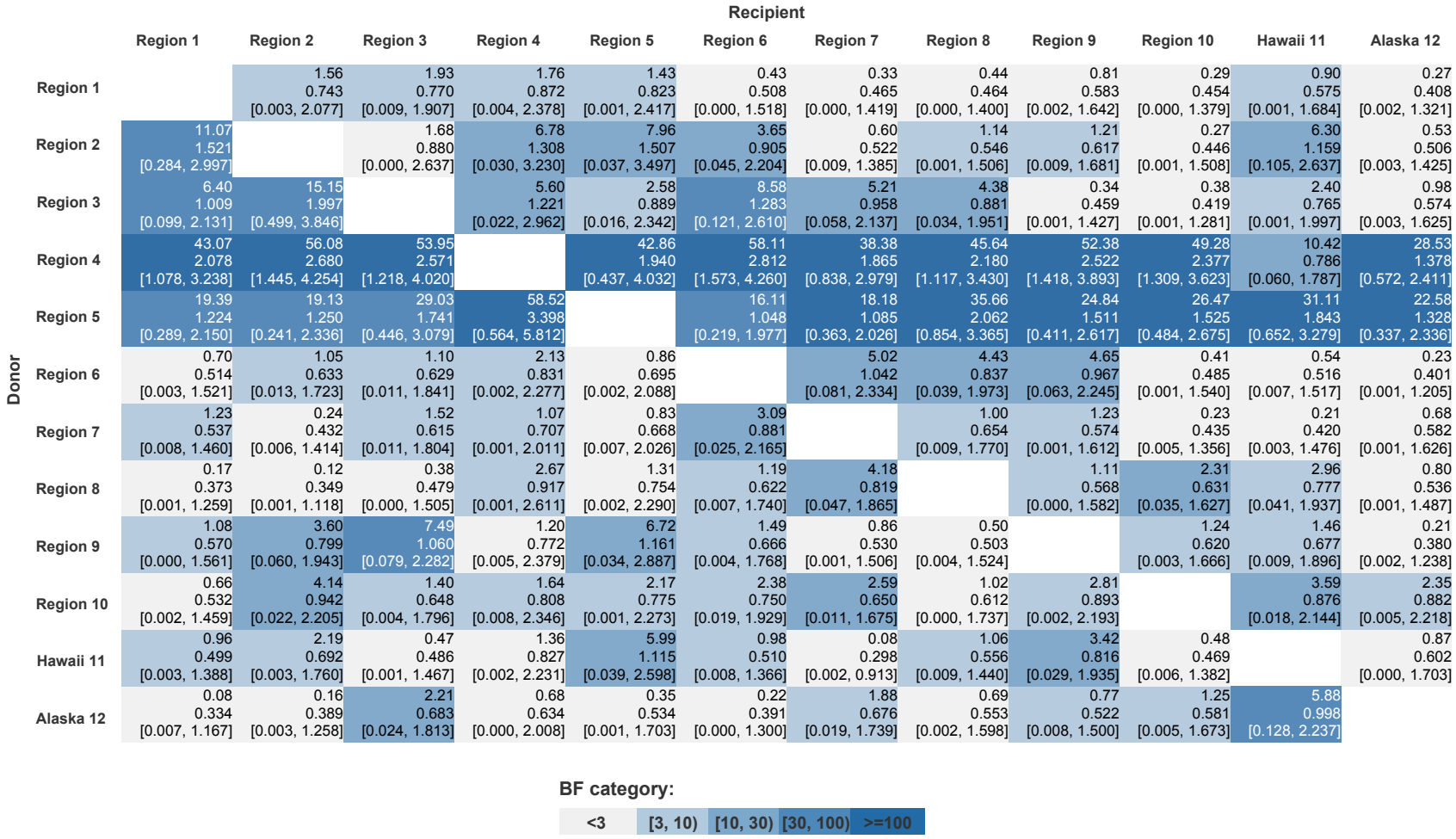


II-5d. B-Yamagata

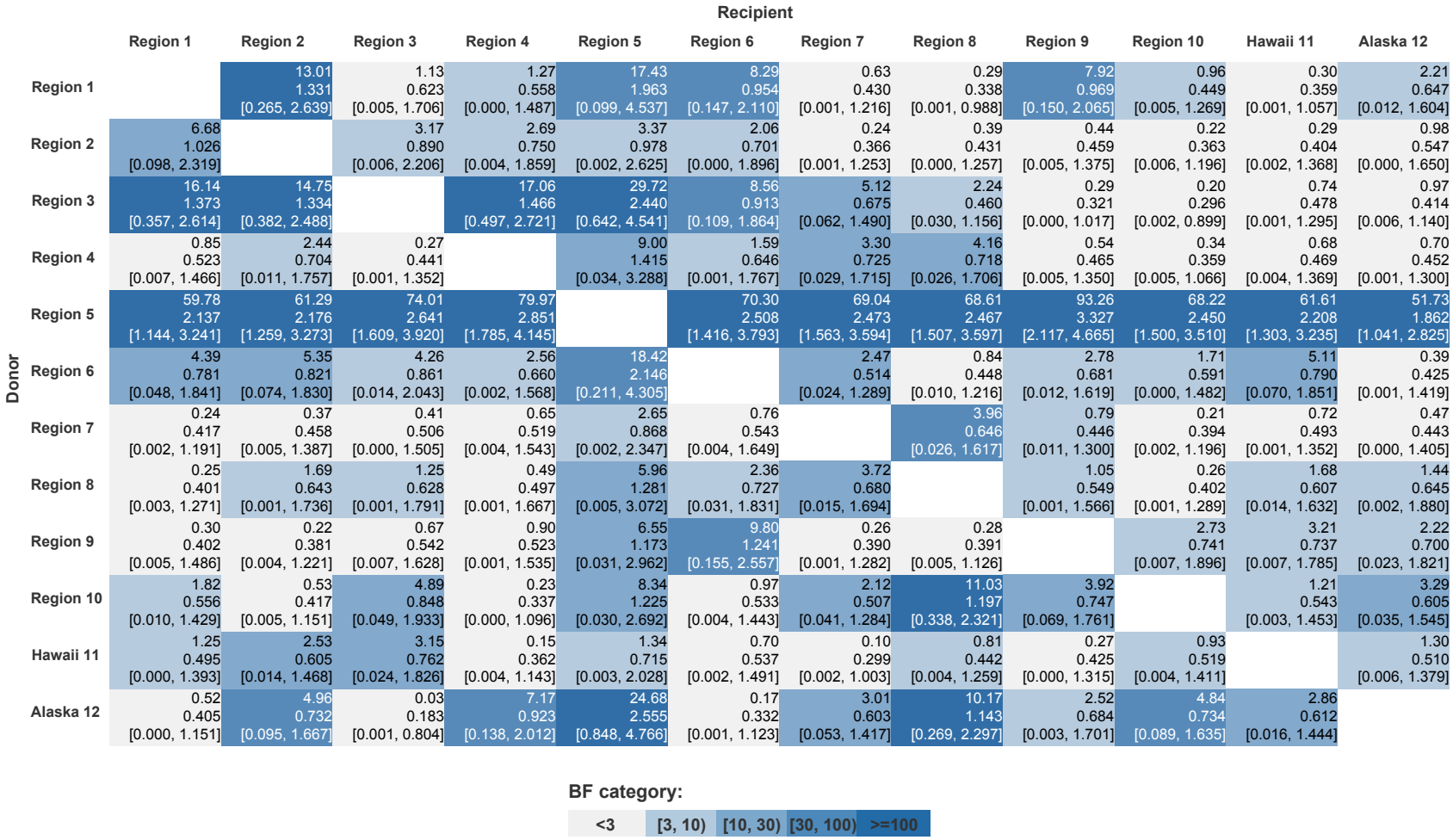


Supplemental Table II-6. The full U.S. region transmission matrix for each influenza type. The three items in each matrix cell in order are count of introductions, migration rate, and its 95% Bayesian Credible Interval (95%BCI). The rows of the table are the donor region and the columns are the recipient region. The cell color indicates BF levels, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

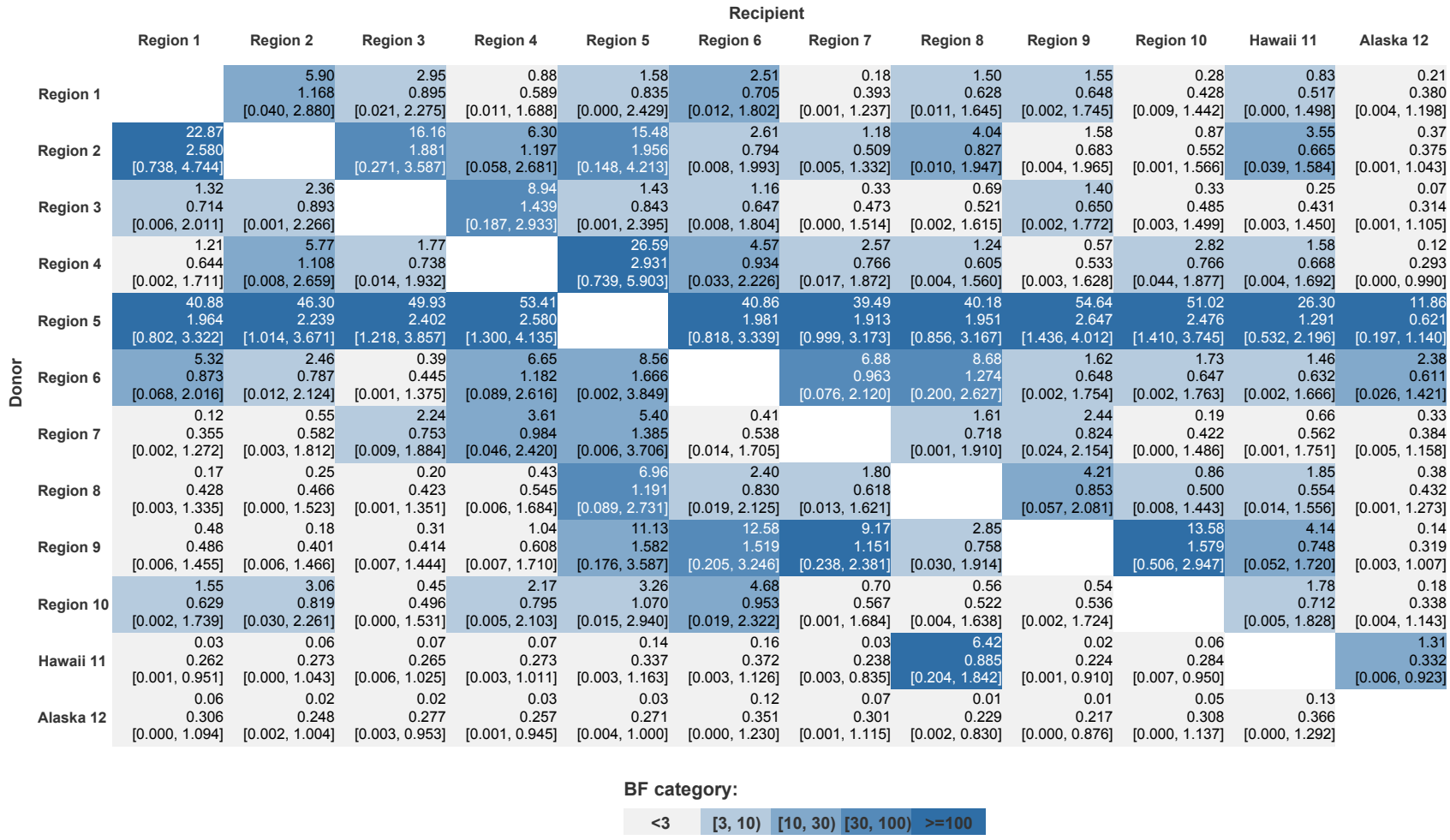
II-6a. A/H1N1


Recipient
Region 1
Region 2
Region 3
Region 4
Region 5
Region 6
Region 7
Region 8
Region 9
Region 10
Hawaii 11
Alaska 12
Donor
Region 1
1.56 0.743 [0.003, 2.077]
1.93 0.770 [0.009, 1.907]
1.76 0.872 [0.004, 2.378]
1.43 0.823 [0.001, 2.417]
0.43 0.508 [0.000, 1.518]
0.33 0.465 [0.000, 1.419]
0.44 0.464 [0.000, 1.400]
0.81 0.583 [0.002, 1.642]
0.29 0.454 [0.000, 1.379]
0.90 0.575 [0.001, 1.684]
0.27 0.408 [0.002, 1.321]
Region 2
11.07 1.521 [0.284, 2.997]
1.68 0.880 [0.000, 2.637]
6.78 1.308 [0.030, 3.230]
7.96 1.507 [0.037, 3.497]
3.65 0.905 [0.045, 2.204]
0.60 0.522 [0.009, 1.385]
1.14 0.546 [0.001, 1.506]
1.21 0.617 [0.009, 1.681]
0.27 0.446 [0.001, 1.508]
6.30 1.159 [0.105, 2.637]
0.53 0.506 [0.003, 1.425]
Region 3
6.40 1.009 [0.099, 2.131]
15.15 1.997 [0.499, 3.846]
5.60 1.221 [0.022, 2.962]
2.58 0.889 [0.016, 2.342]
8.58 1.283 [0.121, 2.610]
5.21 0.958 [0.058, 2.137]
4.38 0.881 [0.034, 1.951]
0.34 0.459 [0.001, 1.427]
0.38 0.419 [0.001, 1.281]
2.40 0.765 [0.001, 1.997]
0.98 0.574 [0.003, 1.625]
Region 4
43.07 2.078 [1.078, 3.238]
56.08 2.680 [1.445, 4.254]
53.95 2.571 [1.218, 4.020]
42.86 1.940 [0.437, 4.032]
58.11 2.812 [1.573, 4.260]
38.38 1.865 [0.838, 2.979]
45.64 2.180 [1.117, 3.430]
52.38 2.522 [1.418, 3.893]
49.28 2.377 [1.309, 3.623]
10.42 0.786 [0.060, 1.787]
28.53 1.378 [0.572, 2.411]
Region 5
19.39 1.224 [0.289, 2.150]
19.13 1.250 [0.241, 2.336]
29.03 1.741 [0.446, 3.079]
58.52 3.398 [0.564, 5.812]
16.11 1.048 [0.219, 1.977]
18.18 1.085 [0.363, 2.026]
35.66 2.062 [0.854, 3.365]
24.84 1.511 [0.411, 2.617]
26.47 1.525 [0.484, 2.675]
31.11 1.843 [0.652, 3.279]
22.58 1.328 [0.337, 2.336]
Region 6
0.70 0.514 [0.003, 1.521]
1.05 0.633 [0.013, 1.723]
1.10 0.629 [0.011, 1.841]
2.13 0.831 [0.002, 2.277]
0.86 0.695 [0.002, 2.088]
5.02 1.042 [0.081, 2.334]
4.43 0.837 [0.039, 1.973]
4.65 0.967 [0.063, 2.245]
0.41 0.485 [0.001, 1.540]
0.54 0.516 [0.007, 1.517]
0.23 0.401 [0.001, 1.205]
Region 7
1.23 0.537 [0.008, 1.460]
0.24 0.432 [0.006, 1.414]
1.52 0.615 [0.011, 1.804]
1.07 0.707 [0.001, 2.011]
0.83 0.668 [0.007, 2.026]
3.09 0.881 [0.025, 2.165]
1.00 0.654 [0.009, 1.770]
1.23 0.574 [0.001, 1.612]
0.23 0.435 [0.005, 1.356]
0.21 0.420 [0.003, 1.476]
0.68 0.582 [0.001, 1.626]
Region 8
0.17 0.373 [0.001, 1.259]
0.12 0.349 [0.001, 1.118]
0.38 0.479 [0.000, 1.505]
2.67 0.917 [0.001, 2.611]
1.31 0.754 [0.002, 2.290]
1.19 0.622 [0.007, 1.740]
4.18 0.819 [0.047, 1.865]
1.11 0.568 [0.000, 1.582]
2.31 0.631 [0.035, 1.627]
2.96 0.777 [0.041, 1.937]
0.80 0.536 [0.001, 1.487]
Region 9
1.08 0.570 [0.000, 1.561]
3.60 0.799 [0.060, 1.943]
7.49 1.060 [0.079, 2.282]
1.20 0.772 [0.005, 2.379]
6.72 1.161 [0.034, 2.887]
1.49 0.666 [0.004, 1.768]
0.86 0.530 [0.001, 1.506]
0.50 0.503 [0.004, 1.524]
1.24 0.620 [0.003, 1.666]
1.46 0.677 [0.009, 1.896]
0.21 0.380 [0.002, 1.238]
Region 10
0.66 0.532 [0.002, 1.459]
4.14 0.942 [0.022, 2.205]
1.40 0.648 [0.004, 1.796]
1.64 0.808 [0.008, 2.346]
2.17 0.775 [0.001, 2.273]
2.38 0.750 [0.019, 1.929]
2.59 0.650 [0.011, 1.675]
1.02 0.612 [0.000, 1.737]
2.81 0.893 [0.002, 2.193]
3.59 0.876 [0.018, 2.144]
2.35 0.882 [0.005, 2.218]
Hawaii 11
0.96 0.499 [0.003, 1.388]
2.19 0.692 [0.003, 1.760]
0.47 0.486 [0.001, 1.467]
1.36 0.827 [0.002, 2.231]
5.99 1.115 [0.039, 2.598]
0.98 0.510 [0.008, 1.366]
0.08 0.298 [0.002, 0.913]
1.06 0.556 [0.009, 1.440]
3.42 0.816 [0.029, 1.935]
0.48 0.469 [0.006, 1.382]
0.87 0.602 [0.000, 1.703]
Alaska 12
0.08 0.334 [0.007, 1.167]
0.16 0.389 [0.003, 1.258]
2.21 0.683 [0.024, 1.813]
0.68 0.634 [0.000, 2.008]
0.35 0.534 [0.001, 1.703]
0.22 0.391 [0.000, 1.300]
1.88 0.676 [0.019, 1.739]
0.69 0.553 [0.002, 1.598]
0.77 0.522 [0.008, 1.500]
1.25 0.581 [0.005, 1.673]
5.88 0.998 [0.128, 2.237]
BF category:
<3
[3, 10)
[10, 30)
[30, 100)
>=100

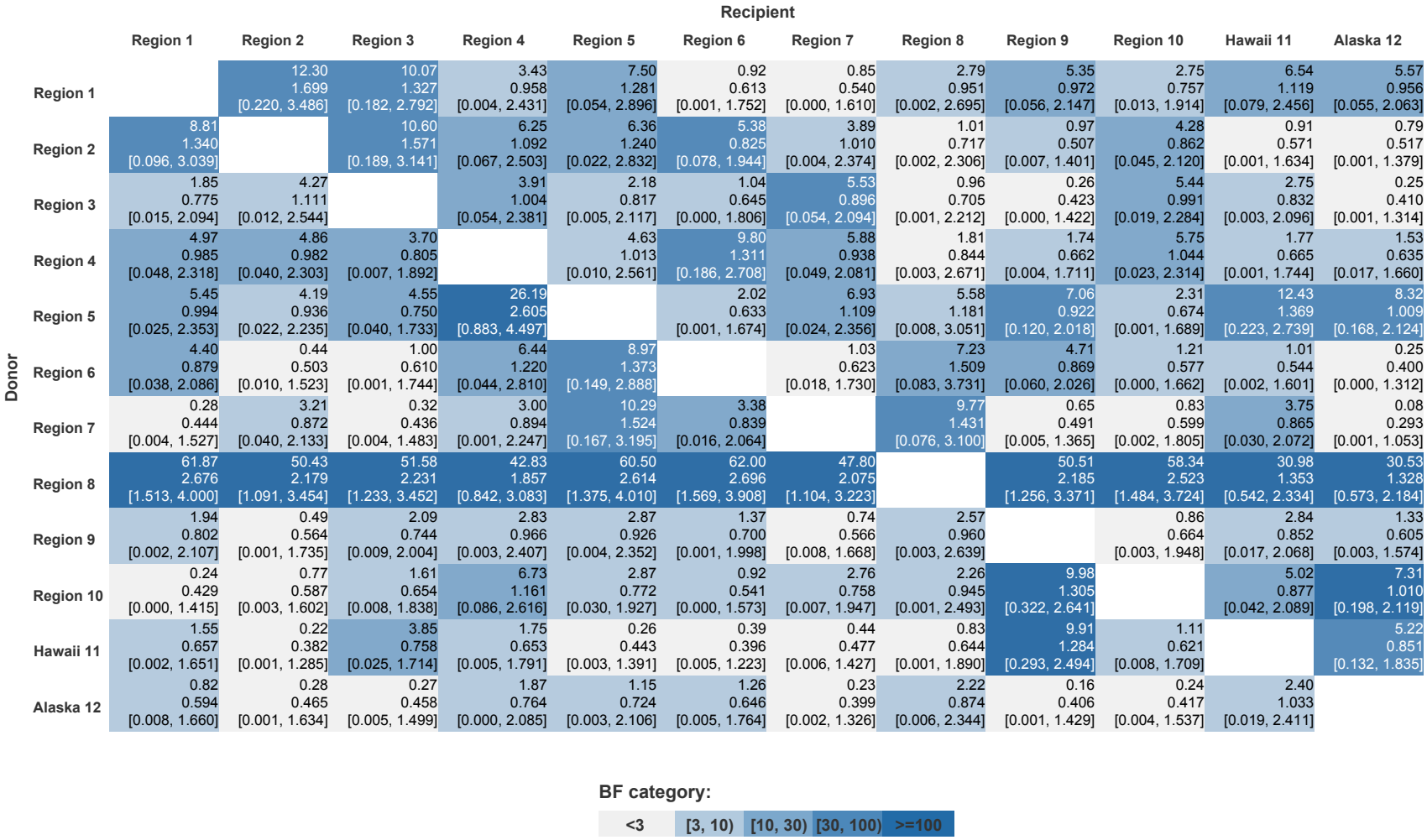
II-6b. A/H3N2


Recipient
Donor
Region 1
Region 2
Region 3
Region 4
Region 5
Region 6
Region 7
Region 8
Region 9
Region 10
Hawaii 11
Alaska 12
13.01 1.331 [0.265, 2.639]
1.13 0.623 [0.005, 1.706]
1.27 0.558 [0.000, 1.487]
17.43 1.963 [0.099, 4.537]
8.29 0.954 [0.147, 2.110]
0.63 0.430 [0.001, 1.216]
0.29 0.338 [0.001, 0.988]
7.92 0.969 [0.150, 2.065]
0.96 0.449 [0.005, 1.269]
0.30 0.359 [0.001, 1.057]
2.21 0.647 [0.012, 1.604]
6.68 1.026 [0.098, 2.319]
3.17 0.890 [0.006, 2.206]
2.69 0.750 [0.004, 1.859]
3.37 0.978 [0.002, 2.625]
2.06 0.701 [0.000, 1.896]
0.24 0.366 [0.001, 1.253]
0.39 0.431 [0.000, 1.257]
0.44 0.459 [0.005, 1.375]
0.22 0.363 [0.006, 1.196]
0.29 0.404 [0.002, 1.368]
0.98 0.547 [0.000, 1.650]
16.14 1.373 [0.357, 2.614]
14.75 1.334 [0.382, 2.488]
17.06 1.466 [0.497, 2.721]
29.72 2.440 [0.642, 4.541]
8.56 0.913 [0.109, 1.864]
5.12 0.675 [0.062, 1.490]
2.24 0.460 [0.030, 1.156]
0.29 0.321 [0.000, 1.017]
0.20 0.296 [0.002, 0.899]
0.74 0.478 [0.001, 1.295]
0.97 0.414 [0.006, 1.140]
0.85 0.523 [0.007, 1.466]
2.44 0.704 [0.011, 1.757]
0.27 0.441 [0.001, 1.352]
9.00 1.415 [0.034, 3.288]
1.59 0.646 [0.001, 1.767]
3.30 0.725 [0.029, 1.715]
4.16 0.718 [0.026, 1.706]
0.54 0.465 [0.005, 1.350]
0.34 0.359 [0.005, 1.066]
0.68 0.469 [0.004, 1.369]
0.70 0.452 [0.001, 1.300]
59.78 2.137 [1.144, 3.241]
61.29 2.176 [1.259, 3.273]
74.01 2.641 [1.609, 3.920]
79.97 2.851 [1.785, 4.145]
70.30 2.508 [1.416, 3.793]
69.04 2.473 [1.563, 3.594]
68.61 2.467 [1.507, 3.597]
93.26 3.327 [2.117, 4.665]
68.22 2.450 [1.500, 3.510]
61.61 2.208 [1.303, 3.235]
51.73 1.862 [1.041, 2.825]
4.39 0.781 [0.048, 1.841]
5.35 0.821 [0.074, 1.830]
4.26 0.861 [0.014, 2.043]
2.56 0.660 [0.002, 1.568]
18.42 2.146 [0.211, 4.305]
2.47 0.514 [0.024, 1.289]
0.84 0.448 [0.010, 1.216]
2.78 0.681 [0.012, 1.619]
1.71 0.591 [0.000, 1.482]
5.11 0.790 [0.070, 1.851]
0.39 0.425 [0.001, 1.419]
0.24 0.417 [0.002, 1.191]
0.37 0.458 [0.005, 1.387]
0.41 0.506 [0.000, 1.505]
0.65 0.519 [0.004, 1.543]
2.65 0.868 [0.002, 2.347]
0.76 0.543 [0.004, 1.649]
3.96 0.646 [0.026, 1.617]
0.79 0.446 [0.011, 1.300]
0.21 0.394 [0.002, 1.196]
0.72 0.493 [0.001, 1.352]
0.47 0.443 [0.000, 1.405]
0.25 0.401 [0.003, 1.271]
1.69 0.643 [0.001, 1.736]
1.25 0.628 [0.001, 1.791]
0.49 0.497 [0.001, 1.667]
5.96 1.281 [0.005, 3.072]
2.36 0.727 [0.031, 1.831]
3.72 0.680 [0.015, 1.694]
1.05 0.549 [0.001, 1.566]
0.26 0.402 [0.001, 1.289]
1.68 0.607 [0.014, 1.632]
1.44 0.645 [0.002, 1.880]
0.30 0.402 [0.005, 1.486]
0.22 0.381 [0.004, 1.221]
0.67 0.542 [0.007, 1.628]
0.90 0.523 [0.001, 1.535]
6.55 1.173 [0.031, 2.962]
9.80 1.241 [0.155, 2.557]
0.26 0.390 [0.001, 1.282]
0.28 0.391 [0.005, 1.126]
2.73 0.741 [0.007, 1.896]
3.21 0.737 [0.007, 1.785]
2.22 0.700 [0.023, 1.821]
1.82 0.556 [0.010, 1.429]
0.53 0.417 [0.005, 1.151]
4.89 0.848 [0.049, 1.933]
0.23 0.337 [0.000, 1.096]
8.34 1.225 [0.030, 2.692]
0.97 0.533 [0.004, 1.443]
2.12 0.507 [0.041, 1.284]
11.03 1.197 [0.338, 2.321]
3.92 0.747 [0.069, 1.761]
1.21 0.543 [0.003, 1.453]
3.29 0.605 [0.035, 1.545]
1.25 0.495 [0.000, 1.393]
2.53 0.605 [0.014, 1.468]
3.15 0.762 [0.024, 1.826]
0.15 0.362 [0.004, 1.143]
1.34 0.715 [0.003, 2.028]
0.70 0.537 [0.002, 1.491]
0.10 0.299 [0.002, 1.003]
0.81 0.442 [0.004, 1.259]
0.27 0.425 [0.000, 1.315]
0.93 0.519 [0.004, 1.411]
1.30 0.510 [0.006, 1.379]
0.52 0.405 [0.000, 1.151]
4.96 0.732 [0.095, 1.667]
0.03 0.183 [0.001, 0.804]
7.17 0.923 [0.138, 2.012]
24.68 2.555 [0.848, 4.766]
0.17 0.332 [0.001, 1.123]
3.01 0.603 [0.053, 1.417]
10.17 1.143 [0.269, 2.297]
2.52 0.684 [0.003, 1.701]
4.84 0.734 [0.089, 1.635]
2.86 0.612 [0.016, 1.444]
BF category:
<3
[3, 10)
[10, 30)
[30, 100)
>=100

II-6c. B-Victoria



II-6d. B-Yamagata



Supplemental Figures

Supplemental Figure II-1a-d. Phylogenetic trees with discrete trait analysis for global datasets. The branch color indicates the inferred ancestor geographic region. All trees showed typical bushy leaves with clear seasonality patterns. II-1a. A/H1N1; II-1b. A/H3N2; II-1c. B-Victoria; II-1d. B-Yamagata.

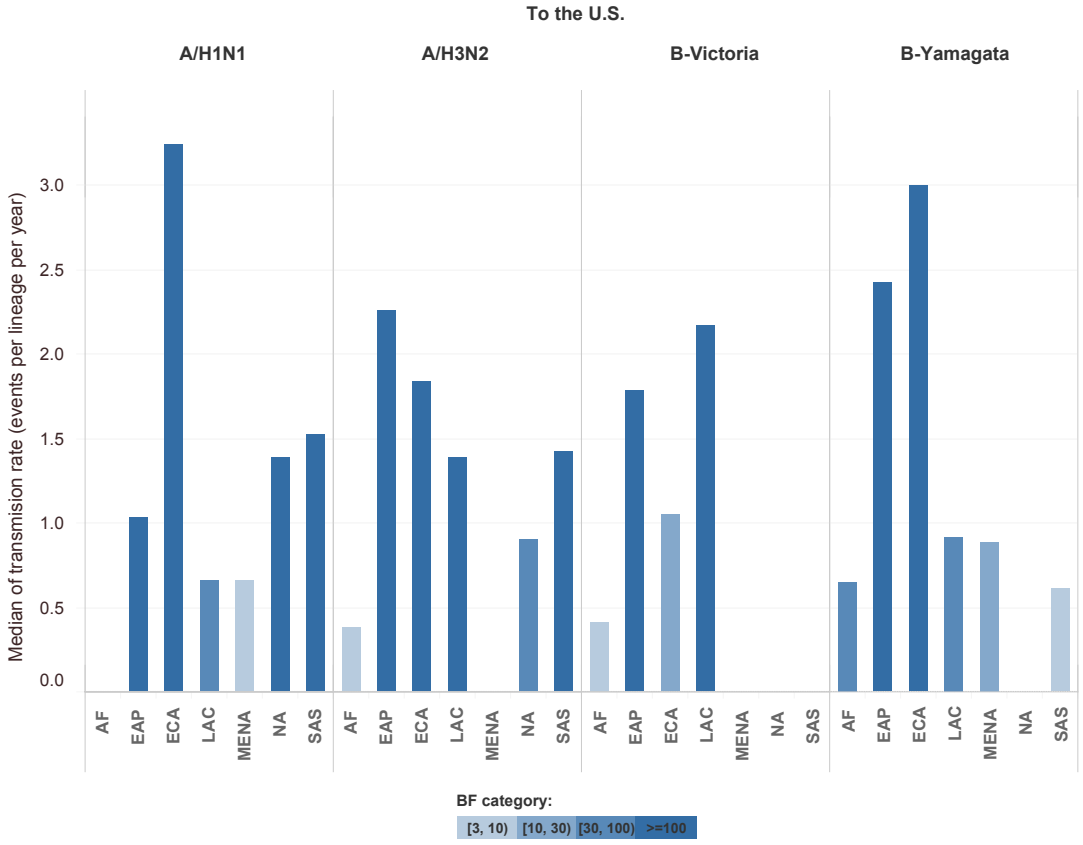
Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

Due to large size of the trees, please find the figures at its original size at Github link:

https://github.com/XuetingQiu/FluDiffusionUS.

Supplemental Figure II-2. Global viral sources into the U.S. for each viral type. Y-axis indicates the median values of transmission rates with unit as events per lineage per year. Only statistically supported viral sources (BF≥ 3) were shown. The bar color indicates BF levels and the legends are below each table, where $3 \leq BF < 10$ indicates substantial support, $10 \leq BF < 30$ indicates strong support, $30 \leq BF < 100$ indicates very strong support, and $BF \geq 100$ implies decisively statistical support. The region with highest transmission rate was ECA for A/H1N1 and B-Yamagata, EAP for A/H3N2, and LAC for B-Victoria.

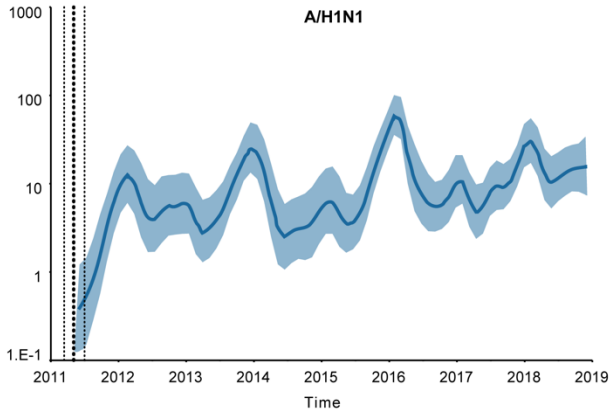
Abbreviations: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), North America (NA, excluding the U.S.), South Asia (SAS), Sub-Saharan Africa (AF).

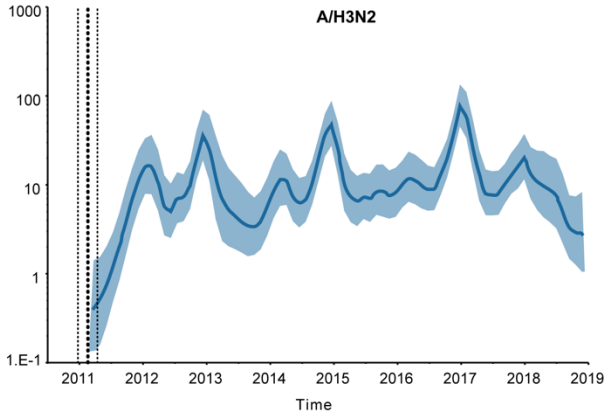

To the U.S.
A/H1N1
A/H3N2
B-Victoria
B-Yamagata
Median of transmision rate (events per lineage per year)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
AF
EAP
ECA
LAC
MENA
NA
SAS
BF category:
[3, 10)
[10, 30)
[30, 100)
>=100

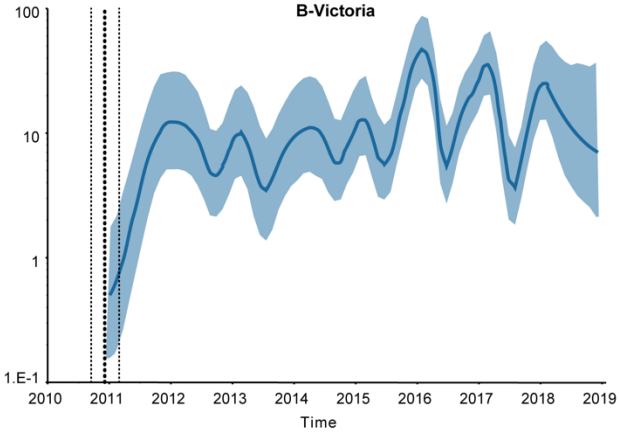
Supplemental Figure II-3a-d. Phylogenetic trees with discrete trait analysis for the U.S. region datasets. The branch color indicates the inferred ancestor geographic region. All trees showed typical bushy leaves with clear seasonality patterns. II-3a. A/H1N1; II-3b. A/H3N2; II-3c. B-Victoria; II-3d. B-Yamagata.

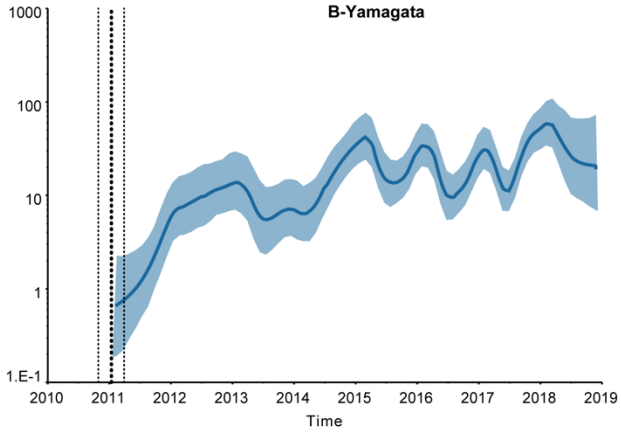
Due to large size of the trees, please find the figures at its original size at Github link: https://github.com/XuetingQiu/FluDiffusionUS.

Supplemental Figure II-4. Estimated effective population size by Skyride coalescent model for the U.S. region samples. The x axis represents time in years and the y axis represents viral effective population size on a log scale. The bolder dashed vertical line indicates the estimated most recent common ancestor time (tMRCA) for the viral population. The thinner dashed vertical lines are the 95% BCI for the tMRCA. The horizontal blue line and blue shadow indicate the median and 95% BCI bands of estimated effective population size, respectively.


A/H1N1
1000
100
10
1
1.E-1
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time

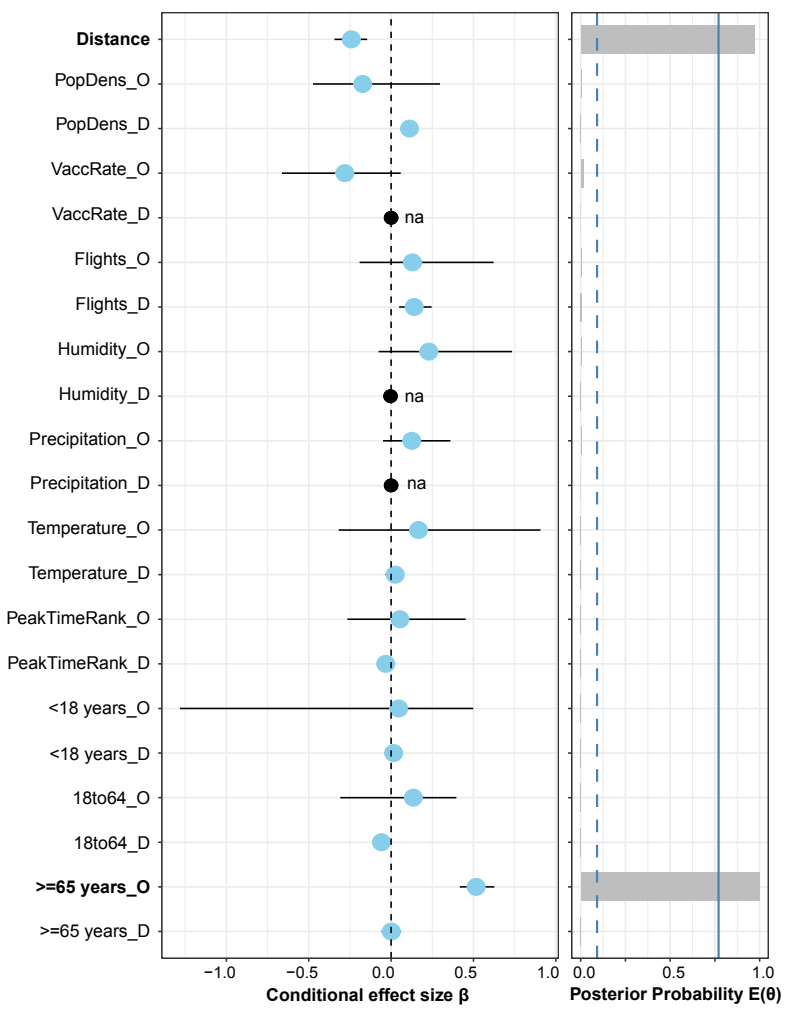

A/H3N2
1000
100
10
1
1.E-1
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time

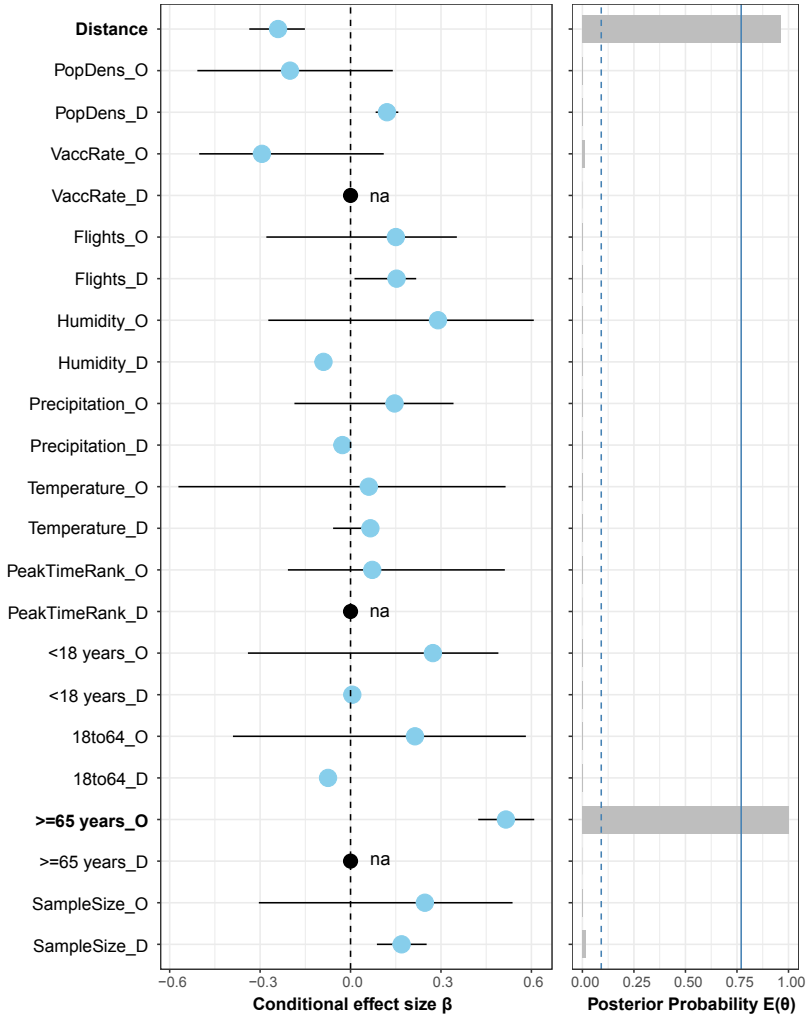

B-Victoria
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time


B-Yamagata
1000
100
10
1
1.E-1
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
Time

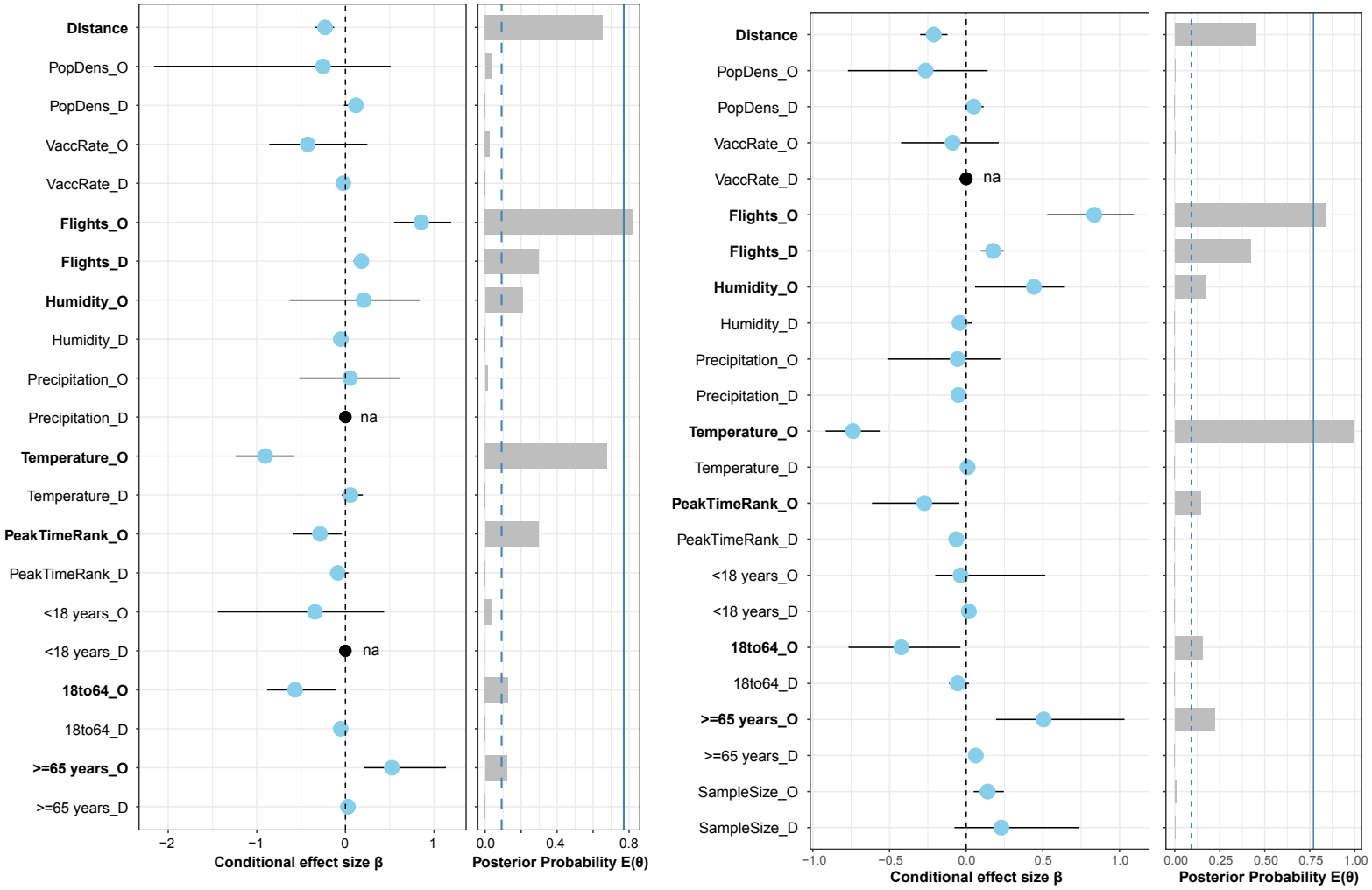
Supplemental Figure II-5. Generalized linear model reported significant predictors to affect diffusion dynamics for each influenza type in the U.S. The conditional effect size represented by the blue dot is the median of the coefficient when this predictor is included in the model. The bar on the blue dot indicates the 95% Bayesian Credible Interval (BCI). The posterior probability is the probability of this factor being included in the model, which is used to calculate Bayes Factor (BF). The dash vertical blue line indicates BF=3, where if the posterior probability is beyond the dashed line, it represents this indicator has statistical support to be included in the model. The solid vertical blue line represents BF=100, where if the posterior probability is beyond the solid line, then the indicator has decisively statistical support to be included in the model. Black dot with "na" indicates the conditional effect size is not available since the predictor is never included in the model during the simulation process. _O: geographic region as origin location; _D: geographic region as destination location. Abbreviation: PopDens – population density; VaccRate: average vaccine rate.
The left panel is the model without sample size of each region; the right panel is the model with sample size included. For all subtypes, the sample size did not affect the significance of these predictors.

II-5a. A/H1N1

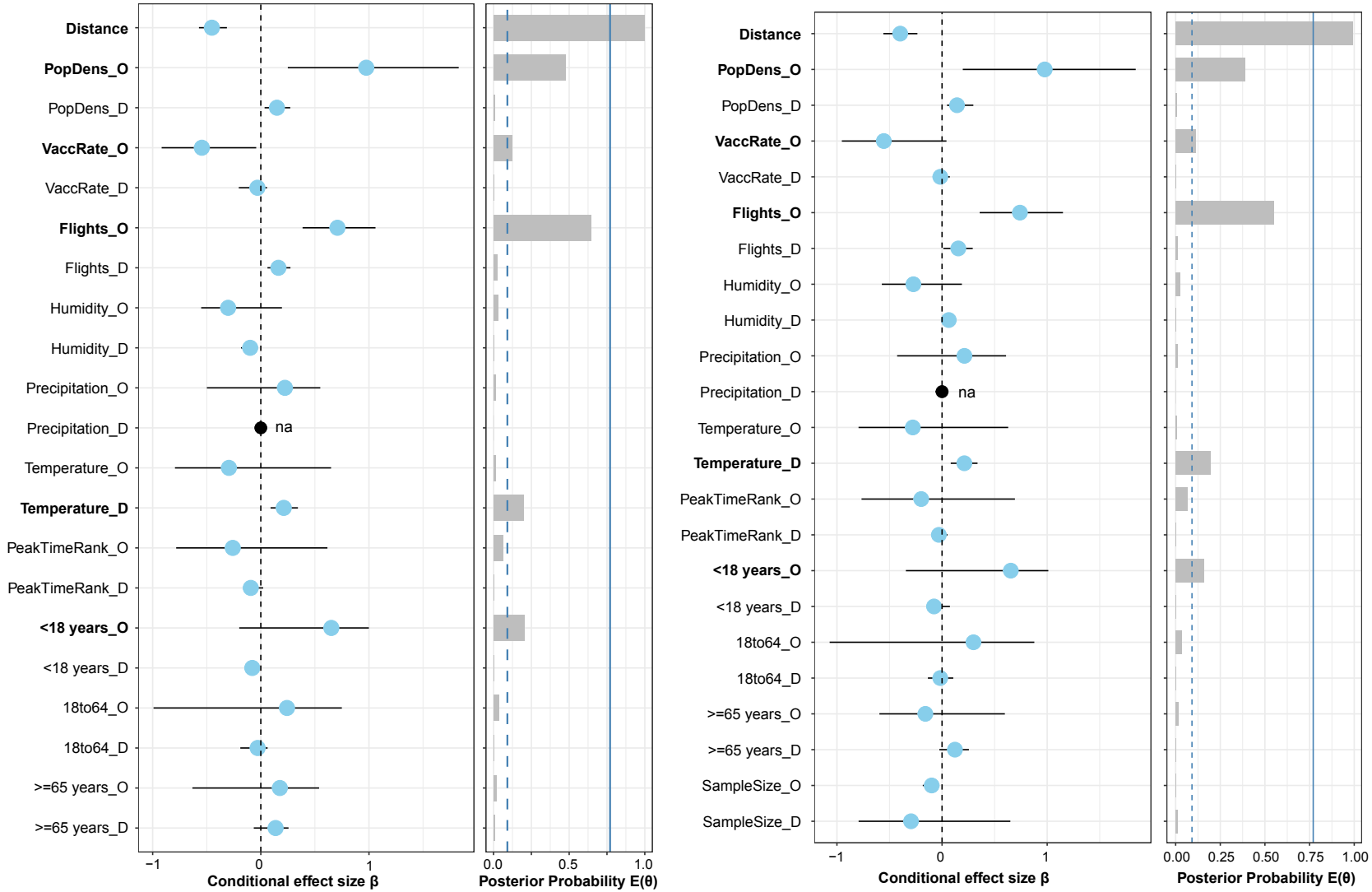

Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
na
−1.0
−0.5
0.0
0.5
1.0
Conditional effect size β
0.0
0.5
1.0
Posterior Probability E(θ)


Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
−0.6
−0.3
0.0
0.3
0.6
Conditional effect size β
0.00
0.25
0.50
0.75
1.00
Posterior Probability E(θ)

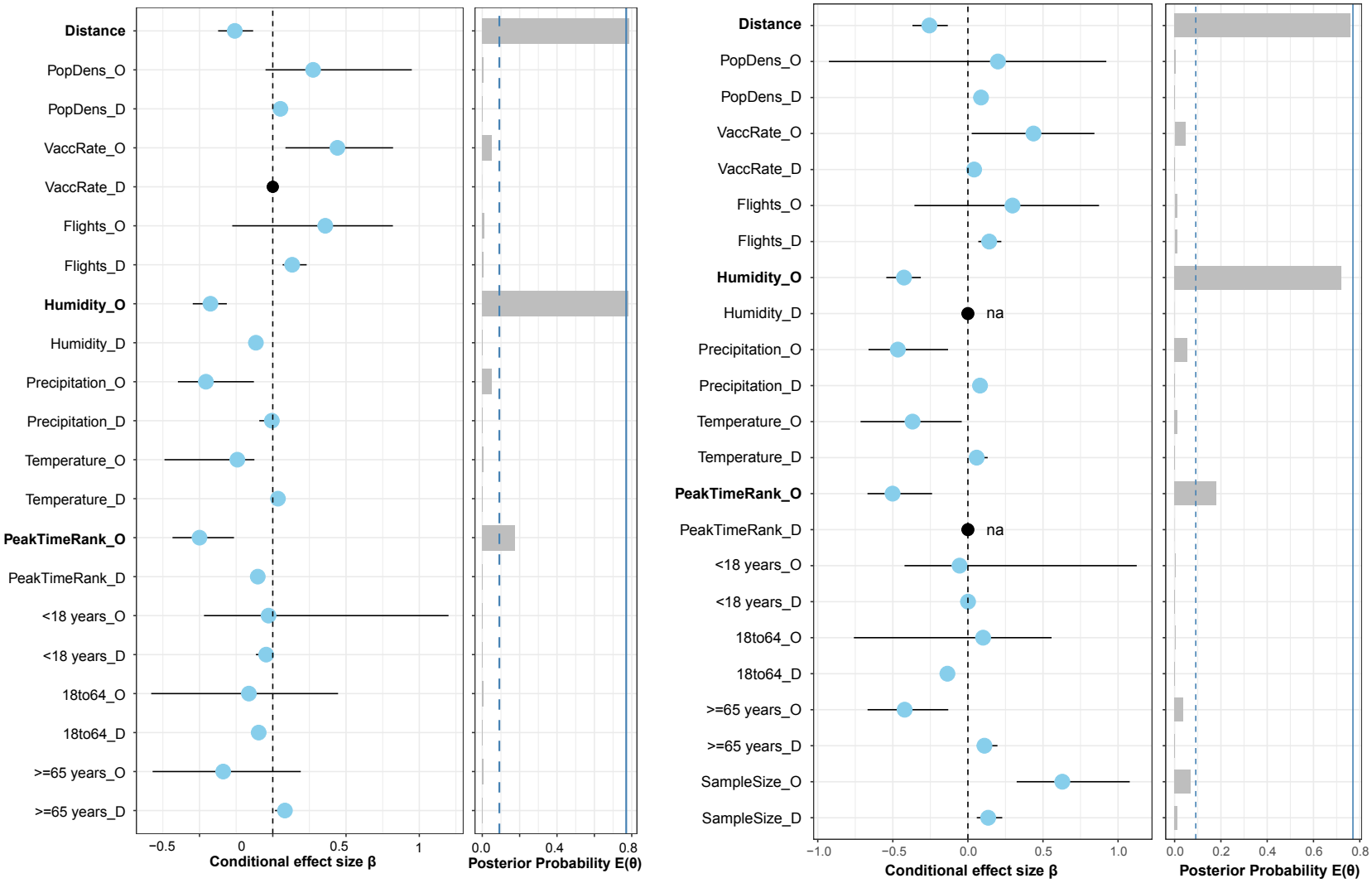
II-5b. A/H3N2


Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
na
na
−2
−1
0
1
0.0
0.4
0.8
Conditional effect size β
Posterior Probability E(θ)
Distance
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
Flights_O
Flights_D
Humidity_O
Humidity_D
Precipitation_O
Precipitation_D
Temperature_O
Temperature_D
PeakTimeRank_O
PeakTimeRank_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
SampleSize_O
SampleSize_D
na
−1.0
−0.5
0.0
0.5
1.0
0.00
0.25
0.50
0.75
1.00
Conditional effect size β
Posterior Probability E(θ)

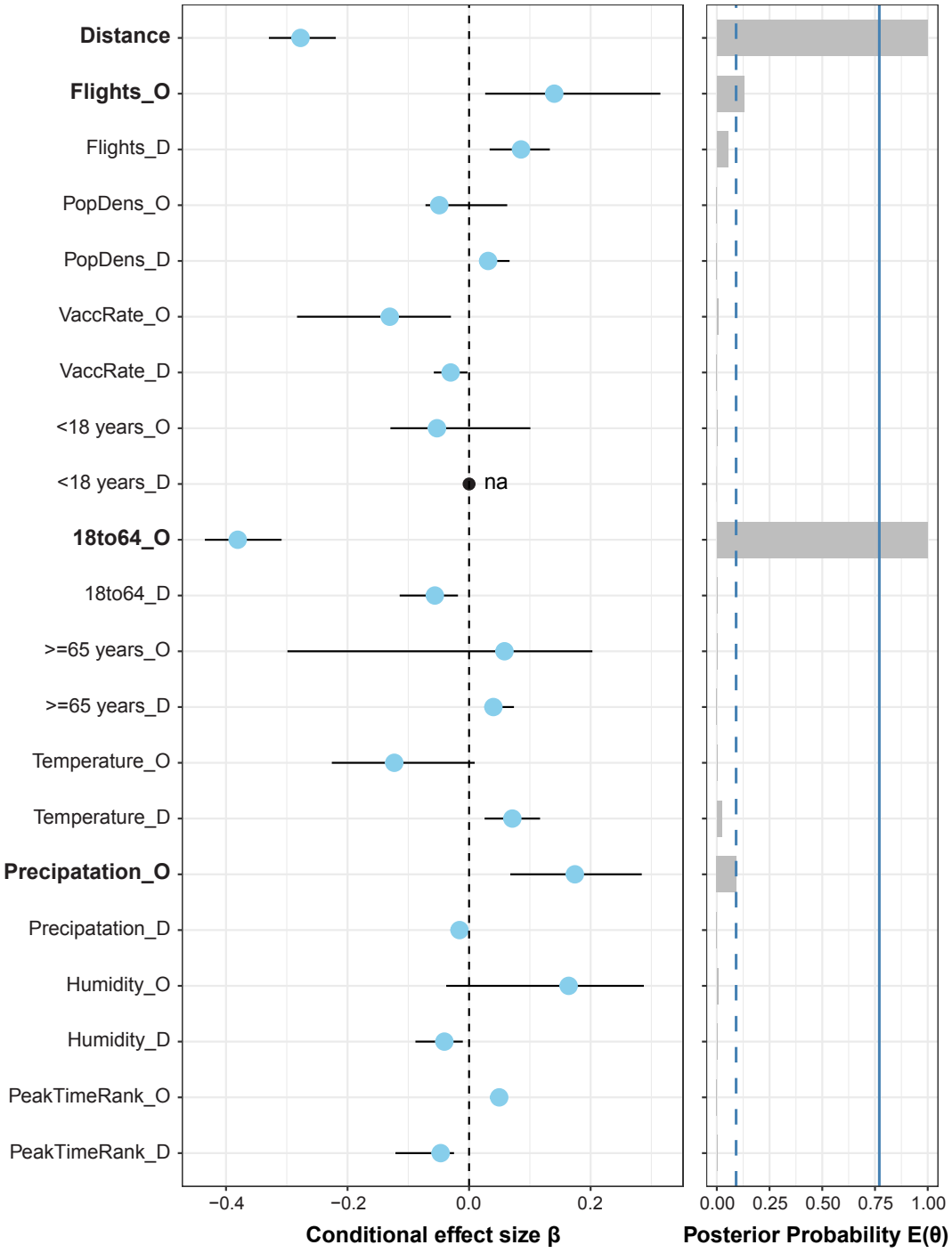
II-5c. B-Victoria



II-5d. B-Yamagata.



Supplemental Figure II-6. Joint generalized linear model (without Epidemic Size) reported significant predictors to affect diffusion dynamics in the U.S. The joint estimation unified four transmission rate matrices of each influenza type into one model to estimate the joint effect size of the predictor. The conditional effect size represented by the blue dot is the median of the coefficient when this predictor is included in the model. The bar on the blue dot indicates the 95% Bayesian Credible Interval (BCI). The posterior probability is the probability of this factor being included in the model, which is used to calculate Bayes Factor (BF). The dash vertical blue line indicates BF=3, where if the posterior probability is beyond the dashed line, it represents this indicator has a statistical support to be included in the model. The solid vertical blue line represents BF=100, where if the posterior probability is beyond the solid line, then the indicator has a decisively statistical support to be included in the model. Black dot with "na" indicates the conditional effect size is not available since the predictor is never included in the model during the simulation process. _O: geographic region as origin location; _D: geographic region as destination location. Abbreviation: PopDens – population density; VaccRate: average vaccine rate.


Distance
Flights_O
Flights_D
PopDens_O
PopDens_D
VaccRate_O
VaccRate_D
<18 years_O
<18 years_D
18to64_O
18to64_D
>=65 years_O
>=65 years_D
Temperature_O
Temperature_D
Precipatation_O
Precipatation_D
Humidity_O
Humidity_D
PeakTimeRank_O
PeakTimeRank_D
na
−0.4
−0.2
0.0
0.2
Conditional effect size β
0.00
0.25
0.50
0.75
1.00
Posterior Probability E(θ)

III. SUPPMENTAL MATERIAL FOR CHAPTER 5

EVALUATE THE IMPACTS OF H3N2 LAIV ON VIRAL DIVERSITY AND DIFFUSION DYNAMICS

Supplemental Table III-1. The estimated medians of the most common ancestor time (tMRCA) and substitution rate are very similar from discrete trait analysis and structured coalescent model. These results showed a slightly wider 95% BCI compared with the full dataset estimations which can be due to smaller sample size increasing uncertainty in the estimations. These results were from the subsampled dataset for global and the U.S. during 2003-2007 and Central Texas Trial samples during 2004-2006. The subsampling strategy is to only capture the global and U.S. isolates that are genetically closely related to Central Texas Trial samples.

	tMRCA		Substitution rate	
	Median	95% BCI	Median	95% BCI
Discrete Trait Analysis	2004.06	[2003.86, 2004.25]	3.72E-3	[3.09E-3, 4.47E-3]
Structured coalescent model	2003.98	[2003.63, 2004.27]	3.56E-03	[3.21E-3, 3.96E-3]

Supplemental Table III-2. Estimated introductions, transmission rate and their 95% BCI between different locations. These results were from the subsampled dataset for global and the U.S. during 2003-2007 and Central Texas Trial samples during 2004-2006. The subsampling strategy is to only capture the global and U.S. isolates that are genetically closely related to Central Texas Trial samples. The items in the cell in order are count of introductions, transmission rate and its 95% BCI. The cell color indicates BF levels and the legends are below each table, where BF < 3 means no statistical support, 3 ≤ BF <10 indicates substantial support, 10 ≤ BF < 30 indicates strong support, 30 ≤ BF <100 indicates very strong support, and BF ≥ 100 implies decisively statistical support.

Geographic region abbreviation: Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).

		Recipients		
		Global	WBC-uv	TB-v
Donors	Global		18.30 1.315 [0.2579, 2.983]	25.59 1.825 [0.3856, 3.969]
	WBC-uv	0.51 0.267 [0.0000, 1.153]		5.33 1.064 [0.0250, 2.756]
	TB-v	1.39 0.420 [0.0009, 1.575]	2.29 0.609 [0.0005, 2.078]	

BF category:

<3	[3, 10)	[10, 30)	>=100

Supplemental Figure III-1. Maximum clade credibility tree of H3N2 phylogeny with global regions, the U.S., and 2004-2006 Central Texas Trial samples including taxa information. This is the full dataset for global and the U.S. during 2003-2007 and Central Texas Trial samples during 2004-2006.

Geographic region abbreviation: East Asia and Pacific (EAP), Europe and Central Asia (ECA), Latin America and Caribbean (LAC), Middle East and North Africa (MENA), South Asia (SAS), Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).

Find the figure at its original size at https://github.com/XuetingQiu/LAIV_impacts.

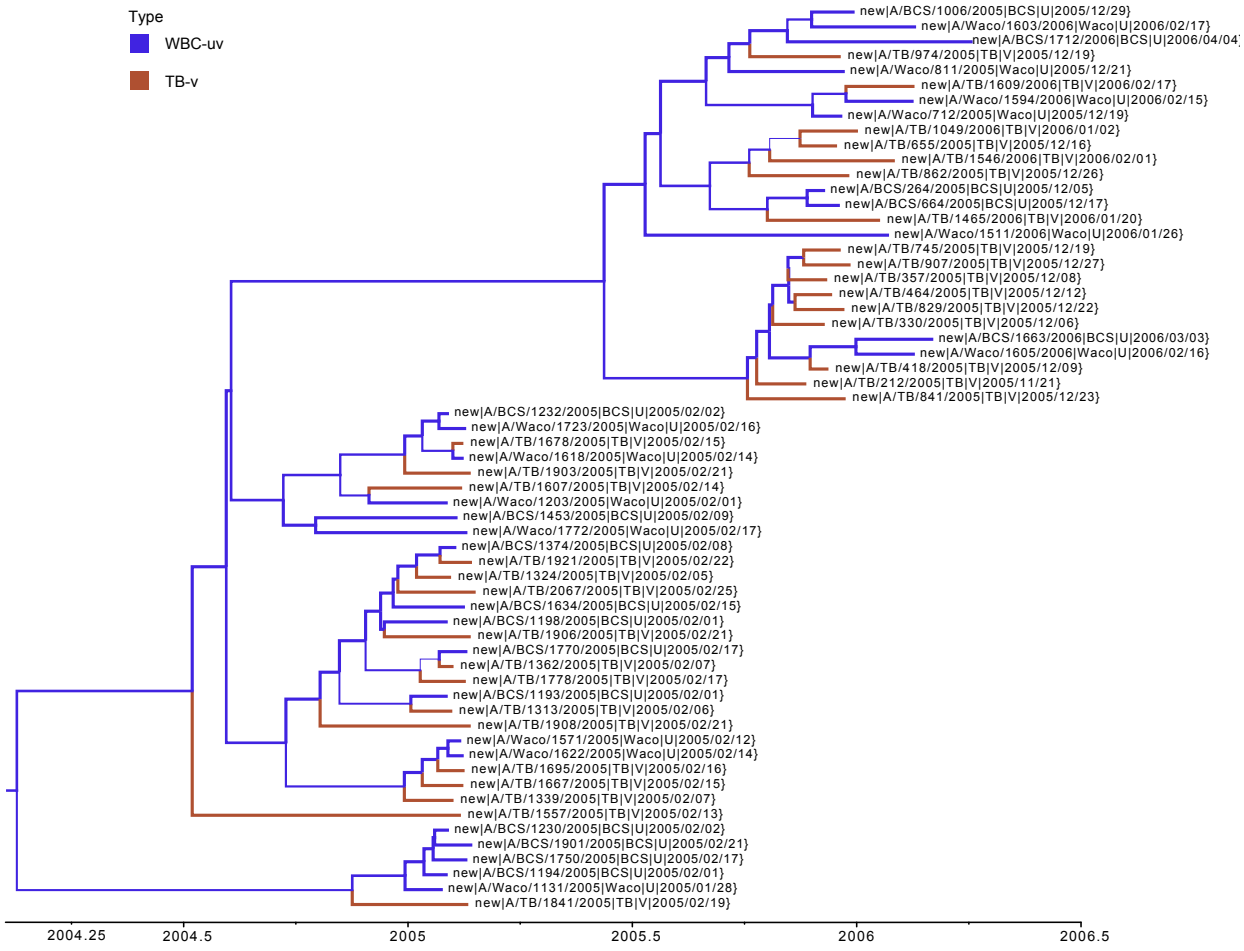
Supplemental Figure III-2. Maximum clade credibility tree from discrete trait analysis of H3N2 phylogeny with subsampled global regions, the U.S., and 2004-2006 Central Texas Trial samples including taxa information. This is the subsampled dataset for global and the U.S. during 2003-2007 and Central Texas Trial samples during 2004-2006. The subsampling strategy is to only capture the global and U.S. isolates that are genetically closely related to Central Texas Trial samples.

Geographic region abbreviations: Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).


Groups:
Global
WBC-uv
TB-v
2004
2005
2006
2007
2008

Supplemental Figure III-3. Maximum clade credibility tree from structured coalescent of H3N2 phylogeny for 2004-2006 Central Texas Trial samples including taxa information. The stroke weight of the branch is proportional to the posterior probability of the assigned type.

Geographic region abbreviation: Texas unvaccinated cities Waco and Bryan College Station (WBC-uv), Texas vaccinated cities Temple and Belton (TB-v).


Type
WBC-uv
TB-v
new|A/BCS/1006/2005|BCS|U|2005/12/29}
new|A/Waco/1603/2006|Waco|U|2006/02/17}
new|A/BCS/1712/2006|BCS|U|2006/04/04}
new|A/TB/974/2005|TB|V|2005/12/19}
new|A/Waco/811/2005|Waco|U|2005/12/21}
new|A/TB/1609/2006|TB|V|2006/02/17}
new|A/Waco/1594/2006|Waco|U|2006/02/15}
new|A/Waco/712/2005|Waco|U|2005/12/19}
new|A/TB/1049/2006|TB|V|2006/01/02}
new|A/TB/655/2005|TB|V|2005/12/16}
new|A/TB/1546/2006|TB|V|2006/02/01}
new|A/TB/862/2005|TB|V|2005/12/26}
new|A/BCS/264/2005|BCS|U|2005/12/05}
new|A/BCS/664/2005|BCS|U|2005/12/17}
new|A/TB/1465/2006|TB|V|2006/01/20}
new|A/Waco/1511/2006|Waco|U|2006/01/26}
new|A/TB/745/2005|TB|V|2005/12/19}
new|A/TB/907/2005|TB|V|2005/12/27}
new|A/TB/357/2005|TB|V|2005/12/08}
new|A/TB/464/2005|TB|V|2005/12/12}
new|A/TB/829/2005|TB|V|2005/12/22}
new|A/TB/330/2005|TB|V|2005/12/06}
new|A/BCS/1663/2006|BCS|U|2006/03/03}
new|A/Waco/1605/2006|Waco|U|2006/02/16}
new|A/TB/418/2005|TB|V|2005/12/09}
new|A/TB/212/2005|TB|V|2005/11/21}
new|A/TB/841/2005|TB|V|2005/12/23}
new|A/BCS/1232/2005|BCS|U|2005/02/02}
new|A/Waco/1723/2005|Waco|U|2005/02/16}
new|A/TB/1678/2005|TB|V|2005/02/15}
new|A/Waco/1618/2005|Waco|U|2005/02/14}
new|A/TB/1903/2005|TB|V|2005/02/21}
new|A/TB/1607/2005|TB|V|2005/02/14}
new|A/Waco/1203/2005|Waco|U|2005/02/01}
new|A/BCS/1453/2005|BCS|U|2005/02/09}
new|A/Waco/1772/2005|Waco|U|2005/02/17}
new|A/BCS/1374/2005|BCS|U|2005/02/08}
new|A/TB/1921/2005|TB|V|2005/02/22}
new|A/TB/1324/2005|TB|V|2005/02/05}
new|A/TB/2067/2005|TB|V|2005/02/25}
new|A/BCS/1634/2005|BCS|U|2005/02/15}
new|A/BCS/1198/2005|BCS|U|2005/02/01}
new|A/TB/1906/2005|TB|V|2005/02/21}
new|A/BCS/1770/2005|BCS|U|2005/02/17}
new|A/TB/1362/2005|TB|V|2005/02/07}
new|A/TB/1778/2005|TB|V|2005/02/17}
new|A/BCS/1193/2005|BCS|U|2005/02/01}
new|A/TB/1313/2005|TB|V|2005/02/06}
new|A/TB/1908/2005|TB|V|2005/02/21}
new|A/Waco/1571/2005|Waco|U|2005/02/12}
new|A/Waco/1622/2005|Waco|U|2005/02/14}
new|A/TB/1695/2005|TB|V|2005/02/16}
new|A/TB/1667/2005|TB|V|2005/02/15}
new|A/TB/1339/2005|TB|V|2005/02/07}
new|A/TB/1557/2005|TB|V|2005/02/13}
new|A/BCS/1230/2005|BCS|U|2005/02/02}
new|A/BCS/1901/2005|BCS|U|2005/02/21}
new|A/BCS/1750/2005|BCS|U|2005/02/17}
new|A/BCS/1194/2005|BCS|U|2005/02/01}
new|A/Waco/1131/2005|Waco|U|2005/01/28}
new|A/TB/1841/2005|TB|V|2005/02/19}
2004.25
2004.5
2005
2005.5
2006
2006.5

IV. REVIEW ON COMPUTATIONAL APPROACHES OF UNIVERSAL INFLUENZA VACCINE

Review

Computational Approaches and Challenges to Develop Universal Influenza Vaccines

Xueting Qiu[1], Venkata R. Duvvuri[1] and Justin Bahl[1,2,3,*]

[1] Center for Ecology of Infectious Diseases, Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA; xueting.qiu@uga.edu, venkata.duvvuri@uga.edu
[2] Department of Epidemiology and Biostatistics, College of Public Health, University of Georgia, Athens, GA
[3] Duke-NUS Graduate Medical School, Singapore
* Correspondence: justin.bahl@uga.edu; Tel.: +1-706-542-3473

Received: 10 March 2019; Accepted: 24 May 2019; Published: 28 May 2019

Abstract: Traditional design of effective vaccines for rapidly evolving pathogens, such as influenza A virus, have failed to provide broad spectrum and long-lasting protection. With low cost whole genome sequencing technology and powerful computing capability, novel computational approaches have demonstrated potential to facilitate design of a universal influenza vaccine. However, few studies have integrated computational optimization in the design and discovery of new vaccines. Understanding the potentials of computational vaccine design is necessary before these approaches can be implemented on a broad scale. This review has summarized some promising computational approaches under current development, including computationally optimized broadly reactive antigens with consensus sequences, phylogenetic model-based ancestral sequence reconstruction, and immunomics to compute conserved cross-reactive T-cell epitopes. Interactions between virus-host-environment determine the evolvability of the influenza population. We propose that with the development of novel technologies that allow integration of data sources such as protein structural modeling, host antibody repertoire analysis and advanced phylodynamic modeling, computational approaches will be crucial for the development of a long-lasting universal influenza vaccine. Taken together, computational approaches are the powerful and promising tools to develop a universal influenza vaccine with durable and broad protection.

Keywords: Universal influenza vaccine; computational design; interactions of virus-host-environment

---

In the history of fighting infectious diseases, vaccination is amongst the most cost-effective approaches available to prevent infection. Traditional approaches to vaccine design have been successful against many pathogens. But vaccines that target rapidly evolving and genetically diverse disease agents have frequently failed to generate long lasting protection for human populations. This is particularly true for influenza viruses, a single-stranded, negative sense RNA virus. One of the important weapons being developed to effectively prevent influenza virus infection is a vaccine that can provide durable and broadly-reactive protection against multiple subtypes, including those that may cause potential pandemics, that is, a universal influenza vaccine [1]. The National Institute of Allergy and Infectious Diseases (NIAID) has defined the criteria for universal influenza vaccine, which includes 1) being at least 75% effective against

symptomatic influenza infection; 2) protecting against group I and group II influenza A viruses (influenza B would be a secondary target); 3) having durable protection that lasts at least 1 year and preferably through multiple seasons [1]. These are challenging, but achievable goals to effectively develop a vaccine that can protect against the globally disseminated virus.

Recent vaccine developments have taken advantage of large-scale viral sequencing platforms, phylogenetic frameworks, protein structural modeling and systems biology to design novel broadly-reactive vaccine candidates, which have been used for influenza and other pathogens [2]. These new approaches have revealed insights of viral evolution, transmission dynamics and biological functions of proteins from mountains of genomic data and metadata [2]–[5]. Novel approaches for rational design in the genomic era can aid in achieving goals of universal influenza vaccine design. However, it has found limited applications in the design and discovery of new vaccines, an area where proper integration of computational support and design is essentially needed [2], [6].

In this review, we aim to briefly summarize the currently applied approaches of seasonal influenza vaccine design and their disadvantages (part 1), gather information on new or potential computational approaches and challenges (part 2), and to propose necessary resources and efforts needed for computational approaches of universal influenza vaccine candidates (part 3). We will explore the important role of computational vaccine design to improve the identification of pathogen-antigens and key components for designing and evaluating a universal vaccine design. Furthermore, we will discuss the potential of incorporating interactions of virus-host-environment to develop models that allow for precise prediction for viral evolution and vaccine candidates. This review provides a framework to integrate computational advances that could help in restructuring the existing seasonal influenza vaccine design and contribute to the development of universal influenza vaccine.

1. Current approach for influenza vaccine design

1.1 Selection of circulating influenza viruses for seasonal vaccine design

To prevent infections from circulating seasonal influenza viruses the annually administrated influenza virus vaccines contain H1N1 (phylogenetic group 1 hemagglutinin), H3N2 (phylogenetic group 2 hemagglutinin), and two influenza B virus components (Victoria-like and Yamagata-like) [7]. The vaccine candidates from natural influenza virus strains are recommended by the World Health Organization (WHO) based on the characterization and prediction of circulating strains likely to dominate in upcoming epidemic seasons. Twice a year, the expert panel from the WHO Collaborating Centers and essential laboratories and academies reviews the evidence of global surveillance, laboratory and clinical studies and evaluate the availability of vaccine strains to make recommendation on the components of influenza vaccine [8]. The evaluations are mainly based on viral antigenic and genetic characterization, which requires tremendous annual surveillance efforts and laboratory tests. After the selection of vaccine strains, it takes at least 6-8 months to produce sufficient global supplies of influenza vaccine via current vaccine production technologies with egg-based, cell-based or recombination-based vaccine [9][10]. For a comprehensive review of traditional approaches for influenza vaccine selection, design, development and challenges refer to this review paper by Wong and Webby [11].

Influenza vaccines selected from natural influenza virus strains predominantly elicit specific antibodies against the globular head domain of the surface protein hemagglutinin (HA) for each subtype or lineage, which is only effective to protect against closely-matched antigenic variants [7]. The HA, however, undergoes rapid antigenic drift that accumulates from point mutations under immune selection pressure in the major antigenic sites, allowing the virus to escape neutralizing antibody responses [12] and resulting in imprecise prediction of circulating strains. Though with large efforts of continuous surveillance and vaccine strain updates, vaccine mismatch has occurred many times [13]. In addition to

potential antigenic mismatch from selection procedure and delays in production, egg-adapted mutations accumulated during egg-based vaccine production can further exacerbate this issue, where the vaccine virus strain obtains relevant functional amino acid changes in the HA protein, resulting in low vaccine effectiveness [14]–[17]. Studies investigating the impact of vaccine mismatch have reported broad ranging vaccine efficacy (10% to 60%) for these annual vaccines, demonstrating severely low and unstable immune protection from influenza infection [18]. Predictive models of viral evolution to forecast dominant circulating influenza viral strains in the upcoming influenza seasons through the analysis of genetic and epidemiological data from influenza surveillance system have been developed to make quantitative predictions of viral evolution and aim to improve the selection of seasonal influenza vaccine candidates [10], [19]. This framework that relies on traditional vaccine design and surveillance has demonstrated potential to integrate multiple data sources to improve influenza vaccine design.

1.2 Universal influenza vaccine design

The seasonal vaccines offer a little or no protection to emerging zoonotic influenza viruses with pandemic potential, as many species, especially wild aquatic birds, are recognized as the natural reservoir of all subtypes of influenza A viruses and have the potential to occur spillover and infect humans directly [20]. As with past pandemics, the surface glycoproteins, HA and neuraminidase (NA) are replaced through reassortments of zoonotic strains where the human population has no pre-existing immune protection and the vaccines in use are not cross-reactive with these new strains [21]–[23]. Experimentally identified conserved and immunogenic M2 protein antigens [24], and HA-stalk design [24]–[26] have potential to elicit broadly protective antibodies against seasonal influenza strain. M2-based universal vaccine design focuses on the conserved antigens that have been experimentally identified on M2 protein. However, the low immunogenicity and epitope density by viral nature has been a fatal limit to make the cross-protection from M2 being effectively applied into vaccine design [24]. To solve this issue, many approaches have been developed to improve M2 immunogenicity, details of which can be found in this review by Zhang et al [24]. Similar with M2-based design, HA-stalk design tries to elicit the conserved and cross-reactive protection from the membrane-proximal stalk domain [25]. While the stalk domain is conserved across multiple influenza subtypes, it is shielded by the immune-dominant head domain. To amplify the broad protection from stalk domain, truncated HA without head domain, concentrated short peptides from stalk domain or recombinant chimeric HA proteins have been employed [24]–[26]. Despite the potential for both M2 and HA-stalk design vaccines to elicit broadly reactive immune response, a number of challenges remain (reviewed in [24] and [27]), including a limited understanding of the full repertoire of potential epitopes. More systematic computational approaches that go beyond circulating strain prediction and incorporate a full profile of antigens stimulating both humoral and cellular immune responses are needed for universal vaccine design [24]–[26]. To overcome these challenges, computational approaches have been employed to rationally and promisingly design vaccine candidates that can induce broadly (ideally universally) cross-protective and durable immunity for all seasonal and even emerging pre-pandemic strains [13], [28], [29].

2. Computational design of universal influenza vaccines

2.1. The rationale of computational design approaches

Traditional approaches have failed to produce stable and protective vaccines for hypervariable and rapidly-evolving viral pathogens, including influenza viruses [30], [31]. Reasons for failure include inherent uncertainty in pathogen evolution [32]. While global surveillance efforts and data sharing agreements have increased available information, vaccine design often ignores the underlying processes of the global influenza meta-population which generates diversity that allows the viral populations to

escape vaccine-induced immune responses and anti-viral treatments. Furthermore, hemagglutination inhibition assay, central to vaccine strain selection, is a poor approximation for the average immune response that does not account for the heterogeneity of immune responses between hosts and pathogens, which cannot provide a full profile of pathogen immunogenic features [33]. The failure to synthesize information across the host-pathogen-environment, including ecological and epidemiological determinants of disease persistence and spread (Figure 1) [2] has resulted in major information gaps that can be addressed by existing computational approaches and a concerted effort to develop a unified framework. Individual immune response to a vaccine is an interplay of genetic, molecular and ecological factors from both host and pathogen populations on large tempo-spatial scales [2]. As a consequence, traditional design inefficiently captures few pathogen features based on a limited input that does not account for the high diversity of pathogen and high heterogeneity of host's immune responses [2][34].

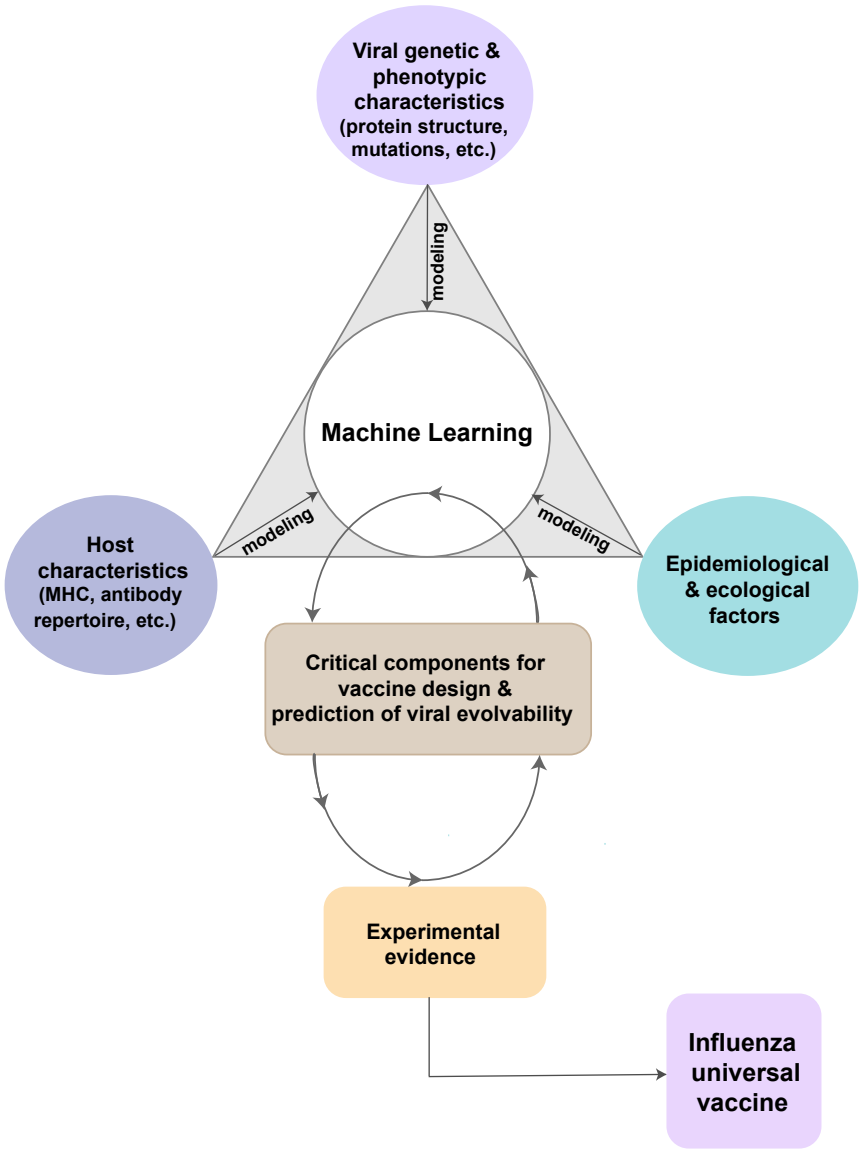


Figure 1. Framework of potential novel computational design. The summarized potential approaches combine the epidemiologic triad of infectious diseases. From host perspective, data on host characteristics are modeled to understand the susceptibility and immune response to influenza viruses by studying the host immunogenetics, for example, the antibody repertoire analysis or the human leukocyte antigen (HLA) structure analysis. From virus perspective, phylogenetic modeling is to understand the evolutionary history and patterns of viral genetic and phenotypic characteristics. Further, developing phylodynamic modeling, generalized linear model (GLM), and other more advanced models are critical to identify important epidemiological and ecological determinants that affect viral evolution and host immunity to influenza virus. In order to generate critical components for vaccine design and accurately predict viral evolvability, all three perspectives are combined to form a machine learning pipeline to incorporate information and learn from these data and models. Evidence from experimental tests on these components can be used into machine learning pipeline to improve outputs. Many iterations are needed with input of more information. The ultimate goal is to generate high-quality information and broad-reactive components for a universal vaccine of influenza viruses.

Fortunately, the growth of databases containing genome sequences sampled throughout global epidemics [35]–[37], increased computational power and theoretical algorithms allow complex data sources to be integrated into a unified framework allowing for a more complete understanding of pathogen and host features. Huge amount of data generated by the high-throughput technologies are currently available with more data regularly being made available. Computational approaches with advance data integration and quantitative empirical analyses fit the needs of universal vaccine design for highly diverse influenza virus in several promising aspects [38], [39]: 1) being able to model and analyze all available viral genomic data over a large tempo-spatial scale and shift from HA only design to cover more antigens on multiple viral proteins; 2) rapidly and cost-effectively screening antigens and epitopes in the early phase of vaccine candidate discovery; 3) capability of incorporating protein functional structure and antibody repertoire analysis via structural biology; 4) machine learning to incorporate viral, ecological, epidemiological and host immunological data to make precise assessment and prediction.

Computational approaches to identify candidates for universal influenza vaccine design have been used with a variety of novel vaccine production strategies in development. These approaches mainly focus on the 'unnatural immunity' [40] induced by more conserved or less immune-dominant domains in the surface proteins, internal proteins or both, to tackle with the high degree of variability in influenza viruses by boosting the immunity from the conserved or less evolvable proteins of the viruses. Current rational vaccine design uses comparative genomic methods to identify these conserved regions. These inferential methods include naïve approaches where conserved regions are identified from multiple sequence alignment comparison [27], [28], [41]–[43], phylogenetic approaches where common ancestry is estimated [44] and peptide engineering based on 3-D protein structure and immunomics. Figure 2 has summarized these current computational approaches. Table 1 has highlighted the advantages, disadvantages and examples of these approaches.

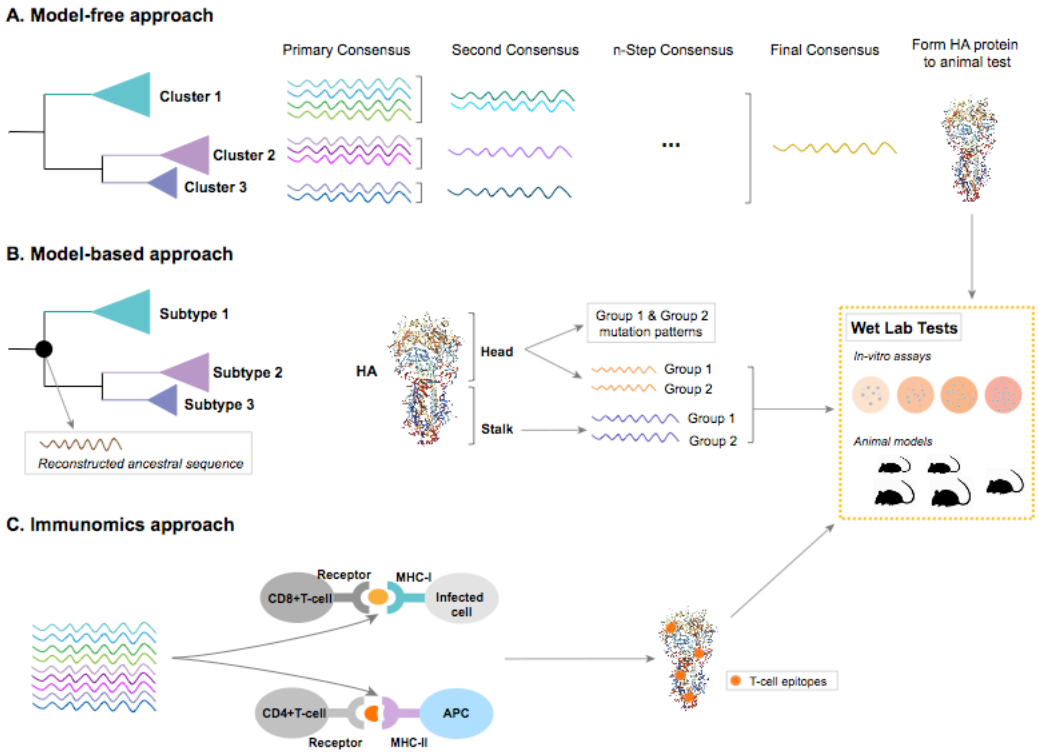


Figure 2. Framework for computational influenza universal vaccine design. ***A. Model-free consensus-based optimized approach.*** Consensus sequences from previously defined clusters are generated by aligning and comparing multiple sequences and selecting the most common residue at each position. It may go through several steps until the generation of a final consensus. ***B. Model-based ancestral sequence reconstruction approach.*** Maximum likelihood approaches and Bayesian framework are the most commonly used methods to reconstruct ancestral sequence at the ancestral node (shown as black dot on the tree)[45]. Statistical phylogenetic approaches (Maximum likelihood and Bayesian) allow for the reconstruction of possible ancestral node sequence [45]–[47]. Evolutionary models that incorporate protein structural domains can be used to separately estimate the evolutionary history on each functional partition as the HA head and stalk domains. Based on the evolutionary relationship among different subtypes of influenza A virus, common ancestral sequences of head and stalk domains can be generated within influenza A virus Group 1 (H1, H2, H5, H6, H7, H8, H9, H11, H12, H13, H16, H17, and H18) and within Group 2 (H3, H4, H7, H10, H14, and H15), respectively. ***C. Immunomics approach.*** The T-cell epitope prediction tools can be used to identify the potential CD4+ T-cell and CD8+ T cell epitopes from the pathogen proteome or protein(s) based on the high binding affinity between epitope-Major Histocompatibility Complex (MHC) complex. Some epitopes will be presented by the MHC-I on the surface of infected cells or by MHC-II on the surface of antigen presenting cells (APC) to the host CD8+ or CD4+ T-cells, respectively. These processes elicit the cellular and/or humoral immunity. The predicted T-cell epitopes that are evolutionarily conserved and common across or within subtypes will be constructed into peptides or proteins. All outputs from these three approaches, like epitopes, peptides, proteins or virus-like particles (VLPs), will be tested at in-vitro and/or in-vivo models to evaluate their immunogenicity. The proposed concept as shown is based on HA gene sequences, but these approaches should be used for all the gene segments of influenza viruses to generate a full profile of viral immunogenicity.

Table 1. Summary and examples of computational influenza universal vaccine design

Approach	Conceptual design	Evidence-level	Advantages	Disadvantages	Examples
Consensus-based optimized approach	Figure 2A	Pre-clinical	1) Efficiently generate a potentially full profile of conserved immunogenicity in viral genome; 2) Induce broad HA inhibition antibody titers that are cross-reactive with diverse strains within the same subtype; 3) Neutralize the receptor binding sites to prevent influenza disease with a clear path towards clinical proof of correlation for protective efficacy in humans	1) Biased viral samples may not generate consensus sequences that represent full profile of conserved immunogenicity; 2) Large efforts on surveillance data required	Pre-clinical tests on H1, H3 and H5 HA [28][41-43][102]
Ancestral sequence reconstruction	Figure 2B	Pre-clinical	1) Induce broad cross-reactive protection within highly diverse influenza subtype 2) Account for sampling bias and the variability of substitution rates among sites; 3) Potentially avoid the detrimental effects of antigenic drift with ancestral sequences; 4) Incorporate protein functional and structural domains	1) More sophisticated and advanced models to incorporate protein domains are still under development; 2) Experimental data on protein function is needed	Pre-clinical tests on ancestral sequence of H5N1 HA and NA [44]
Immunomics	Figure 2C	Pre-clinical & Clinical	1) Account for the heterogeneity of the major histocompatibility complex (MHC) in host; 2) Protections and viral clearance from T-cell response has been distinctively tested	1) Indirect estimation on epitope affinity to MHC; 2) To keep conformational epitopes to be function when designed into vaccine can be challenge	FP-01.1 Flu-v Multimeric-001 See Table 2 for details

2.2 The Host

2.2.1. Immunoinformatics to Immunomics

The field of immunoinformatics or computational immunology received major attention in 2000's from the research and governmental funding agencies [48], [49]. Immunoinformatics research mainly focuses on study and design of high-throughput *in-silico* approaches to explore the immune system at genome level (Figure 2) [50]. These technological developments coupled with pathogen genomes have tremendously contributed to the selection process of optimal vaccine antigens by lessening the time and cost involved in the conventional methodologies that involve pathogen cultivation and protein extractions. This methodology of analyzing pathogen genome to identify potential vaccine antigens is called "reverse vaccinology" [51], [52].

The study of immunomes coined as a new discipline "immunomics", where the 'immunome' is quoted as "the detailed map of immune reactions of a given host interacting with a foreign antigen" [49]. Immunomics tools such as B-cell epitope and T-cell epitope mapping methods mimic the diverse molecular pathways of adaptive immune system that accounts for humoral immunity (B-cells) and cellular immunity (CD4+ T-cells and CD8+ T-cells) to predict potential epitopes or immunomes from the pathogen proteomes [50], [53]. B-cell epitopes are surface exposed clusters of amino acids, which can be categorized as linear (a stretch of amino acids) and conformational (discontinuous) epitopes recognized by B-cell receptors (BCR) [54]. While T-cell epitopes are only linear, and T-cells receptors (TCR) can recognize epitopes when they are bound to the major histocompatibility complex (MHC) molecules. Two distinct subsets, CD4+ T-cells (helper T cells) and CD8+ T-cells (cytotoxic T cells) recognize epitopes when they bind with MHC class II and MHC class I, respectively [55]. MHC genes are highly polymorphic across different ethnicities that determines the fate of an epitope presentation to T-cells [55].

Immunomics can aid in identifying optimal B-cell and T-cell epitopes directly from the pathogen proteomes, while the literature suggested that T-cell predictions are more advanced and reliable than that of B-cell epitope predictions [56], [57]. A workshop on the B-cell epitope prediction tools reported that the

prediction performance of B-cell tools is still far from reality due to a lack of high-quality experimental datasets [56]. Detailed description on the existing epitope mapping tools, and challenges have been discussed in the cited review articles [54], [57]–[61]. Key limitations include: 1) the availability of experimental datasets, essential in training and developing any epitope prediction tool; 2) selection of epitope prediction tools may also introduce discrepancy in the identification of potential T-cell epitopes due to methodological differences. T-cell epitope prediction tools that include sequence- and structure-based methods are reviewed in Patronov et al [55] an Luo et al [62]); 3) the availability of high-quality datasets on the binding affinity of epitope-MHC, which directs the development and success of T-cell epitope prediction tools. A prediction of strong binding affinity suggests that a particular epitope will be presented to T-cells. But, this requires an experimental assessment; 4) The population coverage of an epitope is related to MHC polymorphism that exists in humans. The efficacy of epitope-based vaccine(s) can be limited due to variability of MHC alleles among different ethnicities. This may reduce the maximum population coverage of epitope-based vaccine leading to the failure of the vaccine to elicit T-cell immune responses. The current tools, IEDB population coverage [63] and EPISOPT [64], that are used to predict the population coverage are based on the limited experimental HLA frequency data from world-wide MHC allele frequency database (http://www.allelefrequencies.net/).

T-cell immunity plays a critical role in viral clearance thereby reduction in disease severity. Particularly, memory CD4+ T-cells can provide substantial protection against influenza infection through direct effector mechanisms as well as indirect regulatory and helper functions [65]–[67]. In the absence of neutralizing antibodies, the cross-reactive T-cell immune responses towards the well-conserved T-cell epitopes may play a significant role in promoting clearance of virus and reducing disease severity [68]–[70]. This phenomenon was well documented during the 2009 pandemic H1N1, as its unanticipated milder disease severity was largely attributed to the preexisting cross-reactive T-cell immune responses towards the evolutionarily conserved T-cell epitopes between seasonal H1N1 and 2009 H1N1 strains [71]–[77]. Taken together, these studies suggest that an epitope-based universal influenza vaccine can be developed by selecting the well-conserved and immunodominant epitopes across influenza subtypes using immunomics approach.

A major challenge in the design of epitope-based vaccines is to focus immune response onto multiple well-conserved epitopes in order to elicit broad protective/neutralizing immune responses. Epitope grafting or scaffolding, has been proposed as a solution for epitope-based vaccine design. In this method, minimal epitopes that are highly conserved in pathogen are grafted onto an appropriate heterologous-protein scaffold. Approaches for scaffold selection and design include single algorithm-based tools like MAMMOTH or meta-servers like TM-align and consensus-based designs [30]. Three main criteria have been proposed for the selection of scaffold that include size, where smaller-sized scaffolds help to focus immune responses to grafted epitopes while preventing unwanted responses to scaffold. Second criterion is the flexibility of scaffold with a possible positive correlation between flexibility and immunogenicity. The third criterion is the structural environment of the graft. A well-defined structural boundary between protein scaffold and epitopes enhances the specificity of immune responses [30].

2.2.2. Advanced universal influenza vaccines in clinical development

There are currently three promising epitope-based universal influenza vaccines, FP-01.1, Flu-v and Multimeric-001 (M-001) are at different stages of clinical trials (Table 2). Each vaccine is briefly described below.

Table 2. Promising epitope-based universal influenza vaccines at clinical trials.

Vaccine	Company	Projects	Clinical Phase			Clinical trial registration#	Reference
			I	II	III		
FP-01.1	Immune Targeting Systems Ltd., London, UK.	FP-01.1	completed	completed		NCT01265914, NCT01677676, NCT02071329	Francis 2015 [79]
		FP-01.1-Adjuvant	completed			NCT01677676	unpublished
		FP-01.1 + seasonal TIV + FP-01.1-Adjuvant	completed			NCT01701752	unpublished
Flu-v	PepTcell Limited	Flu-v	completed			NCT01226758, NCT01181336	Pleguezuelos 2015 [81]
		adjuvanted Flu-v		completed		NCT03180801, NCT02962908	van Doorn 2017 [82]
Multimeric-001 (M-001)	BiondVax Pharmaceuticals Ltd	M-001	completed	completed		NCT01146119, NCT01010737	Atsmon 2014 [85]
		M-001 (prime) + seasonal TIV vaccine (boost)	completed	completed		NCT03058692, NCT01419925, NCT02293317	Atsmon 2014 [85]
		M-001 (prime) + H5N1 vaccine (boost)	completed	completed		NCT02691130	unpublished
		M-001 as standalone vaccine			ongoing	NCT03450915	unpublished

FP-01.1 vaccine (also called as Flunisyn™), comprises six different synthetic peptides (length: 35 amino acids) each conjugated to the fluorocarbon moiety C8F17(CH2)2-COOH. These epitopes were derived from the nucleoprotein (NP), matrix protein (M), and polymerase basic proteins (PB1 and PB2) and have high level conservancy across H1-H9 influenza A subtypes with wider population coverage. The phase I clinical trial [78] results observed that vaccine has acceptable safety and tolerable profiles and generate robust CD4+ and CD8+ T-cellular immunity [79].

Flu-v vaccine contains multiple highly conserved T cell epitopes derived from NP, matrix proteins (M1 and M2) from influenza A and NP from influenza B viruses and are conserved across most influenza viruses with high population coverage [80], [81]. The phase II clinical trials with adjuvant+Flu-v triggered the T-cellular responses and also induced antibody response [82].

Multimeric-001 (M-001) is a universal influenza epitope-based vaccine is currently at the pivotal phase III clinical trial to assess the safety and clinical efficacy as a standalone universal flu vaccine in participants with age of older than 50 for a two-year follow-up [83]. M-001 comprised with a single recombinant protein that contains nine linear, conserved and common epitopes from NP, M1, and HA of influenza A and B viruses to activate both humoral and cellular immune system to provide multi-strain protection from the seasonal and pandemic influenza viruses [84]. The predicted population coverage of these selected epitopes is greater than 90%. The epitopes from the HA1 region which is hypervariable were not included in the M-001. At phase II clinical trial in 120 participants aged 65 years and older, M-001 was first administered to the study participants and three weeks later they were immunized with 2011-2012 seasonal trivalent inactivated vaccine. Results reported that M-001 alone elicited cellular responses and enhanced HA inhibition (HAI) seroconversion to 2011/12 vaccine strains and even to certain former vaccine strains [85].

The positive note on the epitope-based universal vaccine efficacy in eliciting the robust immune responses at clinical trials underpins the immunomics in advancing the current vaccine development approaches to prevent infections from remerging or emerging highly evolving influenza viruses.

2.2.3 Computational approaches that incorporate host immunological factors

Immunomics approach indirectly combines host information of MHC structure by computing the epitopes with potentially high affinity to MHC. There is another approach called antibody repertoire

analysis to directly incorporate host immune response for vaccine design. It is to analyze all the antibody affinity and specificities that can be produced by an individual, which can be a valuable tool for quantitative evaluation of vaccine-induced immune responses [30]. Though it is currently used to characterize broadly neutralizing antibody (bnAb) lineages, with the development of next-generation sequencing (NGS) technologies and systems biology, the analysis of antibody repertoire encoded by B cells in the blood or lymphoid organs can be used to understand humoral immune responses and to identify antibodies specific for antigens of interest in animal models and human vaccine trials [30], [86]–[88]. The antibody NGS can have impact on the rational vaccine design by decoding the human immune responses and delineating B and T cell antigen receptors [89], [90]. This approach has been well developed in HIV-1 to identify hypervariants and evolution on neutralization and binding to bnAbs [91], [92] and explore the antibody lineage via phylogenetic modeling [88], [93]. These technologies and bioinformatics tools can be applied to influenza virus vaccine design with creating library of antibody repertoire by NGS. The library then can be used in computational approaches to quantitatively measure the immune responses and further to predict the effects of vaccine candidates without completely relying on costly animal tests.

The main limitation with this approach is that linear sequence may not accurately predict the conformational variations when these antigens are put back in a complete protein context [94]. When the conformational structure of the epitope is not accurate, the corresponding immune response cannot be precisely computed [95]. To solve this issue, some high-performance bioinformatics tools such as molecular dynamics simulations can be used to predict the 3-D structure and stability of proteins or peptides [96], [97]. Furthermore, in the previous section, the successful maintenance of the conformational epitope in these clinical tested vaccines has provided positive evidence for epitope-based universal vaccine design. Taken together, with this antibody repertoire analysis tool, the computational estimation of immune stimulation of these predicted viral antigens in hosts can be more accurate.

2.3. The Pathogen

2.3.1. Model-free Consensus-based optimized approach

Consensus sequences are usually generated by aligning and comparing multiple sequences and selecting the most common residue at each position (Figure 2A). These sequences are expected to effectively capture a profile of conserved genetic and epitope information which can induce cross-reactive cellular immune responses [98]. The outcome of this approach is a sequence alignment with conserved antigens that can be expressed on virus-like particle (VLP), which are similar to intact virions but not pathogenic [41], [99]. Influenza VLP vaccines have advantages that a live virus is not used at any step during vaccine production [100] and they can maintain conformational epitopes by presenting surface antigens in their original structures. Consensus-based studies [29], [98], [100] have generated consensus sequences for NA protein of H1N1 and several influenza proteins of H5N1, including HA, NA and matrix protein M1, which have elicited broadly-reactive immune response. However, the nature of consensus-based antigen design determines that it is highly influenced by the input sequences and thus subject to sampling bias [101]. For example, H5N1 isolates were sampled in different geographical locations and from different hosts, including human and avian. If samples from one location or one host are overrepresented in the sequences used to generate consensus, then it can bias the output consensus sequence, which may not accurately represent the full conserved genetic profile of the whole H5N1 population. To overcome issues from sampling bias, an iterative optimization strategy has been implemented in an approach known as computationally optimized broadly reactive antigens (COBRA) [41].

The critical step in designing COBRAs is to use multiple rounds of consensus generation. Within each phylogenetic subclade of the influenza virus subtype, the primary consensus with the most common amino acid at each position is generated for each individual outbreak group that is defined based on geographic location and collection time. The secondary consensus is generated from the primary consensus to represent the subclade. The third or fourth consensus is generated based on previous round of consensus, until the final consensus is generated and termed COBRAs [41].

The COBRAs generated by multiple rounds of consensus generation are representative of the diversity in the viral population and are able to induce neutralizing antibodies or other immune boosting response to protect against past, current and ideally, future circulating strains of this specific HA subtype [27]. COBRAs-based designed HA protein of H1, H3 and H5 have been tested with *in-vitro* assays and animal models. This preclinical evidence has showed broad HA inhibition antibody titers that were cross-reactive with different strains within the same subtype [28], [41]–[43], [102]. This approach has advantages over other universal vaccine candidates, because COBRA HA-elicited antibodies are able to neutralize the receptor binding site and the design has a clear path towards clinical proof of correlate for protective efficacy in humans [28]. However, there are some major concerns with this approach. To be universally cross-reactive, the ideal CORBA HA protein is to cover all the conserved information within one subtype or multiple subtypes. The conserved immunogenic profile of consensus sequences from COBRA approach is dependent upon the sharing of epidemiological and genetic data collected during public health investigations and surveillance of outbreaks. With biased viral samples, the consensus sequence generated may not represent the full profile of conserved immunogenicity along viral evolving history. Even with increased global efforts to collect data and characterize epidemics it is unlikely that sufficient data could be collected to overcome this challenge. Alternative approaches, such as phylogenetic modeling of viral proteins along a characterized evolutionary trajectory that account for impacts of sample biases and missing data could greatly improve design of COBRA candidates.

2.3.2. Phylogenetic model-based approaches to ancestral sequence reconstruction

Another way to identify potential broadly reactive antigens is ancestral sequence reconstruction, which is to computationally infer ancestral gene sequences and the translated ancestral protein sequence (Figure 2B) [103]. Ancestral sequences can reveal conserved functions of the pathogen protein and evolutionarily favorable traits [104]. These conserved functions may indicate potential immune targets. Phylogenetic evolutionary models have been used to infer influenza viral evolutionary history for decades with molecular data, including the analysis of large phylogenetic trees, complex evolutionary models for more accurate ancestral inference and detection of the imprints of selection pressure in molecular sequences [105], [106]. Phylogenetic algorithms have been developed to reconstruct ancestral sequences for broadly-reactive vaccine design [44][107]. This phylogenetic approach with marginal reconstruction yields the maximum likelihood at the site with a specific amino acid after comparing all probabilities of different amino acids at a site on an internal node [107]. It can more accurately account for sampling bias and the variability of substitution rates among sites that can affect consensus approaches described above.

In detail, Ducatez and colleagues [44] developed an ancestral sequence reconstruction method for highly pathogenic avian influenza (HPAI) H5N1 surface proteins HA and NA. Based on a maximum likelihood tree, several ancestral sequences were reconstructed at the internal nodes of co-circulating HPAI H5N1 viral lineages to capture the conserved genetic characteristics of these viruses. These ancestral sequences were synthesized into attenuated influenza viruses that could replicate. Their cross-reactive protection against H5N1 morbidity and mortality have been confirmed in preclinical experiments with ferret models. These findings provide strong evidence that computationally derived

vaccine candidate sequences and these technologies should be used to explore and enhance the cross-reactivity, which can be easily fit into the current licensed vaccine platform. These computationally derived ancestral sequences as vaccine candidates may help in avoiding the detrimental effects of antigenic drift on the vaccine effectiveness. But this approach can be weakened by phylogenetic uncertainty in particular when trees possess long branches due to insufficient information [108].

The functional and structural domains of pathogen protein can be under disparate immunologic pressures and thus have impacts on the evolutionary phylogeny [109] and the accuracy of ancestral sequence reconstruction. Even though advanced models, including those that account for protein sequence and structure [110], [111] have not been applied for vaccine design, the computational approach is promising. Precise estimation of influenza virus evolution including protein structural and its functional information supported by experimental data [112], may help to efficiently identify and select target antigens for universal vaccine design [30].

The integration of protein functions and structures into evolutionary models has two main challenges: 1) published viral protein structural and functional information may not be available or sufficiently resolved based on current studies; 2) The assumption of nucleotide site independence in the model cannot capture the biological reality that some sites are linked due to shared function [113]. Some modeling approaches with protein structure scoring system or partitioning schemes on the protein sequence [97], [110], [114] can potentially overcome these challenge, for example, protein structure has been explored with coarse-grained models for structure prediction, prediction of protein interaction and molecular dynamics simulations of protein folding [97]. This provides the statistical potential like a scoring system for sequence-structure compatibility, which can be used to evaluate the probability of fixation of a given mutation and improve the precision of ancestral reconstruction [111]. However, few studies have incorporated protein structural information into the evolutionary analyses. Simple representations of protein functional and structural domains have been used so far. Hypothetically, novel models with a more complete representation with a full site mapping of the protein functions and structures would yield a better fit. But in a phylogenetic context, structurally informed models are still outperformed by some site-independent models in terms of fit [111]. Preliminary data suggest that this would become less of a concern with increased sharing of sequence data [110].

High-throughput experiments that quantify the effects of all single substitutions on gene function so that evolutionary model can adequately capture the heterogeneity of selection at different sites, which may improve phylogenetic inference and ancestral sequence reconstruction [112], [115]. The new experimental technique is called deep mutational scanning, where a gene is randomly mutagenized and subjected to functional selection in the laboratory, and then deep sequenced to quantify the relative frequencies of mutations before and after selection [116], [117]. This technique has been used to quantify the impacts of codon changes to several proteins or functional domains [115], [116], [118]–[121]. This information of protein function from rapid high-throughput experiments may greatly improve the precision of ancestral sequence reconstruction [122].

2.4. The Environment

2.4.1 Pathogen Evolvability

Uncovering the important ecological, immunological and environmental determinants on viral evolution is very important to make predictions of the viral emergence, fitness, transmissions and circulating potential after new substitution is introduced [123]. Evolvability, first coined by Kirschner and

Gerhart in 1998, means that the organism's capacity to generate heritable phenotype [124]. The zoonotic nature and complicated ecology of influenza viruses make evolvability more difficult to quantify and predict. But with the advances of phylogenetic algorithms, models can integrate and evaluate the impacts of environmental determinants. For example, an important development in phylogenetic modeling was the field of viral phylodynamics that was introduced in 2004 to study "how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies" [125], [126]. Dynamics of influenza virus infections and transmissions at individual-level (such as viral evolution within an infected host), population-level (individual hosts within a population), or ecology-level (entire populations of different host species) have been studied [125]. Specially, phylodynamics have been used to study factors of interest on some viral phenotypes, including virulence, viral transmissibility, cell or tissue tropism, and antigenic phenotypes that can facilitate immune escape, etc. [108], [125]. Details of methods and examined significant factors can be found in these reviews [108], [125], [126].

Furthermore, the complements between phylodynamic modeling and experimental testing can be integrated together to improve prediction on influenza virus evolvability. For example, experimental studies designed to assess viral evolvability [127], [128] demonstrated that a measured fitness score or estimated tolerance for mutations can be used in phylodynamic modeling to link phenotypes, genetic characteristics and other ecological factors, which can improve the prediction of viral evolvability for natural influenza virus strains [128]. The potential predictors and consequent mutations computed by models can enhance our understanding on viral characteristics, potential immune escape, or influenza antiviral drug resistance [129]. Challenges for this area are how to get accurate and sufficient information on the epidemiological, immunological and ecological factors, how to expand, integrate and enhance phylodynamic models [108], and how to gather the current modeling factors to improve prediction of viral evolvability [123].

3. Resources and efforts needed for computational vaccine design

Computational models with incorporating host-pathogen-environment can efficiently facilitate the understanding of viral evolution and the selection on critical information for vaccine design. With the challenges summarized above, extra resources and efforts are needed for developing computational vaccine design.

3.1. Data collection and sampling efforts

Computational vaccine design highly relies on the input data quality [30]. To be specific, the representative of the collected samples, the completion and precision of recorded data, and the timely manner of data sharing and availability can ameliorate the output from computational modeling [19], [130], [131]. Compared to other infectious disease sampling, influenza viruses have already established an excellent global network of sentinel institutions to monitor outbreaks and collect human samples [132]. With the lower cost of full-genome sequencing, a large amount of genetic data has been available for influenza research. However, three main limitations exist in current surveillance: 1) the imbalanced sampling efforts on different hosts and geographical regions; 2) the incompletion of data records [130]; and 3) the delayed availability of sequence data [19], [131].

The unequal sampling of geographical regions is caused by global and local resource allocation [133]. Policies to globally optimize resource allocation with considering the representative of collected

samples from outbreaks in different regions are needed. But majorly, the unequal sampling in zoonotic hosts is more severe. Human influenza outbreaks have been well monitored and sampled [130]. However, to better understand viral evolvability and predict potential pandemic emerging from zoonotic strains, more sampling efforts are definitely necessary in animal hosts, especially wild aquatic birds [130]. Olson *et al*, examined 11,870 GenBank records and reviewed 50 non-overlapping studies and over 250,000 birds to access the status of historic sampling efforts during 1977 – 2012, where they found that sampling in different hosts, location and viral subtypes are severely imbalanced and there are a high proportion of non-tested samples globally. If we aim to identify a high proportion of the virus subtypes in circulation in a given time period with limited resources, a sample-based accumulation curve can provide an initial rationalization and optimal sample size for AIV surveillance [130], [134], [135].

The affiliated sequence meta-data records have been improved with samples from recent years. But the epidemiological information, viral phenotypic characteristics and host characteristics are not sufficiently recorded. With no accurate information on geographical region, host species and migratory pathways and viral characteristics, we do lose lots of power in our model inference [136], not to say improving the prediction of viral evolvability. GISAID [35] and GenBank [36], these open access database platforms have facilitated the accessibility and sharing of influenza sequence data to the science community. Despite the availability of these platforms, the sharing of viral sequence data is often long after the outbreaks and records are frequently incomplete [137]. Therefore, a standardized protocol on how to record collected samples and what information is needed to report should be established for sharing more complete viral and host-related information.

3.2 Integration of experimental evidence and model development

As shown in Figure 2, these computational models could efficiently compute and select critical components for vaccine design. However, we cannot solely rely on computational design, where computed antigens have uncertain biological effects. Experimental evidence (Figure 1) from animal models or approved human clinical trials are valuable to be incorporated into computational design. The experimental data on pathogen immunogenicity and host immune system can first provide preliminary evidence on natural or computed antigens and further amplify the usage of this new evidence to the computational procedure for more accurate prediction and evaluation [19].

More complicated and realistic models previously limited by computing capability can be developed with the advances of computing power [138]. For example, it becomes possible to develop viral phylodynamic models that can incorporate results from laboratory experiments of viral antigens and host immune responses [108]; The development on structured coalescent for better estimation on viral population and mutation or migration events [139], [140]. Furthermore, to avoid overparameterization, model selection procedure should be applied during the process of novel model development to optimize the balance of biological reality and parameterization [138], [141]. With all these, the next step would be to introduce and apply machine learning to the computational process for vaccine design.

Machine learning is a subset of artificial intelligence in the field of computer science, which usually uses statistical techniques and mathematical models to make computers “learn” with data without being explicitly programmed, that is, performance on a specific task progressively improves [142]. Machine learning algorithms discover patterns in data and construct mathematical models using these prior discoveries. One advantage of machine learning is that the models can be used to make predictions on future data by cumulating from previous evidence and improving on forecasting algorithms [143]–[145]. Though still in an early phase of implementation, the concept of machine learning has been used in viral

evolutionary modeling and has been a rapid way to gather and update information based on known information [146], [147]. Machine learning can incorporate different modeling steps and all available surveillance, genetic and experimental data to keep updating information and make predictions for computational vaccine design (Figure 2). For example, the model of conserved epitope prediction mentioned in previous sections can also be incorporated in the platform with host and environmental factors to make prediction on currently circulating viruses, where broad-reactive vaccine candidates can be rapidly computed and tested.

4. Conclusions

For decades, we have been using the traditional approaches to design and develop influenza vaccines. The rapid genetic changes and antigenic drift of influenza virus populations results in short term protection necessitating continual vaccine updates with novel viral components based on analysis of globally circulating variants. Furthermore, these vaccines do not offer any immunological protection against potential pandemic zoonotic strains, one of the lessons learned through the unprecedent appearance of swine-origin 2009 pandemic H1N1 virus.

Recent decades have witnessed the technological advancements in the viral genetic sequencing and computational modeling in tracing the complexities involved in the interactions of host- pathogen-environment that produced important insights into influenza disease dynamics across biological scales. Integrating these computational and technological pipelines into the vaccine design protocols can facilitate the development of a broadly cross-reactive, evolutionarily resistant universal influenza vaccine.

Author Contributions: conceptualization, X.Q. and J.B.; methodology, X.Q., V.D. and J.B.; validation, X.Q., and V.D.; formal analysis, X.Q. and V.D.; resources, J.B.; writing—original draft preparation, X.Q.; writing—review and editing, X.Q., V.D. and J.B.; visualization, X.Q.; supervision, J.B.; project administration, J.B.; funding acquisition, J.B.

Funding: This research was funded by National Institutes of Health (NIH) Centers of Excellence for Influenza Research and Surveillance (CEIRS, contract #HHSN272201400006C).

Acknowledgments: We thank Lambodhar Damodaran and Jiani Chen for reading through the manuscript and giving valuable comments. We also acknowledge the three anonymous reviewers' valuable comments to improve this review.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] E. J. Erbelding *et al.*, "A Universal Influenza Vaccine: The Strategic Plan for the National Institute of Allergy and Infectious Diseases," *J. Infect. Dis.*, vol. 218, no. 3, pp. 347–354, Jul. 2018.

[2] R. Rappuoli and A. Aderem, "A 2020 vision for vaccines against HIV, tuberculosis and malaria," *Nature*, vol. 473, no. 7348, pp. 463–469, May 2011.

[3] R. A. Medina *et al.*, "Glycosylations in the Globular Head of the Hemagglutinin Protein Modulate the Virulence and Antigenic Properties of the H1N1 Influenza Viruses," *Sci. Transl. Med.*, vol. 5, no. 187, pp. 187ra70-187ra70, May 2013.

[4] D. C. Ekiert *et al.*, "A Highly Conserved Neutralizing Epitope on Group 2 Influenza A Viruses," *Science (80-.).*, vol. 333, no. 6044, pp. 843–850, Aug. 2011.

[5] C. B. Palatnik-de-Sousa, I. da S. Soares, and D. S. Rosa, "Editorial: Epitope Discovery and Synthetic Vaccine Design," *Front. Immunol.*, vol. 9, p. 826, Apr. 2018.

[6] D. R. Flower, I. K. Macdonald, K. Ramakrishnan, M. N. Davies, and I. A. Doytchinova, "Computer aided selection of candidate vaccine antigens," *Immunome Res.*, vol. 6, no. Suppl 2, p. S1, 2010.

[7] I. Margine *et al.*, "H3N2 Influenza Virus Infection Induces Broadly Reactive Hemagglutinin Stalk Antibodies in Humans and Mice," *J. Virol.*, vol. 87, no. 8, pp. 4728–4737, Apr. 2013.

[8] Centers for Disease Control and Prevention, "Selecting Viruses for the Seasonal Influenza Vaccine | CDC," 2018. [Online]. Available: https://www.cdc.gov/flu/about/season/vaccine-selection.htm. [Accessed: 17-Feb-2019].

[9] Centers for Disease Control and Prevention, "Antigenic Characterization | CDC," 2017. [Online]. Available: https://www.cdc.gov/flu/professionals/laboratory/antigenic.htm. [Accessed: 17-Feb-2019].

[10] D. H. Morris *et al.*, "Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology," *Trends in Microbiology*, vol. 26, no. 2. pp. 102–118, Feb-2018.

[11] S.-S. Wong and R. J. Webby, "Traditional and new influenza vaccines.," *Clin. Microbiol. Rev.*, vol. 26, no. 3, pp. 476–92, Jul. 2013.

[12] C. S. Anderson *et al.*, "Natural and directed antigenic drift of the H1 influenza virus hemagglutinin stalk domain.," *Sci. Rep.*, vol. 7, no. 1, p. 14614, Nov. 2017.

[13] F. Berlanda Scorza, V. Tsvetnitsky, and J. J. Donnelly, "Universal influenza vaccines: Shifting to better vaccines," *Vaccine*, vol. 34, no. 26, pp. 2926–2933, Jun. 2016.

[14] D. M. Skowronski *et al.*, "Low 2012–13 Influenza Vaccine Effectiveness Associated with Mutation in the Egg-Adapted H3N2 Vaccine Strain Not Antigenic Drift in Circulating Viruses," *PLoS One*, vol. 9, no. 3, p. e92153, Mar. 2014.

[15] S. J. Zost *et al.*, "Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 47, pp. 12578–12583, Nov. 2017.

[16] N. C. Wu *et al.*, "A structural explanation for the low effectiveness of the seasonal influenza H3N2 vaccine," *PLOS Pathog.*, vol. 13, no. 10, p. e1006682, Oct. 2017.

[17] C. I. Paules, S. G. Sullivan, K. Subbarao, and A. S. Fauci, "Chasing Seasonal Influenza — The Need for a Universal Influenza Vaccine," *N. Engl. J. Med.*, vol. 378, no. 1, pp. 7–9, Jan. 2018.

[18] CDC, "Seasonal Influenza Vaccine Effectiveness, 2004-2018," 2018. [Online]. Available: https://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm.

[19] T. R. Klingen, S. Reimering, C. A. Guzmán, and A. C. McHardy, "In Silico Vaccine Strain Prediction for Human Influenza Viruses," *Trends Microbiol.*, vol. 26, no. 2, pp. 119–131, Feb. 2018.

[20] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses.," *Microbiol. Rev.*, vol. 56, no. 1, pp. 152–79, Mar. 1992.

[21] Y. Guan, D. Vijaykrishna, J. Bahl, H. Zhu, J. Wang, and G. J. D. Smith, "The emergence of pandemic influenza viruses," *Protein Cell*, vol. 1, no. 1, pp. 9–13, Jan. 2010.

[22] G. J. D. Smith *et al.*, "Dating the emergence of pandemic influenza viruses.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, pp. 11709–11712, 2009.

[23] G. Neumann, T. Noda, and Y. Kawaoka, "Emergence and pandemic potential of swine-origin H1N1 influenza virus.," *Nature*, vol. 459, no. 7249, pp. 931–9, Jun. 2009.

[24] H. Zhang, L. Wang, R. W. Compans, and B.-Z. Wang, "Universal Influenza Vaccines, a Dream to Be Realized

Soon," *Viruses*, vol. 6, no. 5, p. 1974, Apr. 2014.
[25] Y. H. Jang and B. L. Seong, "Options and obstacles for designing a universal influenza vaccine.," *Viruses*, vol. 6, no. 8, pp. 3159–80, Aug. 2014.
[26] G. A. Kirchenbaum and T. M. Ross, "Eliciting broadly protective antibody responses against influenza," *Curr. Opin. Immunol.*, vol. 28, pp. 71–76, Jun. 2014.
[27] G. A. Sautto, G. A. Kirchenbaum, and T. M. Ross, "Towards a universal influenza vaccine: different approaches for one goal," *Virol. J.*, vol. 15, no. 1, p. 17, Dec. 2018.
[28] D. M. Carter *et al.*, "Design and Characterization of a Computationally Optimized Broadly Reactive Hemagglutinin Vaccine for H1N1 Influenza Viruses," *J. Virol.*, vol. 90, no. 9, pp. 4720–4734, May 2016.
[29] E. R. Job *et al.*, "Broadened immunity against influenza by vaccination with computationally designed influenza virus N1 neuraminidase constructs," *npj Vaccines*, vol. 3, no. 1, p. 55, Dec. 2018.
[30] L. He and J. Zhu, "Computational tools for epitope vaccine design and evaluation.," *Curr. Opin. Virol.*, vol. 11, pp. 103–12, Apr. 2015.
[31] J. L. Hurwitz, "Respiratory syncytial virus vaccine development.," *Expert Rev. Vaccines*, vol. 10, no. 10, pp. 1415–33, Oct. 2011.
[32] H. Chabas *et al.*, "Evolutionary emergence of infectious diseases in heterogeneous host populations," *PLOS Biol.*, vol. 16, no. 9, p. e2006738, Sep. 2018.
[33] B. C. Long, T. L. Goldberg, S. L. Swenson, G. Erickson, and G. Scherba, "Adaptation and Limitations of Established Hemagglutination Inhibition Assays for the Detection of Porcine Anti—Swine Influenza Virus H1N2 Antibodies," *J. Vet. Diagnostic Investig.*, vol. 16, no. 4, pp. 264–270, Jul. 2004.
[34] B. M. Giles, S. J. Bissel, D. R. Dealmeida, C. A. Wiley, and T. M. Ross, "Antibody breadth and protective efficacy are increased by vaccination with computationally optimized hemagglutinin but not with polyvalent hemagglutinin-based H5N1 virus-like particle vaccines.," *Clin. Vaccine Immunol.*, vol. 19, no. 2, pp. 128–39, Feb. 2012.
[35] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," *Eurosurveillance*, vol. 22, no. 13, p. 30494, Mar. 2017.
[36] NCBI, "Influenza virus database - NCBI." [Online]. Available: https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database. [Accessed: 28-Apr-2019].
[37] WHO, "WHO | FluID - a global influenza epidemiological data sharing platform," *WHO*, 2017.
[38] L. Liljeroos, E. Malito, I. Ferlenghi, and M. J. Bottomley, "Structural and Computational Biology in the Design of Immunogenic Vaccine Antigens.," *J. Immunol. Res.*, vol. 2015, p. 156241, 2015.
[39] A. P. Galvani, "Epidemiology meets evolutionary ecology," *Trends Ecol. Evol.*, vol. 18, no. 3, pp. 132–139, Mar. 2003.
[40] G. J. Nabel and A. S. Fauci, "Induction of unnatural immunity: prospects for a broadly protective universal influenza vaccine," *Nat. Med.*, vol. 16, no. 12, pp. 1389–1391, Dec. 2010.
[41] B. M. Giles and T. M. Ross, "A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets.," *Vaccine*, vol. 29, no. 16, pp. 3043–54, Apr. 2011.
[42] C. J. Crevar, D. M. Carter, K. Y. J. Lee, and T. M. Ross, "Cocktail of H5N1 COBRA HA vaccines elicit protective antibodies against H5N1 viruses from multiple clades," *Hum. Vaccin. Immunother.*, vol. 11, no. 3, pp. 572–583, Mar. 2015.
[43] B. M. Giles *et al.*, "A Computationally Optimized Hemagglutinin Virus-Like Particle Vaccine Elicits Broadly Reactive Antibodies that Protect Nonhuman Primates from H5N1 Infection," *J. Infect. Dis.*, vol. 205, no. 10, pp. 1562–1570, May 2012.
[44] M. F. Ducatez *et al.*, "Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 1, pp. 349–54, Jan. 2011.
[45] D. A. Baum and S. D. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*. Roberts and Co., Greenwood Village, CO, 2012.
[46] P. Lemey, A. Rambaut, A. J. Drummond, and M. a. Suchard, "Bayesian phylogeography finds its roots," *PLoS Comput. Biol.*, vol. 5, no. 9, p. e1000520, 2009.
[47] B. King and M. S. Y. Lee, "Ancestral State Reconstruction, Rate Heterogeneity, and the Evolution of Reptile Viviparity," *Syst. Biol.*, vol. 64, no. 3, pp. 532–544, May 2015.
[48] V. Brusic, "From immunoinformatics to immunomics.," *J. Bioinform. Comput. Biol.*, vol. 1, no. 1, pp. 179–81,

Apr. 2003.
[49] A. Sette, W. Fleri, B. Peters, M. Sathiamurthy, H.-H. Bui, and S. Wilson, "A roadmap for the immunomics of category A-C pathogens.," *Immunity*, vol. 22, no. 2, pp. 155–61, Feb. 2005.
[50] V. Brusic and N. Petrovsky, "Immunoinformatics and its relevance to understanding human immune disease," *Expert Rev. Clin. Immunol.*, vol. 1, no. 1, pp. 145–157, May 2005.
[51] R. Rappuoli, "Reverse vaccinology.," *Curr. Opin. Microbiol.*, vol. 3, no. 5, pp. 445–50, Oct. 2000.
[52] A. Sette and R. Rappuoli, "Reverse Vaccinology: Developing Vaccines in the Era of Genomics," *Immunity*, vol. 33, no. 4, pp. 530–541, Oct. 2010.
[53] A. S. De Groot, "Immunomics: discovering new targets for vaccines and therapeutics," *Drug Discov. Today*, vol. 11, no. 5–6, pp. 203–209, Mar. 2006.
[54] L. Potocnakova, M. Bhide, and L. B. Pulzova, "An Introduction to B-Cell Epitope Mapping and In Silico Epitope Prediction," *J. Immunol. Res.*, vol. 2016, pp. 1–11, 2016.
[55] A. Patronov and I. Doytchinova, "T-cell epitope vaccine design by immunoinformatics.," *Open Biol.*, vol. 3, no. 1, p. 120139, Jan. 2013.
[56] J. A. Greenbaum *et al.*, "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *J. Mol. Recognit.*, vol. 20, no. 2, pp. 75–82, Mar. 2007.
[57] J. L. Sanchez-Trincado, M. Gomez-Perosanz, and P. A. Reche, "Fundamentals and Methods for T- and B-Cell Epitope Prediction," *J. Immunol. Res.*, vol. 2017, pp. 1–14, 2017.
[58] Y. He, R. Rappuoli, A. S. De Groot, and R. T. Chen, "Emerging Vaccine Informatics," *J. Biomed. Biotechnol.*, vol. 2010, pp. 1–26, 2010.
[59] N. Tomar and R. K. De, "Immunoinformatics: an integrated scenario," *Immunology*, vol. 131, no. 2, pp. 153–168, Oct. 2010.
[60] L. Backert and O. Kohlbacher, "Immunoinformatics and epitope prediction in the age of genomic medicine," *Genome Med.*, vol. 7, no. 1, p. 119, Dec. 2015.
[61] N. R. Hegde, S. Gauthami, H. M. Sampath Kumar, and J. Bayry, "The use of databases, data mining and immunoinformatics in vaccinology: where are we?," *Expert Opin. Drug Discov.*, vol. 13, no. 2, pp. 117–130, Feb. 2018.
[62] H. Luo *et al.*, "Machine Learning Methods for Predicting HLA-Peptide Binding Activity," *Bioinform. Biol. Insights*, vol. 9s3, no. Suppl 3, p. BBI.S29466, Jan. 2015.
[63] H.-H. Bui, J. Sidney, K. Dinh, S. Southwood, M. J. Newman, and A. Sette, "Predicting population coverage of T-cell epitope-based diagnostics and vaccines.," *BMC Bioinformatics*, vol. 7, no. 1, p. 153, Mar. 2006.
[64] M. Molero-Abraham, E. M. Lafuente, D. R. Flower, and P. A. Reche, "Selection of conserved epitopes from hepatitis C virus for pan-populational stimulation of T-cell responses.," *Clin. Dev. Immunol.*, vol. 2013, p. 601943, Nov. 2013.
[65] A. J. McMichael, F. M. Gotch, G. R. Noble, and P. A. S. Beare, "Cytotoxic T-Cell Immunity to Influenza," *N. Engl. J. Med.*, vol. 309, no. 1, pp. 13–17, Jul. 1983.
[66] K. K. McKinstry, T. M. Strutt, and S. L. Swain, "Hallmarks of CD4 T cell immunity against influenza," *J. Intern. Med.*, vol. 269, no. 5, pp. 507–518, May 2011.
[67] N. L. La Gruta and S. J. Turner, "T cell mediated immunity to influenza: mechanisms of viral control," *Trends Immunol.*, vol. 35, no. 8, pp. 396–402, Aug. 2014.
[68] R. B. Effros, P. C. Doherty, W. Gerhard, and J. Bennink, "Generation of both cross-reactive and virus-specific T-cell populations after immunization with serologically distinct influenza A viruses.," *J. Exp. Med.*, vol. 145, no. 3, pp. 557–68, Mar. 1977.
[69] J. H. C. M. Kreijtz *et al.*, "Primary influenza A virus infection induces cross-protective immunity against a lethal infection with a heterosubtypic virus strain in mice," *Vaccine*, vol. 25, no. 4, pp. 612–620, Jan. 2007.
[70] H. J. ZWEERINK, S. A. COURTNEIDGE, J. J. SKEHEL, M. J. CRUMPTON, and B. A. ASKONAS, "Cytotoxic T cells kill influenza virus infected cells but do not distinguish between serologically distinct type A viruses," *Nature*, vol. 267, no. 5609, pp. 354–356, May 1977.
[71] V. R. S. K. Duvvuri *et al.*, "Original Article: Highly conserved cross-reactive CD4+ T-cell HA-epitopes of seasonal and the 2009 pandemic influenza viruses," *Influenza Other Respi. Viruses*, vol. 4, no. 5, pp. 249–258, Sep. 2010.
[72] V. R. Duvvuri, B. Duvvuri, V. Jamnik, J. B. Gubbay, J. Wu, and G. E. Wu, "T cell memory to evolutionarily conserved and shared hemagglutinin epitopes of H1N1 viruses: a pilot scale study," *BMC Infect. Dis.*, vol. 13,

no. 1, p. 204, Dec. 2013.
[73] A. S. De Groot, M. Ardito, E. M. McClaine, L. Moise, and W. D. Martin, "Immunoinformatic comparison of T-cell epitopes contained in novel swine-origin influenza A (H1N1) virus with epitopes in 2008–2009 conventional influenza vaccine," *Vaccine*, vol. 27, no. 42, pp. 5740–5747, Sep. 2009.
[74] X. Ge, V. Tan, P. L. Bollyky, N. E. Standifer, E. A. James, and W. W. Kwok, "Assessment of Seasonal Influenza A Virus-Specific CD4 T-Cell Responses to 2009 Pandemic H1N1 Swine-Origin Influenza A Virus," *J. Virol.*, vol. 84, no. 7, pp. 3312–3319, Apr. 2010.
[75] J. A. Greenbaum *et al.*, "Pre-existing immunity against swine-origin H1N1 influenza viruses in the general human population.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 48, pp. 20365–70, Dec. 2009.
[76] J. T. Weinfurter *et al.*, "Cross-Reactive T Cells Are Involved in Rapid Clearance of 2009 Pandemic H1N1 Influenza Virus in Nonhuman Primates," *PLoS Pathog.*, vol. 7, no. 11, p. e1002381, Nov. 2011.
[77] T. M. Wilkinson *et al.*, "Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans," *Nat. Med.*, vol. 18, no. 2, pp. 274–280, Feb. 2012.
[78] ClinicalTrials.gov Identifier:NCT01265914, "A Study to Evaluate the Safety, Tolerability and Immunogenicity of a Universal Influenza A Vaccine," 2010. [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT01265914.
[79] J. N. Francis *et al.*, "A novel peptide-based pan-influenza A vaccine: A double blind, randomised clinical trial of immunogenicity and safety," *Vaccine*, vol. 33, no. 2, pp. 396–402, Jan. 2015.
[80] O. Pleguezuelos, S. Robinson, G. A. Stoloff, and W. Caparrós-Wanderley, "Synthetic Influenza vaccine (FLU-v) stimulates cell mediated immunity in a double-blind, randomised, placebo-controlled Phase I trial," *Vaccine*, vol. 30, no. 31, pp. 4655–4660, Jun. 2012.
[81] O. Pleguezuelos *et al.*, "A Synthetic Influenza Virus Vaccine Induces a Cellular Immune Response That Correlates with Reduction in Symptomatology and Virus Shedding in a Randomized Phase Ib Live-Virus Challenge in Humans," *Clin. Vaccine Immunol.*, vol. 22, no. 7, pp. 828–835, Jul. 2015.
[82] E. van Doorn *et al.*, "Evaluation of the immunogenicity and safety of different doses and formulations of a broad spectrum influenza vaccine (FLU-v) developed by SEEK: study protocol for a single-center, randomized, double-blind and placebo-controlled clinical phase IIb trial," *BMC Infect. Dis.*, vol. 17, no. 1, p. 241, Dec. 2017.
[83] ClinicalTrials.gov Identifier: NCT03450915, "A Pivotal Trial to Assess the Safety and Clinical Efficacy of the M-001 as a Standalone Universal Flu Vaccine." [Online]. Available: https://clinicaltrials.gov/ct2/show/NCT03450915?term=epitope&cond=Influenza&rank=6.
[84] T. Gottlieb and T. Ben-Yedidia, "Epitope-based approaches to a universal influenza vaccine," *J. Autoimmun.*, vol. 54, pp. 15–20, Nov. 2014.
[85] J. Atsmon *et al.*, "Priming by a novel universal influenza vaccine (Multimeric-001)—A gateway for improving immune response in the elderly population," *Vaccine*, vol. 32, no. 44, pp. 5816–5823, Oct. 2014.
[86] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake, "The promise and challenge of high-throughput sequencing of the antibody repertoire," *Nat. Biotechnol.*, vol. 32, no. 2, pp. 158–168, Feb. 2014.
[87] A. Six, B. Bellier, V. Thomas-Vaslin, and D. Klatzmann, "Systems biology in vaccine design.," *Microb. Biotechnol.*, vol. 5, no. 2, pp. 295–304, Mar. 2012.
[88] A. D. Yermanos, A. K. Dounas, T. Stadler, A. Oxenius, and S. T. Reddy, "Tracing Antibody Repertoire Evolution by Systems Phylogeny.," *Front. Immunol.*, vol. 9, p. 2149, 2018.
[89] W. C. Koff *et al.*, "Accelerating Next-Generation Vaccine Development for Global Disease Prevention," *Science (80-.).*, vol. 340, no. 6136, pp. 1232910–1232910, May 2013.
[90] W. C. Koff, I. D. Gust, and S. A. Plotkin, "Toward a Human Vaccines Project," *Nat. Immunol.*, vol. 15, no. 7, pp. 589–592, Jul. 2014.
[91] D. Sok *et al.*, "The Effects of Somatic Hypermutation on Neutralization and Binding in the PGT121 Family of Broadly Neutralizing HIV Antibodies," *PLoS Pathog.*, vol. 9, no. 11, p. e1003754, Nov. 2013.
[92] X. Wu *et al.*, "Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing," *Science (80-.).*, vol. 333, no. 6049, pp. 1593–1602, Sep. 2011.
[93] Y.-G. Zhu *et al.*, "Diverse and abundant antibiotic resistance genes in Chinese swine farms.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 9, pp. 3435–40, Feb. 2013.
[94] R. R. María, C. J. Arturo, J. A. Alicia, M. G. Paulina, and A. O. Gerardo, "The Impact of Bioinformatics on

Vaccine Design and Development," in *Vaccines*, InTech, 2017.
[95] M. H. V. Van Regenmortel, "Structure-Based Reverse Vaccinology Failed in the Case of HIV Because it Disregarded Accepted Immunological Theory," *Int. J. Mol. Sci.*, vol. 17, no. 9, Sep. 2016.
[96] Z. Zhu, C. Zhang, and W. Song, "Rational derivation, extension, and cyclization of self-inhibitory peptides to target TGF-β/BMP signaling in ONFH," *Amino Acids*, vol. 49, no. 2, pp. 283–290, Feb. 2017.
[97] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-Grained Protein Models and Their Applications," *Chem. Rev.*, vol. 116, no. 14, pp. 7898–7936, Jul. 2016.
[98] D. J. Laddy, J. Yan, N. Corbitt, D. Kobasa, G. P. Kobinger, and D. B. Weiner, "Immunogenicity of novel consensus-based DNA vaccines against avian influenza," *Vaccine*, vol. 25, no. 16, pp. 2984–2989, Apr. 2007.
[99] C.-Y. Wu *et al.*, "Mammalian Expression of Virus-Like Particles for Advanced Mimicry of Authentic Influenza Virus," *PLoS One*, vol. 5, no. 3, p. e9784, Mar. 2010.
[100] R. A. Bright *et al.*, "Cross-Clade Protective Immune Responses to Influenza Viruses with H5N1 HA and NA Elicited by an Influenza Virus-Like Particle," *PLoS One*, vol. 3, no. 1, p. e1501, Jan. 2008.
[101] A. Ben-Dor, G. Lancia, R. Ravi, and J. Perone, "Banishing bias from consensus sequences," Springer, Berlin, Heidelberg, 1997, pp. 247–261.
[102] T. M. Wong *et al.*, "Computationally Optimized Broadly Reactive Hemagglutinin Elicits Hemagglutination Inhibition Antibodies against a Panel of H3N2 Influenza Virus Cocirculating Variants.," *J. Virol.*, vol. 91, no. 24, pp. e01581-17, Dec. 2017.
[103] J. W. Thornton, "Resurrecting ancient genes: experimental analysis of extinct molecules," *Nat. Rev. Genet.*, vol. 5, no. 5, pp. 366–375, May 2004.
[104] S. A. Lim, K. M. Hart, M. J. Harms, and S. Marqusee, "Evolutionary trend toward kinetic stability in the folding trajectory of RNases H.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 46, pp. 13045–13050, 2016.
[105] Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites.," *Genetics*, vol. 155, no. 1, pp. 431–49, May 2000.
[106] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology.," *Science*, vol. 294, no. 5550, pp. 2310–4, Dec. 2001.
[107] W. Cai, J. Pei, and N. V Grishin, "Reconstruction of ancestral protein sequences and its applications.," *BMC Evol. Biol.*, vol. 4, no. 1, p. 33, Sep. 2004.
[108] G. Baele, M. A. Suchard, A. Rambaut, and P. Lemey, "Emerging Concepts of Data Integration in Pathogen Phylodynamics.," *Syst. Biol.*, vol. 66, no. 1, pp. e47–e65, Jun. 2017.
[109] E. Kirkpatrick, X. Qiu, P. C. Wilson, J. Bahl, and F. Krammer, "The influenza virus hemagglutinin head evolves faster than the stalk domain," *Sci. Rep.*, vol. 8, no. 1, p. 10432, Dec. 2018.
[110] X. Qiu and J. Bahl, "Structurally informed evolutionary models improve phylogenetic reconstruction for emerging, seasonal, and pandemic influenza viruses," *bioRxiv*, Jan. 2017.
[111] C. L. Kleinman, N. Rodrigue, N. Lartillot, and H. Philippe, "Statistical Potentials for Improved Structurally Constrained Evolutionary Models," 2010.
[112] J. D. Bloom, "An Experimentally Informed Evolutionary Model Improves Phylogenetic Fit to Divergent Lactamase Homologs," *Mol. Biol. Evol.*, vol. 31, no. 10, pp. 2753–2769, Oct. 2014.
[113] T. R. Booker and P. D. Keightley, "Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome," *Mol. Biol. Evol.*, vol. 35, no. 12, pp. 2971–2988, Oct. 2018.
[114] C. L. Kleinman, N. Rodrigue, N. Lartillot, and H. Philippe, "Statistical Potentials for Improved Structurally Constrained Evolutionary Models," *Mol. Biol. Evol.*, vol. 27, no. 7, pp. 1546–1560, Jul. 2010.
[115] J. D. J. D. Bloom, "An Experimentally Determined Evolutionary Model Dramatically Improves Phylogenetic Fit," *Mol. Biol. Evol.*, vol. 31, no. 8, pp. 1956–1978, Aug. 2014.
[116] D. M. Fowler *et al.*, "High-resolution mapping of protein sequence-function relationships," *Nat. Methods*, vol. 7, no. 9, pp. 741–746, Sep. 2010.
[117] C. L. Araya and D. M. Fowler, "Deep mutational scanning: assessing protein function on a massive scale," *Trends Biotechnol.*, vol. 29, no. 9, pp. 435–442, Sep. 2011.
[118] M. W. Traxlmayr *et al.*, "Construction of a Stability Landscape of the CH3 Domain of Human IgG1 by Combining Directed Evolution with High Throughput Sequencing," *J. Mol. Biol.*, vol. 423, no. 3, pp. 397–412, Oct. 2012.
[119] D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, "Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein.," *RNA*, vol. 19, no. 11, pp. 1537–51, Nov.

2013.
[120] B. P. Roscoe, K. M. Thayer, K. B. Zeldovich, D. Fushman, and D. N. A. Bolon, "Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate," *J. Mol. Biol.*, vol. 425, no. 8, pp. 1363–1377, Apr. 2013.
[121] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, "A comprehensive, high-resolution map of a gene's fitness landscape.," *Mol. Biol. Evol.*, vol. 31, no. 6, pp. 1581–92, Jun. 2014.
[122] V. Hanson-Smith, B. Kolaczkowski, and J. W. Thornton, "Robustness of ancestral sequence reconstruction to phylogenetic uncertainty.," *Mol. Biol. Evol.*, vol. 27, no. 9, pp. 1988–99, Sep. 2010.
[123] E. C. Holmes, "What can we predict about viral evolution and emergence?," *Curr. Opin. Virol.*, vol. 3, no. 2, pp. 180–4, Apr. 2013.
[124] M. Kirschner and J. Gerhart, "Evolvability.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 15, pp. 8420–7, Jul. 1998.
[125] E. M. Volz, K. Koelle, and T. Bedford, "Viral Phylodynamics," *PLoS Comput. Biol.*, vol. 9, no. 3, p. e1002947, Mar. 2013.
[126] B. T. Grenfell *et al.*, "Unifying the Epidemiological and Evolutionary Dynamics of Pathogens," *Science (80-.).*, vol. 303, no. 5656, pp. 327–332, Jan. 2004.
[127] E. Visher, S. E. Whitefield, J. T. McCrone, W. Fitzsimmons, and A. S. Lauring, "The Mutational Robustness of Influenza A Virus," *PLOS Pathog.*, vol. 12, no. 8, p. e1005856, Aug. 2016.
[128] B. Thyagarajan and J. D. Bloom, "The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin," *Elife*, vol. 3, Jul. 2014.
[129] J. D. Bloom, L. I. Gong, and D. Baltimore, "Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance," *Science (80-.).*, vol. 328, no. 5983, pp. 1272–1275, Jun. 2010.
[130] S. H. Olson *et al.*, "Sampling Strategies and Biodiversity of Influenza A Subtypes in Wild Birds," *PLoS One*, vol. 9, no. 3, p. e90826, Mar. 2014.
[131] T. R. Klingen *et al.*, "Sweep Dynamics (SD) plots: Computational identification of selective sweeps to monitor the adaptation of influenza A viruses," *Sci. Rep.*, vol. 8, no. 1, p. 373, Dec. 2018.
[132] W. K. Ampofo *et al.*, "Strengthening the influenza vaccine virus selection and development process," *Vaccine*, vol. 33, no. 36, pp. 4368–4382, Aug. 2015.
[133] Institute of Medicine (US) Forum on Microbial Threats, *The Domestic and International Impacts of the 2009-H1N1 Influenza A Pandemic: Global Challenges, Global Solutions: Workshop Summary*. Washington (DC): National Academies Press (US), 2010.
[134] B. J. Hoye, V. J. Munster, H. Nishiura, M. Klaassen, and R. A. M. Fouchier, "Surveillance of wild birds for avian influenza virus.," *Emerg. Infect. Dis.*, vol. 16, no. 12, pp. 1827–34, Dec. 2010.
[135] R. B. Squires *et al.*, "Influenza research database: an integrated bioinformatics resource for influenza research and surveillance.," *Influenza Other Respi. Viruses*, vol. 6, no. 6, pp. 404–16, Nov. 2012.
[136] G. Gunnarsson *et al.*, "Disease dynamics and bird migration--linking mallards Anas platyrhynchos and subtype diversity of the influenza A virus in time and space.," *PLoS One*, vol. 7, no. 4, p. e35679, 2012.
[137] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Glob. Challenges*, vol. 1, no. 1, pp. 33–46, Jan. 2017.
[138] G. Baele, P. Lemey, T. Bedford, A. Rambaut, M. A. Suchard, and A. V. Alekseyenko, "Improving the accuracy of demographic and molecular clock model comparison while accomodating phylogenetic uncertainty," *Mol. Biol. Evol.*, Jan. 2012.
[139] G. Dudas, L. M. Carvalho, A. Rambaut, and T. Bedford, "MERS-CoV spillover at the camel-human interface," *Elife*, vol. 7, Jan. 2018.
[140] N. F. Mu, D. A. Rasmussen, and T. Stadler, "The Structured Coalescent and Its Approximations," vol. 34, no. 11, pp. 2970–2981, 2017.
[141] S. Duchene, R. Bouckaert, D. A. Duchene, T. Stadler, and A. J. Drummond, "Phylodynamic Model Adequacy Using Posterior Predictive Simulations," *Syst. Biol.*, vol. 68, no. 2, pp. 358–364, Mar. 2019.
[142] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, Jul. 1959.
[143] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Appl. Comput. Informatics*, vol. 15, no. 1, pp. 27–33, Jan. 2019.
[144] B. A. Fritz *et al.*, "Using machine learning techniques to develop forecasting algorithms for postoperative complications: protocol for a retrospective study.," *BMJ Open*, vol. 8, no. 4, p. e020124, 2018.
[145] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine learning for

integrating data in biology and medicine: Principles, practice, and opportunities," *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019.

[146] M. A. Salama, A. E. Hassanien, and A. Mostafa, "The prediction of virus mutation using neural networks and rough set techniques.," *EURASIP J. Bioinform. Syst. Biol.*, vol. 2016, no. 1, p. 10, Dec. 2016.

[147] H. Shim, "Feature Learning of Virus Genome Evolution With the Nucleotide Skip-Gram Neural Network," *Evol. Bioinforma.*, vol. 15, p. 117693431882107, Jan. 2019.