PLOS PATHOGENS

RESEARCH ARTICLE

Host diversity and behavior determine patterns of interspecies transmission and geographic diffusion of avian influenza A subtypes among North American wild reservoir species

Joseph T. Hicks[1], **Kimberly Edwards**[2¤], **Xueting Qiu**[1], **Do-Kyun Kim**[3], **James E. Hixson**[3], **Scott Krauss**[2], **Richard J. Webby**[2], **Robert G. Webster**[2], **Justin Bahl**[1]*

1 Center for Ecology of Infectious Diseases, Department of Infectious Diseases, College of Veterinary Medicine, Department of Epidemiology and Biostatistics, College of Public Health, Institute of Bioinformatics, University of Georgia, Athens, Georgia, United States of America, **2** Department of Infectious Disease, St. Jude Children's Research Hospital, Memphis, Tennessee, United States of America, **3** University of Texas Health Science Center at Houston School of Public Health, Houston, Texas, United States of America

¤ Current address: School of Public Health, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China, HKU-Pasteur Research Pole, School of Public Health, The University of Hong Kong, Hong Kong, China

* justin.bahl@uga.edu

OPEN ACCESS

Citation: Hicks JT, Edwards K, Qiu X, Kim D-K, Hixson JE, Krauss S, et al. (2022) Host diversity and behavior determine patterns of interspecies transmission and geographic diffusion of avian influenza A subtypes among North American wild reservoir species. PLoS Pathog 18(4): e1009973. https://doi.org/10.1371/journal.ppat.1009973

Editor: Adi Stern, Tel Aviv University, ISRAEL

Received: September 14, 2021

Accepted: March 7, 2022

Published: April 13, 2022

Data Availability Statement: Newly sequenced AIV nucleotide sequences have been deposited in GenBank with accession numbers provided in Table M in S1 Text.

Funding: This work was funded in part by the National Institute of Allergy and Infectious Diseases (HHSN272201400006C (RJW, JB) and HHSN272201400007C (JB), NIAID Centers of Excellence for Influenza Research and Surveillance, CEIRS; 75N93021C00018 (JB) NIAID Centers of Excellence for Influenza Research and Response,

Abstract

Wild birds can carry avian influenza viruses (AIV), including those with pandemic or panzootic potential, long distances. Even though AIV has a broad host range, few studies account for host diversity when estimating AIV spread. We analyzed AIV genomic sequences from North American wild birds, including 303 newly sequenced isolates, to estimate interspecies and geographic viral transition patterns among multiple co-circulating subtypes. Our results show high transition rates within Anseriformes and Charadriiformes, but limited transitions between these orders. Patterns of transition between species were positively associated with breeding habitat range overlap, and negatively associated with host genetic distance. Distance between regions (negative correlation) and summer temperature at origin (positive correlation) were strong predictors of transition between locations. Taken together, this study demonstrates that host diversity and ecology can determine evolutionary processes that underlie AIV natural history and spread. Understanding these processes can provide important insights for effective control of AIV.

Author summary

Avian influenza viruses (AIV) maintained in wild birds provide much of the genetic diversity for emerging panzootic and pandemic influenza A viruses. AIV's wide-ranging host and geographic distribution complicates understanding the determinants of interspecies transmission and distribution. We estimated geographic, ecological, and host

CEIRR) and ALSAC (RGW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

characteristics associated with AIV movement and cross species transmission using newly sequenced and publicly available AIV genome data sampled from North American ducks, geese, shorebirds, and gulls between 2005 and 2016, We found AIV dispersal among hosts are associated with host genetic relatedness, breeding distribution overlap, and migratory behaviors. Geographic dispersal is strongly limited by physical distance despite the long distances many host birds migrate. Higher geographic movement rates were associated with higher summer temperatures, likely associated with timing of bird behaviors such as timing of breeding. Taken together, this study demonstrates that host diversity and ecology can determine AIV natural history, spread and emergence risk.

Introduction

Avian influenza viruses (AIV) are globally distributed pathogens maintained within wild waterfowl (order Anseriformes) and shorebirds (order Charadriiformes) [1]. Despite being largely asymptomatic within wild birds, AIV provide cause for global concern as sources of influenza A viral diversity for domestic avian and mammalian hosts [2]. AIV hemagglutinin (HA) subtypes H5 and H7 have repeatedly evolved into highly pathogenic viruses in domestic poultry causing severe losses [3]. Furthermore, all modern pandemic influenza A viruses contain gene segments of avian origin, suggesting reassortment with avian viruses plays a crucial role in pandemic emergence [4]. The segmented genome is an important characteristic of influenza viruses because it facilitates continual reassortment and promotes diversity of AIV within wild avian populations [1,5,6]. Due to the unlinked nature of the AIV genes, each segment can be considered as an independent hereditary particle with its own evolutionary history [7].

Understanding the host behavior and environmental drivers of AIV susceptibility and dispersal remain a top priority for avian influenza surveillance, but the vast array of susceptible host species and ecological variables hampers the prediction of AIV emergence and incidence [8]. Surveillance data and spatial analysis have begun to assess the association between avian influenza prevalence and environmental variables, including land use [9,10], temperature measures [9,11,12], altitude [10], distance to water [10], and precipitation [11]. Fewer studies have assessed the impact of host characteristics on the prevalence of AIV within individual avian species although migration distance, habitat water salinity, and surface foraging methods have been implicated as important predictors in one such study [13]. Further, host community species composition and host phylogenetic relatedness may help explain spatial patterns of highly pathogenic avian influenza H5N1 outbreaks among wild birds in Europe [14]. Sequence data acquired by viral surveillance provide further information to understand AIV dynamics. Because viral evolution, host ecology, and environmental factors necessarily interact, phylogenetic studies can help elucidate the paths of AIV dispersal (Fig 1). For example, previous phylogeographic analysis [7] of AIV within North America provided evidence that migratory flyways are not as strong a barrier to viral dispersal as previously believed [15].

Although phylogenetic studies to date have been able to interrogate the impact of broad ecological patterns such as migratory flyways and interhemispheric viral exchange, few incorporate characteristics of the location or host from which the virus was sampled. Prevalence studies include these characteristics into regression and spatial models but are limited due to the long-distance migration of wildlife hosts. Generalized linear models (GLM) implemented within a Bayesian phylogenetic framework have made it possible to include environmental and ecological covariates into phylogenetic models [16,17]. This allows the simultaneous

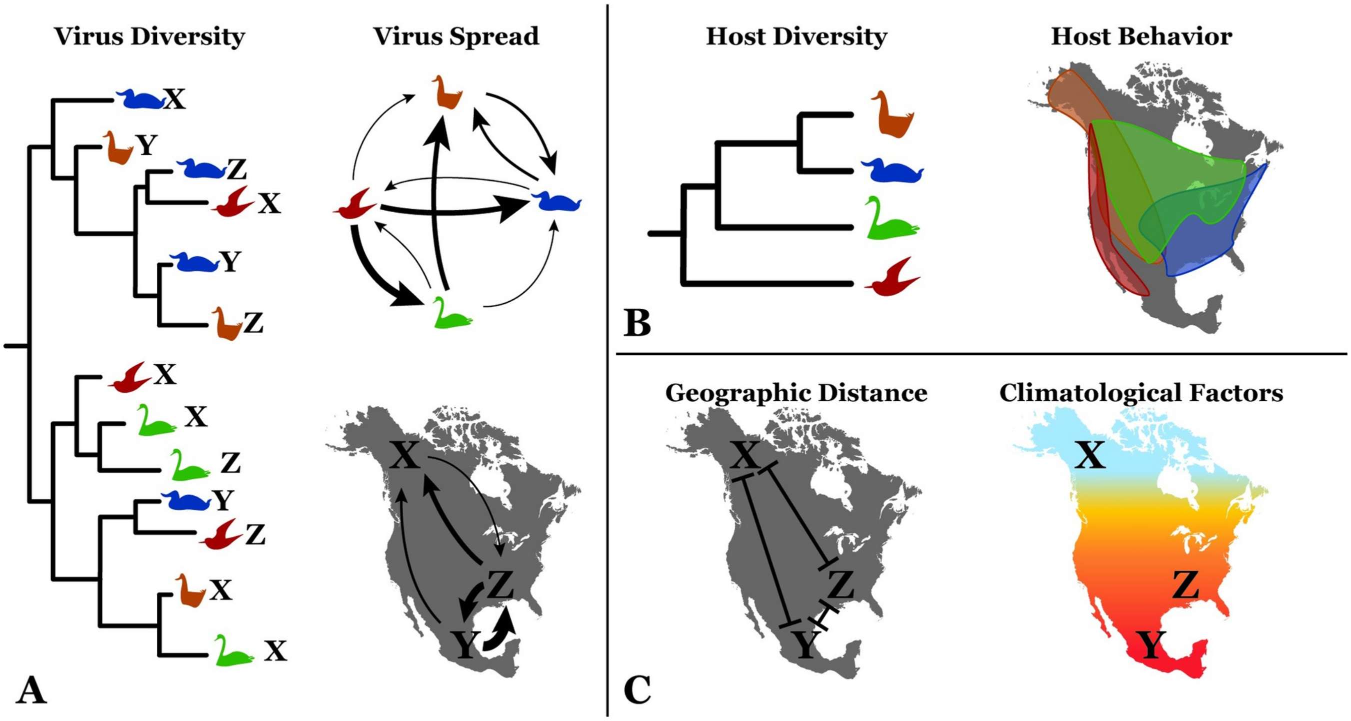

Fig 1. Interaction of viral evolution, host ecology and the environment. Viral genetic sequences contain information regarding virus evolution and diversity (A). Because their evolution occurs at a rapid pace, evolutionary patterns can be used in conjunction with location and species data to infer rates of viral dispersal among sampled geographic regions and host species. Many factors may influence observed virus transmission and spread. For instance, host factors (B) such as relatedness of host species and overlap of habitat distributions may be associated with viral transitions between host species. Further, environmental factors (C) may also play a role in the spatial diffusion of the virus. By incorporating viral, host and environmental information into computational models, the impact of host and environmental characteristics on virus spread can be estimated. Public domain map of North America was accessed from *The World Factbook 2021*, Central Intelligence Agency (https://www.cia.gov/the-world-factbook/static/09f14262b8d528e66631646c85e0edc0/north_america_pol.pdf).

https://doi.org/10.1371/journal.ppat.1009973.g001

inference of viral transition rates among specified traits (i.e., hosts or locations) and their association with covariates that may drive viral movement. This approach has been adapted to investigate the role of anthropogenic and environmental variables on the diffusion of avian influenza within China [18] as well as of avian influenza subtype H9N2 on a global scale [19]. The GLM has also been used to uncover the impact of host behavior on the dispersal of rabies virus among bat species [20]. Previous analyses have also demonstrated the importance of environmental transmission on AIV prevalence and evolution [21,22], suggesting ecological factors may influence AIV transmission. Understanding how avian host characteristics and environmental variables impact zoonotic transmission and geographic dispersal will be key to identify surveillance priorities among species and locations. In the presented analysis, using an extensive publicly available dataset of multiple AIV subtypes collected from North American wild birds supplemented with newly sequenced surveillance samples, we implemented the GLM to assess the impact of ecological and environmental characteristics on the dispersal of AIV across the North American continent and among frequently sampled Anseriformes and Charadriiformes.

Results

Summary of sequenced data

The sequenced samples are drawn from a long-term systematic influenza surveillance program in ducks in Alberta, Canada since 1976, and in shorebirds and gulls at Delaware Bay (Delaware and New Jersey) since 1985. This publicly available data represent research and surveillance efforts to collect, isolate, and sequence AIV across locations and host species to support research into risks of zoonotic transmission, pandemic influenza preparedness, and influenza vaccine development. The newly sequenced AIV isolates are part of this continued effort, especially targeting under-collected shorebird species. Table A in S1 Text describes the characteristics of 303 newly sequenced AIV isolates, which originated from samples collected from wild birds between 2003 and 2016 in Delaware Bay, New Jersey, United States (86.5%) and Alberta, Canada (13.5%). All sequenced samples from Alberta were exclusively of waterfowl origin (order Anseriformes). Delaware Bay samples originated from shorebirds (order Charadriiformes), except for a single Canada goose (*Branta canadensis*) sample. Among all newly sequenced viral isolates, most (60.4%) were isolated from samples collected from the ruddy turnstone (*Arenaria interpres*), a migratory shorebird of the wader family with near global distribution and intercontinental migration patterns. Nine samples were found to be co-infected with avian paramyxovirus and were excluded from further analysis. The most frequently isolated hemagglutinin (HA) subtype was H10 (27.4%), followed by H12 (18.8%) and H3 (9.2%). Most HA subtypes were collected in Delaware Bay, including H1, H3, H5, H6, H7, H8, H9, H10, H11, H12, H13, and H16. Only H4 was exclusively isolated from Alberta. The most frequent neuraminidase (NA) subtypes were N5 (20.1%), N7 (13.9%), and N8 (13.5%). All but two NA subtypes were isolated from both Delaware Bay and Alberta; N3 and N9 were only recovered from Delaware Bay.

Evolutionary comparison between segments and subtypes

The newly sequenced data were aligned with publicly available sequences and subsampled in two methods to help address sampling biases in surveillance: a phylogenetic diversity-based analysis method (PDA sample) and a simple stratified random sample method (stratified sample). Evolutionary models were constructed separately for each gene segment; HA, NA, and NS segment datasets were further subdivided by subtype or allele. In general, the two samples produced similar comparative relationships of evolutionary parameters among the analyzed gene segments; however, the stratified sample had consistently lower molecular clock rates and effective population sizes compared to the PDA sample (Fig A and Table B in S1 Text). Compared with the HA and NA surface proteins, the internal gene segments tended to have older times to the most recent common ancestor (TMRCA) (Fig A and Table B in S1 Text), except the included sequences of the NS gene B allele, which shared a common ancestor around 1965 (95% Highest Posterior Density Bayesian Credibility Interval (HPD) 1958.5–1969.7). As compared to the HA and NA surface proteins which contend with greater selection pressure, the internal gene segments tended to have slower evolutionary rates as measured by the mean substitution rate of the uncorrelated relaxed molecular clock.

In general, the effective population size of gene segments appeared dependent on the viral genetic diversity present in the sampled sequence data. For example, the larger internal gene segment data sets produced larger effective population sizes as compared to the surface proteins. With approximately half the sample size, the NS alleles differed from the remaining internal gene segments, with median effective population sizes around half that of PB2, PB1, PA, NP, and MP segments. Similarly, the effective population sizes of the various HA and NA subtypes were considerably lower than that of the non-subdivided internal gene segments. While the sum genetic diversity of the NS, HA, and NA gene segments respectively, was divided between datasets (allele or subtype), the difference in effective population size estimates likely reflects the flat fitness landscape of the internal gene segments where segments frequently reassort among subtypes [5].

Discrete trait diffusion models

Two discrete trait diffusion models were estimated for each of 22 gene segment or subtype datasets to assess how AIV disperses among host species and geographic regions of North America. North American regions were categorized into eight Canadian provinces and territories, ten United States climate regions, one Mexican state, and Guatemala. Hosts were of 16 species of order Anseriformes (waterfowl) and five species of order Charadriiformes (gulls and shorebirds). Geographic regions and host species are listed in Fig 2. Henceforth, all host species will be referenced using common names.

Across all included gene segments and subtypes, the distribution of hosts statistically differed between the PDA and stratified subsampling strategies (p = 0.047), but no statistical difference was noted in geographic distribution (p = 0.08). The distribution of hosts and geographic regions were similar among the internal gene segments by both PDA and stratified subsampling strategies (Tables C-E in S1 Text). The subsampling strategy had an effect on the temporal distribution of host and geographic region variables; proportions were more consistent between 2005 and 2016 in the stratified sample compared with the PDA sample (Figs C-F in S1 Text). Within both the PDA and stratified samples, Alaska, the Ohio Valley, and the Northeast were the most frequently represented regions between 2005 and 2016 among the internal genes. While regional distribution varied considerably across HA and NA subtypes Alaska and Northeast were frequently represented. Although host species distribution differed among gene segments and subtypes, all shared mallard as the most frequently sampled avian species (28.1–52.7%). The stratified sample attempted to counteract the oversampling of mallards resulting in lower percentages of these hosts within the sample (20.6–35.7%). The stratified method also tended to increase the frequency of the more sparsely represented regions and hosts within the models.

Asymmetrical diffusion models allow directionality to be inferred so that each viral transition is characterized by a source (i.e., origin of the virus) and a sink (i.e., destination). In this analysis, we estimated transition rates between discrete trait categories. The transition rate is the number of times per year across the phylogenetic history that an ancestral virus was estimated to move between regions or host species. The transition rate from location A to location B may differ from the rate in the opposing direction (from location B to A). Across all gene segments and subtypes, mallards were supported as the source of AIV for green-winged teals and northern shovelers within the stratified sample (Fig E in S1 Text). These rates were also supported in the PDA sample in all gene segments and subtypes except N6. The rates from mallards to blue-winged teals were also supported among all gene segments and subtypes except H7 in both samples. Within the PDA sample, American black ducks, blue-winged teals, Canada geese, greater white-fronted geese, ring-necked ducks, and snow geese were only supported to receive viral diversity from mallards. In contrast, only American black ducks and Canada geese exclusively received virus from mallards in the stratified sample. Similarly, ruddy turnstones were the exclusive source of viral diversity for laughing gulls and sanderlings in both samples, as well as red knots in the stratified sample.

A single host diffusion model was also jointly estimated across all internal gene segments, all HA subtypes, and all NA subtypes (Figs 2 and F in S1 Text). For each joint host model, the highest transition rate occurred from blue-winged teals to mallards (PDA internal gene model: 40.7 transitions/year, 95% HPD 32.5–49.2; PDA HA model: 22.0 transitions/year, 95% HPD 16.6–27.5; Stratified NA model: 22.6 transitions/year, 95% HPD 14.2–31.7; Tables G-I in S1 Text). In all three joint host models, all species were supported as receiving virus from at least one other species, except snow geese within the NA models. Not all species acted as a source, however. Cinnamon teals, gadwalls, and red knots were included in all three joint host models,

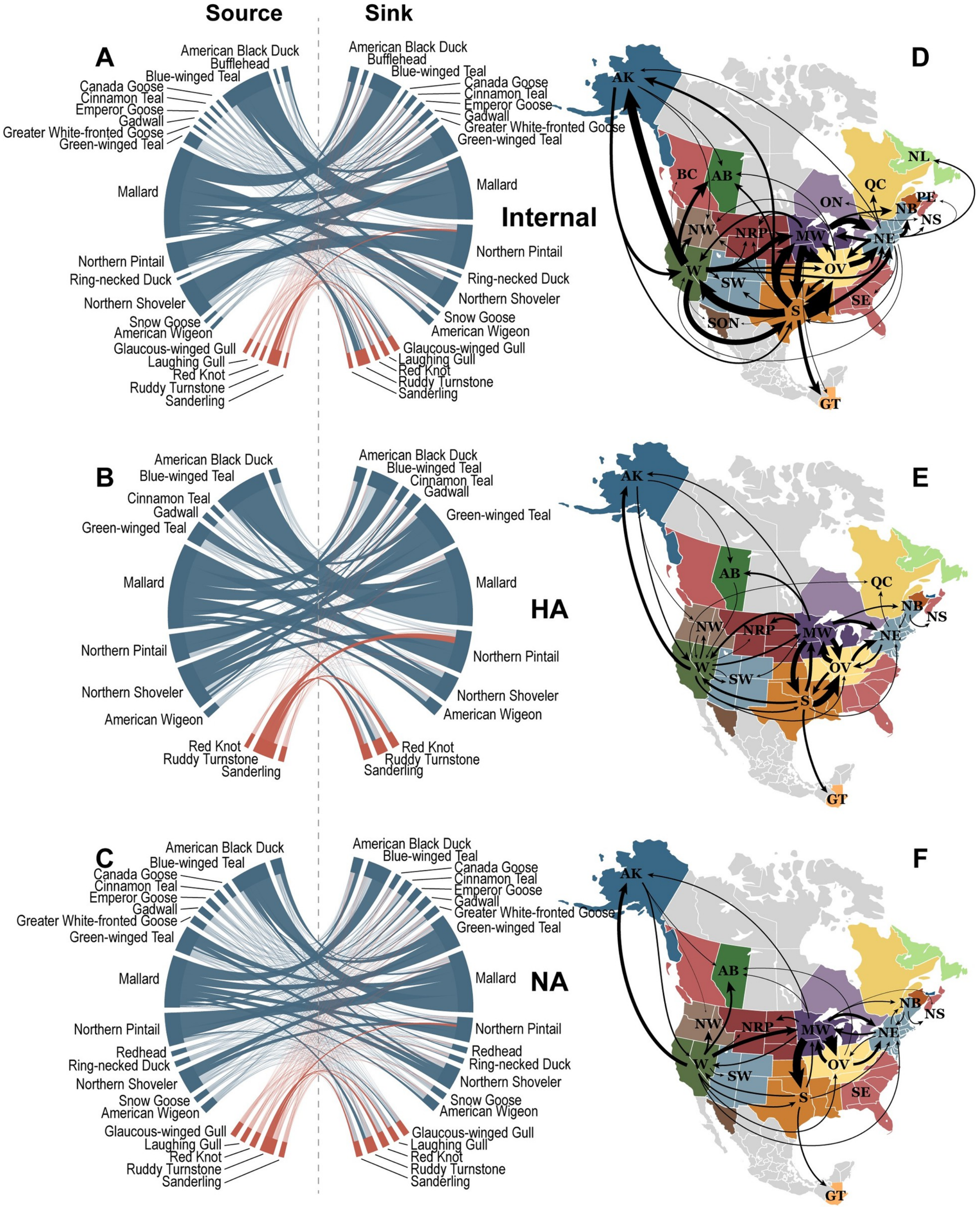

Fig 2. Discrete trait diffusion models of North American avian influenza using a sample of genetic sequences based on phylogenetic diversity. Host models (left) are presented for combined internal gene segment (A), hemagglutinin gene subtype (B), and neuraminidase gene subtype (C) models. Source host species on the left of the chord diagrams contribute viral diversity to sink host species on the right. The magnitude of the viral transition rate is proportional to the width of the band, and statistically supported rates are darkened. Bands are colored by the host order of the source species (Charadriiformes–red; Anseriformes–blue). Similarly, geographic models (right) are summarized for combined internal gene

segment (D), hemagglutinin gene subtype (E), and neuraminidase gene subtype (F) models. Arrow width is proportional to the magnitude of the transition rate, and only statistically supported rates are displayed. (AK–Alaska, AB–Alberta, BC–British Columbia, GT–Guatemala, MW–Midwest, NB–New Brunswick, NE–Northeast, NL–Newfoundland and Labrador, NRP–Northern Rockies and Plains, NS–Nova Scotia, NW–Northwest, OV–Ohio Valley, PE–Prince Edward Island, QC–Quebec, S–South, SE–Southeast, SON–Sonora, SW–Southwest, W–West). Public domain maps of United States, Canada, and Mexico states, and provinces were accessed from Wikimedia Commons, author Alex Covarrubias (https://commons.wikimedia.org/wiki/File:North_America_second_level_political_division.svg). Public domain map of Guatemala was accessed from *The World Factbook 2021*, Central Intelligence Agency (https://www.cia.gov/the-world-factbook/static/09f14262b8d528e66631646c85e0edc0/north_america_pol.pdf).

https://doi.org/10.1371/journal.ppat.1009973.g002

but none were supported to contribute AIV genetic diversity to any other host species. In addition, Canada geese, emperor geese, ring-necked ducks, snow geese, laughing gulls, and sanderlings, which were only included in the internal gene and NA models, were also not supported as viral sources. A marked difference between the PDA and stratified samples can be noted in this regard. Whereas green-winged teals were not supported as a source of virus for any other species within the PDA internal gene segment model, the stratified sample estimates green-winged teals as the source of viral genetic diversity for nine other avian species within the internal gene model. This provides evidence that sampling methods can influence discrete trait diffusion model results.

Among the North American regional models, no single transition rate was supported across all gene segments or subtypes (Fig G in S1 Text). The internal gene segments and the HA and NA subtype models differed in regard to support for the Northeast region of the United States as a source of AIV for other North American regions. Across the HA and NA subtypes, there is only sporadic support for the Northeast as a source of AIV, with only three rates among the subtypes supported in the PDA sample, and four supported in the stratified sample. In contrast, each internal gene segment model within the PDA sample has at least six rates in support of the Northeast as a source. Support for a Northeastern source is less consistent across the stratified sample internal gene segments: while nine rates are supported in the PB2 model, no rates in the PB1 model are supported. The internal genes further differ between PDA and stratified samples in terms of their support for New Brunswick as a viral source. No New Brunswick source rates are supported within the PDA internal gene segment models, yet 16 rates are supported in the stratified models among four sink regions (Northeast, Nova Scotia, Ohio Valley, and Prince Edward Island).

Among the three joint geographic models, the internal gene model has the largest number of decisively supported transition rates between regions (Figs 2 and F in S1 Text). The highest rate among the internal genes occurred from the South to the Ohio Valley (PDA sample: 48.5 transitions/year, 95% HPD 42.5–55.0; Table J in S1 Text). The highest transition rate among both HA and NA models occurred from the Midwest to the Ohio Valley (stratified HA: 17.8 transitions/year, 95% HPD 12.7–23.4; PDA NA: 21.3 transitions/year, 95% HPD 16.7–26.2; Tables K and L in S1 Text). Similar patterns can be observed across the three models. For instance, due to their frequent support and large transition rates, the West, Midwest, South, and Ohio Valley all appear to be important regions in the dispersal of AIV across the North American continent. Furthermore, while most decisively supported rates are between neighboring regions, longer distance transitions are also observed in all three models, including between the West and Alaska, the South and Guatemala, and the West and the Ohio Valley. Many supported rates also align with an East-West axis, suggesting viral exchange across migratory flyways.

Generalized linear model

The discrete trait diffusion models were extended with a GLM to evaluate the impact of host and geographic ecological characteristics on AIV dispersal among host species and geographic regions within North America. Table 1 summarizes host species and regional characteristics included in the GLM. Genetic distance of host species ranged widely, from 0.3 to 196.7 million years (Fig 3). Variables such as host genetic relatedness, habitat overlap, and geographic distance reflect the relationship between two variables, whereas the remaining ecological variables summarize aggregate measurements. For this reason, the relational variables were only included in the GLM once, but the remaining characteristics were each included twice to capture directionality of the viral transition rate. For instance, the average temperature during the summer months was included twice to assess if the summer temperature of the source region was associated with viral transition or if the summer temperature of the sink region impacted viral transition.

The host and region GLM models tested the same covariates across all gene segments and subtypes, individually. Overall, the internal gene segments held higher support for the inclusion of both host and region covariates as compared to the HA and NA subtype models (Fig 4). On average 20 of the 32 tested variables were supported for inclusion among the internal gene segments compared to five and seven supported variables among HA and NA subtypes, respectively. In the PDA sample, the H5 and N7 subtype models each supported only one variable across both host and region GLMs. Nonbreeding distribution overlap and migratory distance of the sink host species, as well as summer distribution overlap of the source region, winter temperature of both source and sink regions, winter precipitation of both source and sink regions, and winter humidity of source regions tended to have lower support among internal gene segment models. Overlap of host breeding distribution was supported across all gene segments and subtypes except H5 in both PDA and stratified samples and N6 in the PDA sample. Regional distance was also frequently supported across the gene segments and subtypes, with support in all but H10, N1, and N7 in both samples as well as H11 in the stratified sample. Summer temperature of the source region was supported in both samples for all

Table 1. Summary of host and geographic variables used to inform the Bayesian discrete diffusion generalized linear model describing avian influenza virus dispersal among North American wild birds.

	Mean	Standard Deviation	Range
Genetic Distance (Myr)	92.9	83.4	0.3–196.7
Summer Distribution Overlap (%)	26.6	32.0	0.0–99.99
Winter Distribution Overlap (%)	40.9	32.2	0.0–99.6
Breeding Latitude	56.4	11.7	28.0–76.7
Nonbreeding Latitude	34.1	8.3	21.6–56.6
Migration Distance (km)	2469.2	1447.9	473.8–5651.8
Migration Propensity (%)	84.4	22.1	18.1–100.0
Geographic Distance (km)	2716.9	1342.2	138.5–7087.5
Summer Proportion (%)	46.7	20.3	0.0–85.7
Winter Proportion (%)	48.1	28.2	0.0–90.5
Summer Temperature (C)	19.7	5.1	11.0–28.9
Winter Temperature (C)	-2.7	9.4	-15.6–19.7
Summer Precipitation (kg/m^2)	2.7	1.4	0.3–7.5
Winter Precipitation (kg/m^2)	1.9	0.9	0.4–3.4
Summer Humidity (%)	66.9	15.7	31.4–86.1
Winter Humidity (%)	78.1	10.6	47.4–86.3
Summer NDVI*	0.63	0.16	0.31–0.82
Winter NDVI*	0.29	0.16	0.02–0.75

*NDVI–Normalized difference vegetation index

https://doi.org/10.1371/journal.ppat.1009973.t001

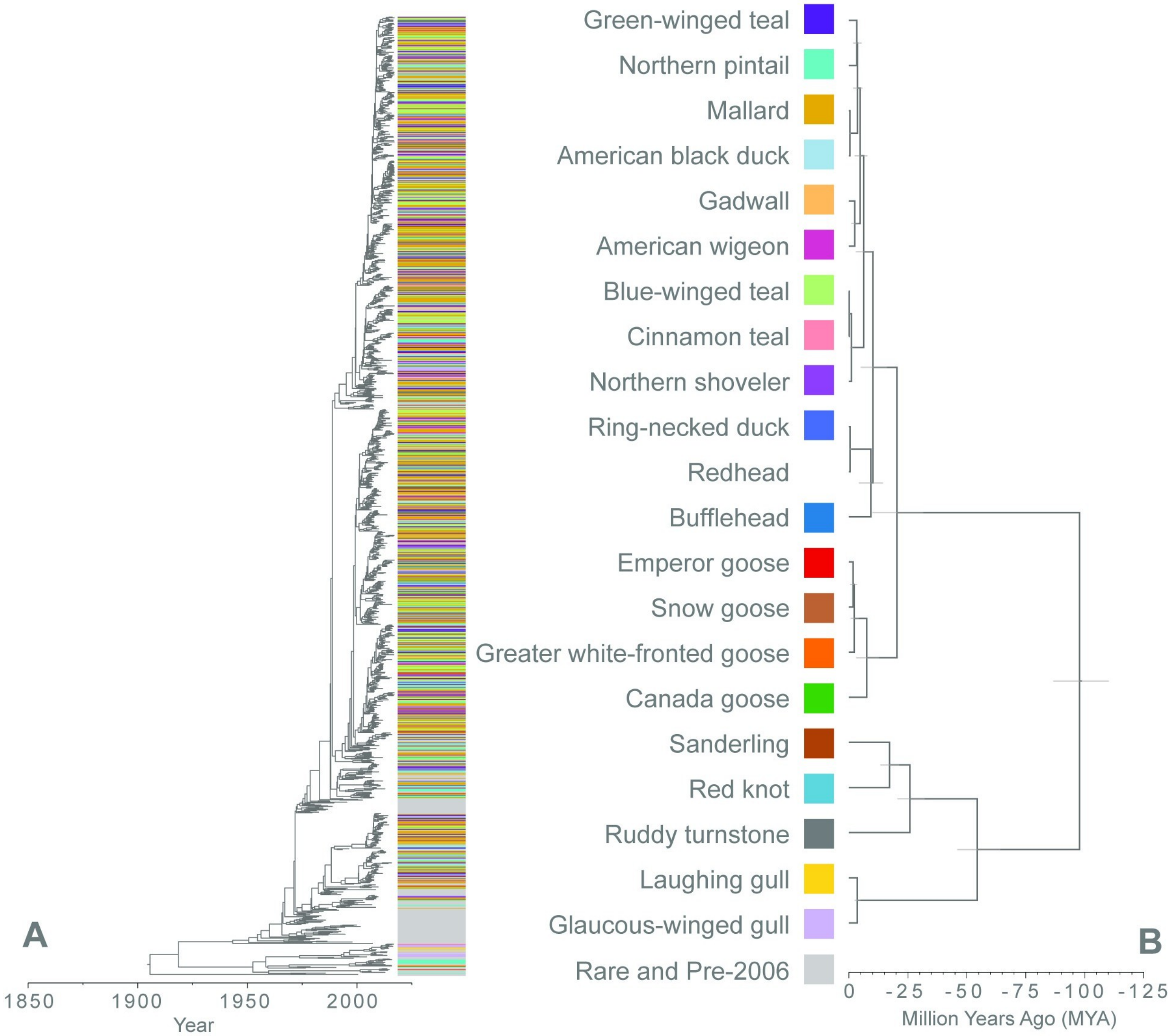

Fig 3. Viral and host phylogenetic diversity of North American AIV. (A) Estimation of the phylogenetic history of the PB2 AIV gene segment within North American wild birds. Color bands at the tips of the tree denote the host species distribution. This is contrasted with the phylogenetic history of the avian host species included in this analysis (B). Avian host phylogenetic history was summarized from 1,000 phylogenetic trees previously published by Jetz, et al. Light gray node bars represent the 95% highest posterior density of the node height. The redhead species was not categorized in the internal gene segment models and is therefore not included.

https://doi.org/10.1371/journal.ppat.1009973.g003

internal gene segments as well as subtypes H1, H3, H4, H11, N2, and N3. Host genetic relatedness was supported in all internal gene segments and all but two NA subtypes, N7 and N9, yet there was no support for this variable among HA subtypes. While non-ecological forces might explain lower support for GLM variables within the HA and NA gene segments compared to the internal genes, the smaller sample sizes of HA and NA data sets might also lead to less power to detect such associations.

The magnitude and direction of variable effect size differed among the various gene segments although most variables demonstrated the same directional effect across multiple gene segments. Among the 30 variables with support within two or more gene segments, 22 had the same directional effect (positive or negative) without regard to gene segment or subtype. Host variables which were consistently positively associated with interspecies transmission included

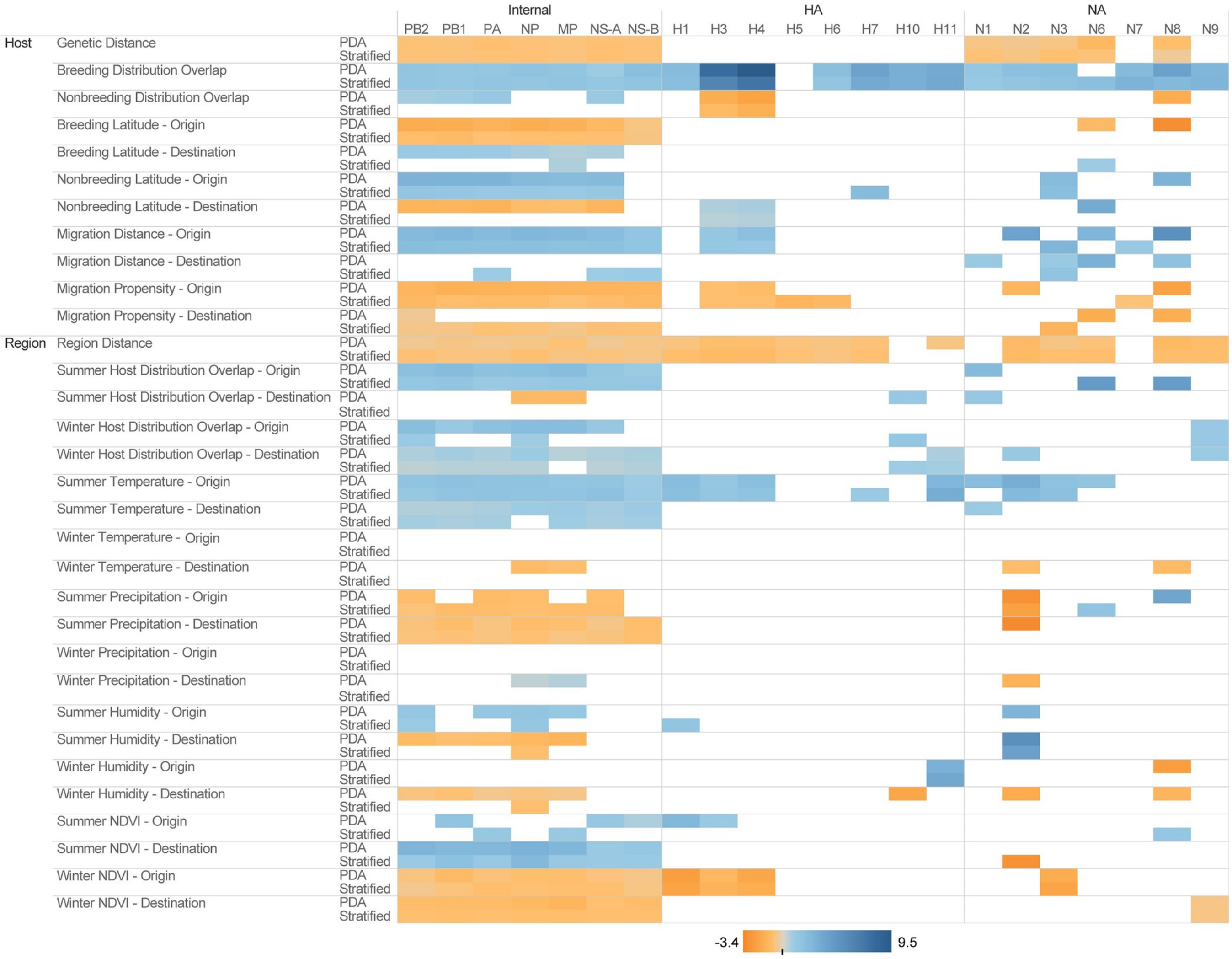

Fig 4. Heat map of conditional coefficient values for host and region generalized linear models of North American avian influenza discrete trait diffusion models. Conditional coefficient effect sizes are presented for each supported ecological variable across all gene segment and subtype datasets and both subsampling strategies (phylogenetic diversity analyzer (PDA) vs. stratified random sample). Only supported coefficients are displayed. Color darkness is proportional to the magnitude of the effect. Orange represents a negative correlation and blue represents a positive correlation.

https://doi.org/10.1371/journal.ppat.1009973.g004

breeding distribution overlap, nonbreeding latitude, and migration distance of the source host. Consistently negatively associated host characteristics included genetic distance between host species and migration propensity of both the source and sink host species. Among North American regional geographic models, the proportion of avian hosts in source regions with summer distribution overlap, the proportion of avian hosts in sink regions with winter distribution overlap, both source and sink summer temperature measures, and source summer normalized difference vegetation index (NDVI) were positively associated with viral dispersal across all segments and subtypes in which the variables had support for inclusion. The following region characteristics were consistently negatively associated with viral dispersal: distance between regions, summer precipitation of the sink region, and both winter NDVI measures. Eight covariates' directional effects conflicted among subtypes and gene segments, including nonbreeding distribution overlap, nonbreeding latitude of the sink host, summer habitat overlap in the sink region, summer precipitation of the source region, winter precipitation of the sink region, summer humidity of the sink region, winter humidity of the source region, and summer NDVI of the sink region. These conflicts tended to be observed in variables with infrequent support, especially within the stratified sample.

Discussion

The presented analysis provides insight into the potential impact of ecological variables that influence AIV dispersal and diversity within North American wild birds. While the evolution and dispersal of AIV within North America has been previously examined [7,8,15,23,24], this study employs a discrete trait diffusion GLM to incorporate ecological data into such estimates. Using new and historical AIV sequences, we demonstrate that host and geographic characteristics are associated with viral movement among avian species and North American regions. Because AIV gene segments can be treated as independent hereditary particles, genetic similarities can be used to infer information regarding the ecological pressures experienced by viral populations. By estimating ecological models separately for each AIV gene segment, dispersal patterns and their associations with ecological characteristics can be tested independently and compared. Consistent support for a variable across multiple gene segments and subtypes, such as breeding distribution overlap and geographic distance between regions, suggest that these host habitat characteristics play an important role in the evolution and ecology of AIV. Although AIV hosts often migrate and potentially carry virus over long distances, a geographic distance effect can be noted: as the distance between two regions increases, the frequency of AIV transition decreases. The importance of proximity is reinforced by the consistent support of distribution overlap, particularly in the summer breeding season. A similar finding has been observed in bats, in which viral transmission of rabies virus was associated with host distribution overlap in North America [20]. Species that have greater overlap during the breeding months tend to have a higher frequency of AIV transition due to a larger population of immunologically naïve juvenile hosts.

Another frequently supported host characteristic is the genetic distance or relatedness between two species, a characteristic that has been suggested to influence the rate of interspecies transmission of pathogens in general [25]. Genetic relatedness may be a proxy for a suite of shared characteristics that would increase the likelihood of two hosts sharing a pathogen. For instance, viruses that infect multiple species are most likely targeting conserved molecular mechanisms, and related hosts will most likely have similar physiological responses [26]. Furthermore, related species typically share similar ecology, i.e., breeding and feeding behavior or habitat, which can increase the likelihood of contact between the two species, a prerequisite for pathogen transmission. Experimental studies [27] and mathematical models [28] have shown that host relatedness is associated with successful host transition. Genetic distance, however, was not supported among any of the HA models. The HA models in general tended to have lower frequency of support for the included GLM models. This may suggest that the host immune pressure exerted on the HA supersedes influence of ecological determinants. In other words, because HA subtypes exist as a constellation of fitness peaks, these genes may be unable to provide information on ecological factors that affect viral transmission. Rather they are coerced by immune pressure to constantly accumulate mutations that provide fitness advantages to evade host immune systems. Lack of statistical support of ecological variables within the HA gene data sets should be interpreted with respect to their smaller sample sizes compared to the internal gene segment data. However, NA models had similar sample sizes to HA models and still showed a negative association in all NA subtypes except N7 (n = 262) and N9 (n = 237), the two smallest NA data sets.

Somewhat surprisingly, summer temperature of the originating region was positively associated with viral dispersal among regions in multiple gene segments. Environmental durability experiments [29,30] and AIV prevalence studies [11,12] have demonstrated evidence that colder temperatures increase risk of AIV infection due to environmental persistence of the virus. In contrast, our geographic model suggests that regions that are warmer on average during the summer are more likely to act as sources of the virus to other regions. It should be noted that causality cannot be established for this association. Proper interpretation of this result is that warmer regions are merely associated with viral dispersal, not that virus is more likely to arise from regions during summer. Summer temperature may be a proxy for other environmental, temporal or behavioral characteristics. The effect of temperature on AIV dispersal can also be observed in the host models in which latitude of the breeding distribution was negatively associated with viral transitions between host species. In other words, species that breed farther south were more likely to act as sources of AIV diversity to other host species. Importantly, the seasonal temperature peak in breeding zones coincides with the increase in juvenile Anseriformes [7]. Similarly, species that overwinter farther north were also more likely to act as sources of the virus. In corroboration with our model, one prevalence study [9] revealed an earlier thaw date of a location to be associated with higher AIV prevalence. Our results may be best explained by timing of breeding and migration rather than environmental persistence alone. Those locations that thaw first (i.e., are warmer in general) become available as breeding habitat sooner than regions farther north. Breeding marks the influx of new, immunologically naïve juveniles, populations that breed earlier tend to become infected earlier, which may increase their capacity to serve as a source of virus to other hosts and locations.

As with analyses reliant on publicly available data, these results are limited by potential sampling bias of available surveillance and sequence data. As demonstrated, mallards markedly dominate the diffusion models as sources of virus to other species. Mallards are the most populous of the dabbling ducks and therefore are more frequently included in AIV surveillance, but they are often also the species with the highest prevalence of AIV [31]. While one explanation for the estimation of mallards as frequent viral sources is their predominance in surveillance, the analysis methods were intended to limit the effects of sampling biases. Sequences collected prior to 2005 when sequencing efforts were irregular were not permitted to influence the discrete trait diffusion models. Further, by subsampling the datasets based on phylogenetics, we preserved the genetic diversity of the sequence data. The fact that mallards predominate in the PDA sample suggests that, as a primary reservoir species, mallards harbor a large diversity of AIV. A second subsampling technique (stratified random sample) was also performed in attempt to limit oversampling bias and increase the frequency of underrepresented hosts and regions. Both sampling strategies have limitations, however. Because the PDA method is blind to the sequences' location and host distribution, closely related viruses from different locations and hosts may have been removed from the analysis. Conversely, the stratified random sample ignores viral phylogenetic history, potentially resulting in loss of viral genetic diversity within the sample in favor of maintaining geographic and host diversity. By comparing results between the two datasets, the influence of sampling schemes on the observed results can be approximated, and the limitations of each strategy can be mitigated. Estimating the models across multiple gene segments and subtypes also allowed the host and regional proportions to vary, which is more apparent among the HA and NA subtypes. It should be noted that the magnitude of the effective population size across all segments tracked closely with the sample size of sequences included within the analysis. As the sample size was proportional to the number of available sequences that met inclusion criteria, the sample size may indicate the overall genetic diversity available for analysis, which would then be reflected in the estimated effective population size. This is supported by the use of phylogenetic diversity as a means to sub-sample the data, which ensures that both the full available sequences and the sampled sequences would have equivalent genetic diversity, unlike a simple random sample which, by chance, may remove some genetic diversity. In addition, the two sampling strategies produced data sets with differing host distributions. For these reasons, it is pertinent to consider multiple strategies when subsampling large data sets. Ideally, a scheme that takes both viral phylogenetic diversity and the discrete traits of interest should be used to insure neither type of diversity is lost.

While oversampling of specific hosts or locations can be adjusted for in the model, gaps in sampling are more difficult to address. Due to resource limitations, sampling of wild birds cannot occur in all hosts or locations where AIV infection occurs. Some species that may occur in multiple regions may only be sampled at a single location or region due to the convenience of congregation during migration or breeding. This is reflected in the data, for example, by wading bird species being collected exclusively in the Northeastern US. This might drive the limited viral transitions noted between Charadriiformes species and those of Anseriformes and thus overestimate the negative association between viral transition rates and genetic distance.

The GLM framework is also subject to several limitations. First, because the GLM attempts to explain variation among rates between discrete categories, the included covariates can only describe the categories themselves, not the individual viral isolates, hosts, or locations. Variables such as the exact temperature, humidity, or season at sample collection or age of the host might help explain associations observed in this analysis. For example, blue-winged teals are often sampled early in spring in the Southern US. By controlling for season and temperature at collection, the influence of blue-winged teals as a source of virus for mallards might be diminished. Similarly, host behavior variables could not be included in the geographic model and vice versa. This limits the ability to control for confounders such as the potential for breeding and migratory behavior to explain the positive association between viral transition and summer temperature. Second, the number of possible rates between discrete categories limits the number of covariates that can be estimated within the model with confidence. This results not only in the restriction of potential variables to be investigated, but the need to aggregate variables such as climatic measures across multiple years and large geographic areas. Finally, the GLM framework used BSSVS as a model selection tool to identify statistical significance of the investigated covariates. This method uses an indicator to include or remove variables within the model at each step of the Markov process. However, this prevents the evaluation of the impact of these variables. Shrinkage techniques are alternative statistical tools to evaluate covariates within a GLM that allow statistically insignificant covariates to shrink to zero, but still influence the model.

Although causality between ecologic factors and AIV diffusion cannot be inferred from this analysis, our results provide further evidence of the association of geographic and host characteristics with AIV diversity and dispersal. In general, dispersal among hosts of AIV genetic diversity appears driven by host genetic relatedness, breeding distribution overlap, and migratory behaviors. Geographic dispersal of AIV is strongly limited by physical distance, but further study into the importance of environmental and climatic factors as predictors of viral movement are still needed. Continued AIV surveillance, especially in undersampled regions and hosts, provides valuable information on AIV evolution and diffusion. Furthermore, the inclusion of detailed environmental and host measures within AIV sequence databases will help add granularity to future models.

Materials and methods

Ethics statement

All animal experiments were performed following Protocol Number 081 approved on August 19, 2011 by the St. Jude Children's Research Hospital Institutional Animal Care and Use

Committee in compliance with the Guide for the Care and Use of Laboratory Animals, 8th Ed. These guidelines were established by the Institute of Laboratory Animal Resources and approved by the Governing Board of the U.S. National Research Council.

Sample collections

Systematic avian influenza surveillance of wild birds has been performed in Alberta, Canada and Delaware Bay, New Jersey, United States since 1976 and 1985 respectively. Surveillance efforts, viral isolation, and genomic sequencing methods were performed as previously described [7]. In short, ducks were sampled post-breeding and prior to southern migration during July through early September at various wetlands in the prairie pothole regions Alberta. Sampling occurred during duck banding operations conducted by the Canadian Wildlife Service after ducks were captured in swim-in bait traps. Sample collections were permitted under agreement between St Jude Children's Research Hospital Center of Excellence for Influenza Research and Surveillance and the Canadian Wildlife Service. Samples were imported for analysis under the United States Veterinary Permit for Importation and Transportation of Controlled Materials and Organisms and Vectors (Permit No. 106760). Birds banded in Alberta have been recovered in all four North American flyways but most mallards are recovered in the Central and pacific flyways. Fecal samples from *Charadriiformes*–shorebirds and gulls—were collected in May at Delaware Bay. It is during this period in May that shorebirds (waders) are migrating north from South America to their breeding grounds in the Canadian Arctic. Delaware Bay serves as a stopover point where the birds can re-fuel on the abundance of eggs deposited by the coincident spawning of horseshoe crabs (*Limulus polyphemus*). Sample collections were permitted under agreement between St Jude Children's Research Hospital Center of Excellence for Influenza Research and Surveillance and the Delaware Bay Shorebird Project of the New Jersey Division of Fish and Wildlife, Endangered and Nongame Species Program.

Sequence data sets

Newly sequenced genomes from 303 viral isolates were deposited in GenBank (Table M in S1 Text) and were aligned using MUSCLE (v. 3.8) [32] with publicly available AIV sequences from within North America between 1970 and 2016, which were downloaded from the Influenza Research Database (IRD; www.fludb.org) on March 26, 2018. Gene segment sequences with vague host (e.g., "avian," "bird," "duck," etc.) or location (i.e., only country level data for the United States, Canada, or Mexico), more than 10% missing nucleotide sites within the coding region of the gene segment, or isolated from domestic poultry were removed from the dataset. Internal gene segments PB2, PB1, PA, NP, and MP were aligned separately. Alignments of gene segments NS, HA, and NA were further subdivided: NS by allele group (A and B) and HA and NA by subtype. HA and NA subtypes with sparse representation (<250 sequences between 2005 and 2016) were excluded from the analysis (subtypes H2, H8, H9, H12, H13, H14, H16, N4, N5). Initial maximum likelihood phylogenetic trees were estimated using RAxML v8 [33] with a general time reversible nucleotide substitution model and gamma distribution of sites. TempEst v1.5 [34] was used to identify sequences with a rate of evolutionary divergence out of the expected bounds as compared to the remaining sequences in the dataset. This helps to identify poor quality sequences or viruses under unexpected evolutionary pressure. Eurasian lineages with little continued North American circulation or associated with highly pathogenic avian influenza viruses were removed from the dataset. For each sequence, host of origin was categorized based on host species. Location of origin was categorized based on United States National Oceanic and Atmospheric Administration historical climate region for United States isolates, province and territory for Canadian isolates, state for

Mexican isolates, and country for Guatemalan isolates. Categories to be included in the discrete trait models were determined separately for the internal genes, HA subtypes, and NA subtypes. For internal genes, the 20 most common host species and regions (based on average rank among the segments) were chosen to be included in the discrete trait diffusion models. For both HA and NA subtypes, categories with greater than one sequence in a majority of the HA or NA subtypes were included in the models.

All sequences collected before 2005 were combined into a single category that was masked in the diffusion models to prevent the influence of inconsistent sampling and to focus diffusion summaries on the most recent years. Although the analytical period of interest was from 2005 to 2016, sequences collected before 2005 were retained in the data set to improve phylogenetic estimation and maintain information about event times that might affect more recent samples. To mitigate oversampling, two subsampling schemes were used: simple stratified random sampling and phylogenetic diversity-based sampling. In the simple random sample, sequences were stratified by region, host species, and year, and a maximum sample size of three sequences for each stratum were maintained in the dataset. Developed to help make economic decisions for conservation purposes, Phylogenetic Diversity Analyzer (PDA; http://www.cibiv.at/software/pda/) was used to select a subsample for each segment or subtype that maximized the represented genetic diversity [35]. This process was weighted to prevent over-representation of samples before 2005 which, though diverse, were masked in the diffusion model. As PDA allows the user to select the desired sample size, the number of selected sequences was specified to match the stratified sample and ensure datasets were proportional. Differences in host species and geographic region composition between the two sampling schemes were assessed with permutational multivariate analysis of variance (PERMANOVA) [36] using the vegan package (version 2.5, https://github.com/vegandevs/vegan) in R (version 3.5). A statistical difference between samples was defined as $p < 0.05$.

Phylogenetic analysis

Using ModelFinder algorithm [37] implemented in the program IQTree (http://www.iqtree.org/), the best fit nucleotide substitution model was determined. The empirical sets of phylogenetic trees were estimated under the same model assumptions for all sequence datasets in BEAST v1.10.4 [38]. A general time reversible (GTR) nucleotide substitution model [39–41] with a 4-category gamma distribution of variation among sites and a proportion of invariant sites [42,43] was implemented with a lognormal uncorrelated relaxed molecular clock [44] (mean clock rate prior distribution: uniform 0–1, initial value = 0.0033) and a constant coalescent population model [45,46] (population size prior distribution: lognormal distribution with mean = 50 and standard deviation = 50). At least four independent Markov chain Monte Carlo (MCMC) runs of 100 million state length and sampling every 10,000 states were performed. To ensure proper convergence and parameter mixing with an effective sample size (ESS) of at least 200, a minimum of 10% burn-in was removed. Non-convergent runs were discarded, larger burn-in percentages were removed, and additional MCMC runs were performed to achieve ESS > 200. Empirical tree sets were obtained by combining and resampling tree log files from non-discarded runs with LogCombiner to achieve a tree file length of at least 1,500 trees.

Discrete trait diffusion models

With the ability to incorporate ecological and epidemiological metadata, the discrete trait diffusion model uses a continuous-time Markov chain as its basis to estimate the ancestral history of trait changes across a phylogenetic tree, in essence treating the trait as a characteristic that evolves over time [47,48]. To investigate recent movement of AIV among avian hosts and North American regions, discrete trait diffusion models based on the empirical tree sets described above were estimated using BEAST v1.10.4. Posterior distributions of phylogenetic trees based on sequence data alone were estimated separately from the discrete trait diffusion models because the discrete trait model has an insignificant impact on phylogenetic estimation [17]. Furthermore, this approach enables the inference of a single diffusion model across multiple empirical tree sets, allowing the genetic information from multiple gene segments to inform the model. Due to the high level of reassortment of gene segments within low pathogenic AIV in wild birds [5], each gene segment can be treated as an independent hereditary particle, providing separate evolutionary and ecological information within its phylogenetic history [7]. Asymmetrical discrete trait diffusion models were estimated across empirical tree sets for the following: 1) each gene segment or subtype dataset individually, 2) all internal gene segments together, 3) all represented HA subtypes together, and 4) all represented NA subtypes together. Discrete host and geographic traits were specified as described above. Pre-2005 sequences and rare categories were masked from the discrete trait diffusion model, providing an estimate of viral transitions between common host species and regions between 2005 and 2016.

The discrete trait diffusion models were extended using a generalized linear model (GLM) to evaluate predictors associated with the discrete trait transition rates among host species and geographic regions. Using the transition rates as the outcome to a log-linear combination of covariate predictors, BEAST v1.10 estimates the GLM at each state in the MCMC simulation, integrating across the empirical phylogenetic tree space. Host diffusion predictors included genetic distance between species, habitat distribution overlap, migration distance, migration propensity, and latitudinal distribution. Host species genetic distances were calculated from a subsample of phylogenetic trees retrieved from www.birdtree.org, selecting the 21 bird species included in the discrete trait analysis and sampling 1000 trees from the "Stage 2 FP Trees Ericson" set [49]. These phylogenetic trees represent an extensive analysis into bird evolutionary history based on multiple gene loci, topological constraints, and fossil constraints. Patristic distances between species were calculated for each sampled tree using dendropy v.4.0 (www.dendropy.org) and then averaged across the 1000 tree sample [49]. Habitat overlap, migration distance, migration propensity, and latitudinal distribution were summarized from BirdLife species range maps [50] using ArcGIS Pro software. Habitat distribution overlap was calculated as the percentage of a source host's geographic distribution shared with that of a sink host. Migration distance was estimated by the difference between the mean breeding distribution latitude and the mean wintering distribution latitude [13]. Migration propensity was estimated as the percentage of total summer distribution range considered to be migratory as opposed to resident [13]. Latitudinal distribution was the average latitude for breeding and wintering ranges and served as an estimate of habitat temperature. Geographic diffusion predictors included distance between regions as well as summer and winter summaries of each of the following: average temperature, average precipitation, average relative humidity, average normalized difference vegetation index (NDVI), and proportion of included host species that reside in the region. All geographic variables were summarized between 2005 and 2016 and aggregated in ArcGIS Pro. Climatological data originated from the National Centers for Environmental Prediction North American Regional Reanalysis [51] provided by the National Oceanic and Atmospheric Administration Oceanic and Atmospheric Research Earth System Research Laboratory's Physical Sciences Division, Boulder, Colorado, USA, from their website at https://www.esrl.noaa.gov/psd/. NDVI data originated from the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13A3) Version 6 [52]. All covariates were log-transformed and standardized before inclusion in the GLM; therefore, a

GLM coefficient of 1.0 can be interpreted as an increase of one transition per year for every one unit increase in the log-transformed, standardized covariate. Each discrete trait diffusion model and GLM were performed with at least three independent MCMC runs of 1 million chain length sampling every 100 states.

Statistical analysis

For both the discrete trait diffusion model and the GLM, Bayesian stochastic search variable selection (BSSVS) was used to estimate the statistical support for diffusion rates and coefficients, respectively, by enabling the calculation of a Bayes factor (BF) [53]. A larger BF indicates stronger support that the parameter (i.e., transition rate between two hosts or GLM coefficient) is non-zero. Due to the number of models tested, a stringent statistical cutoff was implemented, only allowing those with BF > 100 to signify statistical support. Median conditional transition rates, median conditional coefficients, 95% highest posterior density (HPD) intervals, and BF were calculated from BEAST posterior samples with personalized Python scripts using the PyMC3 package for Bayesian statistical modeling [54].

Supporting information

S1 Text. Fig A. Evolutionary parameter estimation for North American avian influenza viruses of wild birds. Estimated parameters include A) time to most recent common ancestor (TMRCA), B) molecular clock rate, and C) effective population size. Parameters are compared across internal gene segments (blue), hemagglutinin gene subtypes (orange), and neuraminidase gene subtypes (purple) as well as between subsampling strategies, phylogenetic diversity-based sample (left, dark grey) and stratified random sample (right, light grey). Median values (black midline) indicated as well as the 95% highest posterior density (whiskers). Fig B. Host species temporal distribution of sampled North American avian influenza virus PB2 gene segment sequences, 2005–2016. Proportions of represented host species are compared between the original, unsampled data set (All), the phylogenetic diversity-based sample (PDA) and the stratified random sample (stratified). Fig C. Geographic region temporal distribution of sampled North American avian influenza virus PB2 gene segment sequences, 2005–2016. Proportions of represented host species are compared between the original, unsampled data set (All), the phylogenetic diversity-based sample (PDA) and the stratified random sample (stratified). Fig D. Host species and geographic region temporal distribution of sampled North American avian influenza virus PB2 gene segment sequences, 2005–2016. Proportions of represented host species are compared between the original, unsampled data set (All), the phylogenetic diversity-based sample (PDA) and the stratified random sample (stratified). Fig E. Heat map of supported viral transition rates among host species across avian influenza virus gene segments and subtypes. Colored cells represent the magnitude of the transition rate from the species in the first column (source) to the species in the second column (sink). White cells were transition rates that were not supported (Bayes factor < 100). Results from both subsampling strategies (phylogenetic diversity-based sample (PDA) and stratified random sample (stratified)) are presented for comparison. Fig F. Discrete trait diffusion models of North American avian influenza using a stratified random sample of genetic sequences. Host models (left) are presented for combined internal gene segment (A), hemagglutinin gene subtype (B), and neuraminidase gene subtype (C) models. Source host species on the left of the chord diagrams contribute viral diversity to sink host species on the right. The magnitude of the viral transition rate is proportional to the width of the band, and statistically supported rates darkened. Bands are colored by the host order of the source species (Charadriiformes–red; Anseriformes–blue). Similarly, geographic models (right) are summarized for combined internal gene segment (D), hemagglutinin gene subtype (E), and neuraminidase gene subtype (F) models. Arrow width is proportional to the magnitude of the transition rate, and only statistically supported rates are displayed. (AK–Alaska, AB–Alberta, BC–British Columbia, GT–Guatemala, MW–Midwest, NB–New Brunswick, NE–Northeast, NL–Newfoundland and Labrador, NRP–Northern Rockies and Plains, NS–Nova Scotia, NW–Northwest, OV–Ohio Valley, PE–Prince Edward Island, QC–Quebec, S–South, SE–Southeast, SON–Sonora, SW–Southwest, W–West). Public domain maps of United States, Canada, and Mexico states, and provinces were accessed from Wikimedia Commons, author Alex Covarrubias (https://commons.wikimedia.org/wiki/File:North_America_second_level_political_division.svg). Public domain map of Guatemala was accessed from *The World Factbook 2021*, Central Intelligence Agency (https://www.cia.gov/the-world-factbook/static/09f14262b8d528e66631646c85e0edc0/north_america_pol.pdf). Fig G. Heat map of supported viral transition rates among geographic regions across avian influenza virus gene segments and subtypes. Colored cells represent the magnitude of the transition rate from the region in the first column (source) to the region in the second column (sink). White cells were transition rates that were not supported (Bayes factor < 100). Results from both subsampling strategies (phylogenetic diversity-based sample (PDA) and stratified random sample (stratified)) are presented for comparison. Table A. Demographic characteristics of 303 wild bird surveillance samples with newly sequenced avian influenza isolates, 2003–2016. Table B. Evolutionary parameters of avian influenza virus gene segments collected from North American wild birds between 1970 and 2016. Datasets were sampled so as to maintain the total phylogenetic diversity of the original publicly available sequence sample. Table C. Host and regional distribution of phylogenetic diversity-based subsample of influenza virus gene segments isolated from North American wild birds. Table D. Host and regional distribution of stratified subsample of influenza virus gene segments isolated from North American wild birds. Table E. Cross-tabulation of host and regional distribution of influenza virus gene segments isolated from North American wild birds. Counts and proportions of represented host species by geographic location are compared between the original, unsampled data set (all), the phylogenetic diversity-based sample (PDA) and the stratified random sample (stratified). Table F. Hemagglutinin and neuraminidase subtype by host order. Counts across all segments and subtypes are compared between the original, unsampled data set (all), the phylogenetic diversity-based sample (PDA) and the stratified random sample (stratified). Table G. Host species transition rate matrix from combined internal gene model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically supported (Bayes factor > 100). (ABD–American black duck, BUF–bufflehead, BWT–blue-winged teal, CAN–Canada goose, CIN–cinnamon teal, EMP–emperor goose, GAD–gadwall, GWF–greater white-fronted goose, GWG–glaucous-winged gull, GWT–green-winged teal, LAU–laughing gull, MAL–mallard, PIN–northern pintail, RED–redhead, RKN–red knot, RND–ring-necked duck, RUD–ruddy turnstone, SHO–northern shoveler, SND–sanderling, SNO–snow goose, WIG–American wigeon). Table H. Host species transition rate matrix from combined hemagglutinin subtype model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically supported (Bayes factor > 100). (ABD–American black duck, BUF–bufflehead, BWT–blue-winged teal, CAN–Canada goose, CIN–cinnamon teal, EMP–emperor goose, GAD–gadwall, GWF–greater white-fronted goose, GWG–glaucous-winged gull, GWT–green-winged teal, LAU–laughing gull, MAL–mallard, PIN–northern pintail, RED–redhead, RKN–red knot, RND–ring-necked duck, RUD–ruddy turnstone, SHO–northern shoveler, SND–sanderling, SNO–snow goose, WIG–American wigeon). Table I. Host species transition rate matrix from combined neuraminidase subtype model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically

supported (Bayes factor > 100). (ABD–American black duck, BUF–bufflehead, BWT–blue-winged teal, CAN–Canada goose, CIN–cinnamon teal, EMP–emperor goose, GAD–gadwall, GWF–greater white-fronted goose, GWG–glaucous-winged gull, GWT–green-winged teal, LAU–laughing gull, MAL–mallard, PIN–northern pintail, RED–redhead, RKN–red knot, RND–ring-necked duck, RUD–ruddy turnstone, SHO–northern shoveler, SND–sanderling, SNO–snow goose, WIG–American wigeon). Table J. Geographic region transition rate matrix from combined internal gene model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically supported (Bayes factor > 100). (AK–Alaska, ALB–Alberta, BCO–British Columbia, GUA–Guatemala, MW–Midwest, NBR–New Brunswick, NE–Northeast, NFL–Newfoundland and Labrador, RP–Northern Rockies and Plains, NSC–Nova Scotia, NW–Northwest, OV–Ohio Valley, PEI–Prince Edward Island, QUE–Quebec, S–South, SE–Southeast, SON–Sonora, SW–Southwest, W–West). Table K. Geographic region transition rate matrix from combined hemagglutinin subtype model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically supported (Bayes factor > 100). (AK–Alaska, ALB–Alberta, BCO–British Columbia, GUA–Guatemala, MW–Midwest, NBR–New Brunswick, NE–Northeast, NFL–Newfoundland and Labrador, RP–Northern Rockies and Plains, NSC–Nova Scotia, NW–Northwest, OV–Ohio Valley, PEI–Prince Edward Island, QUE–Quebec, S–South, SE–Southeast, SON–Sonora, SW–Southwest, W–West). Table L. Geographic region transition rate matrix from combined neuraminidase subtype model. Median rates and 95% highest posterior density intervals are displayed for both subsampling strategies. Rates colored in blue are statistically supported (Bayes factor > 100). (AK–Alaska, ALB–Alberta, BCO–British Columbia, GUA–Guatemala, MW–Midwest, NBR–New Brunswick, NE–Northeast, NFL–Newfoundland and Labrador, RP–Northern Rockies and Plains, NSC–Nova Scotia, NW–Northwest, OV–Ohio Valley, PEI–Prince Edward Island, QUE–Quebec, S–South, SE–Southeast, SON–Sonora, SW–Southwest, W–West). Table M. Names and GenBank accession numbers of 303 newly sequenced AIV nucleotide sequences.
(PDF)

Acknowledgments

We would like to thank Karlie Woodard and Angela Danner for performing the RNA extraction as well as Nichola Hill for her insight into measures of avian phylogenetic diversity.

Author Contributions

Conceptualization: Joseph T. Hicks, Robert G. Webster, Justin Bahl.

Data curation: Kimberly Edwards, Do-Kyun Kim, Scott Krauss.

Formal analysis: Joseph T. Hicks.

Funding acquisition: Justin Bahl.

Investigation: Xueting Qiu, Scott Krauss, Justin Bahl.

Methodology: Xueting Qiu, Do-Kyun Kim, James E. Hixson, Scott Krauss, Richard J. Webby.

Resources: James E. Hixson, Scott Krauss, Richard J. Webby, Robert G. Webster, Justin Bahl.

Supervision: James E. Hixson, Richard J. Webby, Justin Bahl.

Validation: Kimberly Edwards, Do-Kyun Kim.

Visualization: Joseph T. Hicks.

Writing – original draft: Joseph T. Hicks.

Writing – review & editing: Joseph T. Hicks, Kimberly Edwards, Xueting Qiu, James E. Hixson, Scott Krauss, Richard J. Webby, Robert G. Webster, Justin Bahl.

References

1. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev [Internet]. 1992 Mar; 56(1):152–79. Available from: http://mmbr.asm.org/content/56/1/152.abstract https://doi.org/10.1128/mr.56.1.152-179.1992 PMID: 1579108
2. Webby RJ, Webster RG. Emergence of influenza A viruses. Philos Trans R Soc London Ser B Biol Sci [Internet]. 2001 Dec; 356(1416):1817–28. Available from: http://rstb.royalsocietypublishing.org/content/356/1416/1817.abstract https://doi.org/10.1098/rstb.2001.0997 PMID: 11779380
3. Dhingra MS, Artois J, Dellicour S, Lemey P, Dauphin G, Dobschuetz S Von, et al. Geographical and Historical Patterns in the Emergences of Novel Highly Pathogenic Avian Influenza (HPAI) H5 and H7 Viruses in Poultry. Front Vet Sci [Internet]. 2018; 5:84. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29922681 https://doi.org/10.3389/fvets.2018.00084 PMID: 29922681
4. Runstadler J, Hill N, Hussein IT, Puryear W, Keogh M. Connecting the study of wild influenza with the potential for pandemic disease. Infect Genet Evol. 2013 Jul; 17:162–87. https://doi.org/10.1016/j.meegid.2013.02.020 PMID: 23541413
5. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Ghedin E, et al. The Evolutionary Genetics and Emergence of Avian Influenza Viruses in Wild Birds. PLoS Pathog [Internet]. 2008 May; 4(5): e1000076. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18516303 https://doi.org/10.1371/journal.ppat.1000076 PMID: 18516303
6. Gultyaev AP, Fouchier RAM, Olsthoorn RCL. Influenza Virus RNA Structure: Unique and Common Features. Int Rev Immunol [Internet]. 2010 Nov; 29(6):533–56. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20923332 https://doi.org/10.3109/08830185.2010.507828 PMID: 20923332
7. Bahl J, Krauss S, Kühnert D, Fourment M, Raven G, Pryor SP, et al. Influenza a virus migration and persistence in North American wild birds. PLoS Pathog [Internet]. 2013; 9(8):e1003570. Available from: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003570#s4 https://doi.org/10.1371/journal.ppat.1003570 PMID: 24009503
8. Spackman E, Stallknecht DE, Slemons RD, Winker K, Suarez DL, Scott M, et al. Phylogenetic analyses of type A influenza genes in natural reservoir species in North America reveals genetic variation. Virus Res. 2005 Dec; 114(1–2):89–100. https://doi.org/10.1016/j.virusres.2005.05.013 PMID: 16039745
9. Fuller TL, Saatchi SS, Curd EE, Toffelmier E, Thomassen HA, Buermann W, et al. Mapping the risk of avian influenza in wild birds in the US [Internet]. Vol. 10, BMC Infectious Diseases. 2010. p. 187. Available from: https://www.ncbi.nlm.nih.gov/pubmed/20573228 https://doi.org/10.1186/1471-2334-10-187 PMID: 20573228
10. Belkhiria J, Alkhamis MA, Martínez-López B. Application of Species Distribution Modeling for Avian Influenza surveillance in the United States considering the North America Migratory Flyways [Internet]. Vol. 6, Scientific reports. 2016. p. 33161. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27624404 https://doi.org/10.1038/srep33161 PMID: 27624404
11. Herrick KA, Huettmann F, Lindgren MA. A global model of avian influenza prediction in wild birds: the importance of northern regions [Internet]. Vol. 44, Veterinary research. 2013. p. 42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23763792 https://doi.org/10.1186/1297-9716-44-42 PMID: 23763792
12. Farnsworth ML, Miller RS, Pedersen K, Lutman MW, Swafford SR, Riggs PD, et al. Environmental and demographic determinants of avian influenza viruses in waterfowl across the contiguous united states [Internet]. Vol. 7, PLoS ONE. 2012. p. e32729. Available from: https://www.ncbi.nlm.nih.gov/pubmed/22427870 https://doi.org/10.1371/journal.pone.0032729 PMID: 22427870
13. Garamszegi LZ, Møller AP. Prevalence of avian influenza and host ecology. Proc R Soc B Biol Sci [Internet]. 2007 Aug; 274(1621):2003–12. Available from: http://rspb.royalsocietypublishing.org/content/274/1621/2003.abstract?cited-by=yes&legid=royprsb;274/1621/2003 https://doi.org/10.1098/rspb.2007.0124 PMID: 17537707
14. Huang ZYX, Xu C, van Langevelde F, Ma Y, Langendoen T, Mundkur T, et al. Contrasting effects of host species and phylogenetic diversity on the occurrence of HPAI H5N1 in European wild birds. J Anim Ecol. 2019 Jul; 88(7):1044–53. https://doi.org/10.1111/1365-2656.12997 PMID: 31002194

15. Lam TT, Ip HS, Ghedin E, Wentworth DE, Halpin RA, Stockwell TB, et al. Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds. Ecol Lett [Internet]. 2012 Jan; 15(1):24–33. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2011.01703.x PMID: 22008513

16. 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. PLoS Pathog [Internet]. 2014; 10(2):e1003932. Available from: http://tmclibrary.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwtV1ba9swFBZpYdCXsfuyG3ov7mRLtuyHPWRhpRuluzRlezOSLbWDxQ65DLpfv3Mk2XFKB9vDSDBBsWVF57POp5NzIYQnRyy6sSYUGOEDul0wU1SxBrVZKM210oxzLVxc7XTCp9_Szx_k19Goq3C4bfuvgoc2ED0G0v6D8PtOoQE-AwTgCCCA41_BADil https://doi.org/10.1371/journal.ppat.1003932 PMID: 24586153

17. Baele G, A. MS, Rambaut A, Lemey P. Emerging Concepts of Data Integration in Pathogen Phylodynamics. Syst Biol. 2016; 66(1):e65.

18. Lu L, Brown AJL, Lycett SJ. Quantifying predictors for the spatial diffusion of avian influenza virus in China. BMC Evol Biol [Internet]. 2017; 17(1):16. Available from: https://doi.org/10.1186/s12862-016-0845-3 PMID: 28086751

19. Wei K, Li Y. Global genetic variation and transmission dynamics of H9N2 avian influenza virus. Transbound Emerg Dis. 2018 Apr; 65(2):504–17. https://doi.org/10.1111/tbed.12733 PMID: 29086491

20. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints [Internet]. Vol. 368, Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences. 2013. p. 20120196. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23382420 https://doi.org/10.1098/rstb.2012.0196 PMID: 23382420

21. Breban R, Drake JM, Stallknecht DE, Rohani P. The Role of Environmental Transmission in Recurrent Avian Influenza Epidemics. Fraser C, editor. PLoS Comput Biol [Internet]. 2009 Apr 10 [cited 2019 Apr 18]; 5(4):e1000346. Available from: https://doi.org/10.1371/journal.pcbi.1000346 PMID: 19360126

22. Roche B, Drake JM, Brown J, Stallknecht DE, Bedford T, Rohani P. Adaptive Evolution and Environmental Durability Jointly Structure Phylodynamic Patterns in Avian Influenza Viruses. Fraser C, editor. PLoS Biol [Internet]. 2014 Aug 12 [cited 2019 Apr 18]; 12(8):e1001931. Available from: https://doi.org/10.1371/journal.pbio.1001931 PMID: 25116957

23. Bahl J, Vijaykrishna D, Holmes EC, Smith GJD, Guan Y. Gene flow and competitive exclusion of avian influenza A virus in natural reservoir hosts. Virology [Internet]. 2009; 390(2):289–97. Available from: https://www.sciencedirect.com/science/article/pii/S0042682209002876?via%3Dihub https://doi.org/10.1016/j.virol.2009.05.002 PMID: 19501380

24. Fourment M, Darling AE, Holmes EC. The impact of migratory flyways on the spread of avian influenza virus in North America. BMC Evol Biol. 2017; 17(1). https://doi.org/10.1186/s12862-017-0965-4 PMID: 28545432

25. Woolhouse MEJ. Population biology of emerging and re-emerging pathogens. Trends Microbiol [Internet]. 2002; 10(10):s7. Available from: https://www.sciencedirect.com/science/article/pii/S0966842X02024289 https://doi.org/10.1016/s0966-842x(02)02428-9 PMID: 12377561

26. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM. The evolution and genetics of virus host shifts [Internet]. Vol. 10, PLoS pathogens. 2014. p. e1004395. Available from: https://www.ncbi.nlm.nih.gov/pubmed/25375777 https://doi.org/10.1371/journal.ppat.1004395 PMID: 25375777

27. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. Host phylogeny determines viral persistence and replication in novel hosts [Internet]. Vol. 7, PLoS Pathogens. 2011. p. e1002260. Available from: https://www.ncbi.nlm.nih.gov/pubmed/21966271

28. Cuthill JH, Charleston MA. A SIMPLE MODEL EXPLAINS THE DYNAMICS OF PREFERENTIAL HOST SWITCHING AMONG MAMMAL RNA VIRUSES. Evolution (N Y) [Internet]. 2013 Apr; 67(4):980–90. Available from: https://www.jstor.org/stable/23463853 https://doi.org/10.1111/evo.12064 PMID: 23550750

29. Keeler SP, Dalton MS, Cressler AM, Berghaus RD, Stallknecht DE. Abiotic factors affecting the persistence of avian influenza virus in surface waters of waterfowl habitats [Internet]. Vol. 80, Applied and Environmental Microbiology. 2014. p. 2910–7. Available from: https://www.ncbi.nlm.nih.gov/pubmed/24584247 https://doi.org/10.1128/AEM.03790-13 PMID: 24584247

30. Dalziel AE, Delean S, Heinrich S, Cassey P. Persistence of low pathogenic influenza A virus in water: a systematic review and quantitative meta-analysis [Internet]. Vol. 11, PLoS One. 2016. p. e0161929. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27736884 https://doi.org/10.1371/journal.pone.0161929 PMID: 27736884

31. Bevins SN, Pedersen K, Lutman MW, Baroch JA, Schmit BS, Kohler D, et al. Large-Scale Avian Influenza Surveillance in Wild Birds throughout the United States. Yoon K-J, editor. PLoS One [Internet].

2014 Aug 12 [cited 2019 Apr 18]; 9(8):e104360. Available from: http://dx.plos.org/10.1371/journal.pone.0104360 PMID: 25116079

32. RE C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5):1792–7. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

33. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

34. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol [Internet]. 2016; 2(1):vew007. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27774300 https://doi.org/10.1093/ve/vew007 PMID: 27774300

35. Chernomor O, Minh BQ, Forest F, Klaere S, Ingram T, Henzinger M, et al. Split diversity in constrained conservation prioritization using integer linear programming. Methods Ecol Evol [Internet]. 2015; 6 (1):83–91. Available from: https://doi.org/10.1111/2041-210X.12299 PMID: 25893087

36. Anderson MJ. Permutational Multivariate Analysis of Variance (PERMANOVA) [Internet]. Wiley StatsRef: Statistics Reference Online. 2017. p. 1–15. (Major Reference Works). Available from: https://doi.org/10.1002/9781118445112.stat07841

37. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun; 14(6):587–9. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

38. 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; 4(1):vey016. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29942656 https://doi.org/10.1093/ve/vey016 PMID: 29942656

39. Tavaré S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. In Amer Mathematical Society; 1986. p. 57–86. (American Mathematical Society: Lectures on Mathematics in the Life Sciences; vol. 17). Available from: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0821811673

40. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. J Mol Evol. 1984; 20(1):86–93. https://doi.org/10.1007/BF02101990 PMID: 6429346

41. Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. J Theor Biol [Internet]. 1990; 142(4):485–501. Available from: https://www.sciencedirect.com/science/article/pii/S0022519305801043 https://doi.org/10.1016/s0022-5193(05)80104-3 PMID: 2338834

42. Uzzell T, Corbin KW. Fitting Discrete Probability Distributions to Evolutionary Events. Science (80-) [Internet]. 1971 Jun; 172(3988):1089–96. Available from: http://www.sciencemag.org/cgi/content/abstract/172/3988/1089 https://doi.org/10.1126/science.172.3988.1089 PMID: 5574514

43. Jin L, Nei M. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol [Internet]. 1990 Jan; 7(1):82. Available from: https://www.ncbi.nlm.nih.gov/pubmed/2299983 https://doi.org/10.1093/oxfordjournals.molbev.a040588 PMID: 2299983

44. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence [Internet]. Vol. 4, PLoS Biology. 2006. p. e88. Available from: https://www.openaire.eu/search/publication?articleId=dedup_wf_001::2701d63ad55f5ea9e00b918241d07440 https://doi.org/10.1371/journal.pbio.0040088 PMID: 16683862

45. Kingman JFC. On the Genealogy of Large Populations. J Appl Probab [Internet]. 1982 Jan; 19(A):27–43. Available from: http://www.jstor.org/stable/3213548

46. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. Genetics [Internet]. 2002; 161(3):1307–20. Available from: http://tmclibrary.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV1NS_QwEB5UELyIr5-rr5CTJ3fpZ5IePMiiiDdB0Vto86HCtrvY7mH99c6kqbwrnt5j27S0mcxknuTpMwBpMonGP2JCjlhIW5zbrEuESbRx2qtNSWe4kYZw4_Q6nb7kD_fieQMGFhmRLNeoipPm_c3TLUO_fvh6axNHuJFOmSv0fJy15EU3n8-uFrW_ https://doi.org/10.1093/genetics/161.3.1307 PMID: 12136032

47. Minin VN, Suchard MA. Fast, accurate and simulation-free stochastic mapping. Philos Trans R Soc B Biol Sci [Internet]. 2008 Dec; 363(1512):3985–95. Available from: http://rstb.royalsocietypublishing.org/content/363/1512/3985.abstract https://doi.org/10.1098/rstb.2008.0176 PMID: 18852111

48. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. J Math Biol [Internet]. 2007; 56(3):391–412. Available from: http://tmclibrary.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwlR1ZS_QwcBBB0Afvox4QfFOoNJsmbR5lUUTwQTy-7y20OUA-7Iq7K_jvnenluvrw-dTCpCXJTGYyN4AYnCXxHE_IUWvmNkt9qvEpi8CDSkpOhShlSGq_-_BcDP_K2-

vszwIMektG9e-sc1DWfLtPfSNlQca1oY2jBkTpviQrSV-_e-z9CJlseuYho46p https://doi.org/10.1007/s00285-007-0120-8 PMID: 17874105

49. Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. The global diversity of birds in space and time. Nature [Internet]. 2012 Nov [cited 2019 Mar 4]; 491(7424):444–8. Available from: http://www.nature.com/articles/nature11631 https://doi.org/10.1038/nature11631 PMID: 23123857

50. BirdLife International and Handbook of the Birds of the World. Bird species distribution maps of the world. Version 6.0. [Internet]. Available from: http://datazone.birdlife.org/species/requestdis

51. Mesinger F, DiMego G, Kalnay E, Mitchell K, Shafran PC, Ebisuzaki W, et al. NORTH AMERICAN REGIONAL REANALYSIS [Internet]. Vol. 87, Bulletin of the American Meteorological Society. American Meteorological Society; 2006 [cited 2019 May 9]. p. 343–60. Available from: https://www.jstor.org/stable/26217151

52. Didan K. MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC; 2015.

53. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol [Internet]. 2009 Sep; 5(9):e1000520. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19779555 https://doi.org/10.1371/journal.pcbi.1000520 PMID: 19779555

54. Salvatier J, Wiecki T V., Fonnesbeck C. Probabilistic programming in Python using PyMC3. PeerJ Comput Sci [Internet]. 2016 Apr 6 [cited 2019 Mar 4]; 2:e55. Available from: https://peerj.com/articles/cs-55