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

Short Title: Ecology and Evolution of Avian Influenza A virus in North American wild birds

AUTHORS: Joseph T. Hicks[1], Kimberly Friedman[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[1,4]

AFFILIATIONS: [1]Center for Ecology of Infectious Diseases, Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, Georgia; [2]Department of Infectious Disease, St. Jude Children's Research Hospital, Memphis, Tennessee; [3]University of Texas Health Science Center at Houston School of Public Health, Houston, Texas; [4]Duke-NUS Graduate Medical School, Singapore.

CORRESPONDING AUTHOR: Justin Bahl justin.bahl@uga.edu

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 transmission and geographic diffusion patterns among multiple co-circulating subtypes. Our results show high transition rates within Anseriformes and Charadriiformes, but limited transitions between these orders. Patterns of interspecies transmission 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 diffusion. 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.

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 devastating loss (3). Furthermore, all modern pandemic influenza 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). AIV segments are habitually interchanging, existing as functionally equivalent arrangements (5). 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). 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 (Figure 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 (14).

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 (15,16). This allows the simultaneous 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 (17) as well as of avian influenza subtype H9N2 on a global scale (18). The GLM has also been used to uncover the impact of host behavior on the dispersal of rabies virus among bat species (19). Previous analyses have also demonstrated the importance of environmental transmission on AIV prevalence and evolution (20,21), 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 Newly Sequenced Data

Supplementary Table S1 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 almost exclusively 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. Because the PDA sample maintains the total genetic diversity of the original sample, the evolutionary parameters of the PDA sample are discussed here. 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 (Supplementary Figure S1). Compared with the HA and NA surface proteins, the internal gene segments tended to have older times to the most recent common ancestor (TMRCA) (Supplementary Figure S1A; Supplementary Table S2), 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). HA and NA genes tended to have a TMRCA within the mid- to latter-half of the twentieth century, although this pattern deviated for H3 and N3, which had TMRCA older than other HA and NA subtypes (H3: 1929.0, 95% HPD 1900.0 – 1948.5; N3: 1895.6, 95% HPD 1830.8 – 1944.7). The uncertainty of the N3 subtype TMRCA is most likely due to the lack of older sequences which would help calibrate the divergence time of apparent Eurasian-origin viruses that circulate in North America.

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 (Supplementary Figure S1B; Supplementary Table S2). The two NS alleles had differing substitution rates with the A allele evolving faster (3.4×10^{-3} substitutions/site/year; 95% HPD 3.0×10^{-3} – 3.7×10^{-3}) compared to the

B allele ($2.6x10^{-3}$ substitutions/site/year; 95% HPD $2.3x10^{-3}$ – $2.9x10^{-3}$). In comparison to the internal gene segments, the HA and NA surface proteins were estimated to have more variable substitution rates with median rates ranging from $2.5x10^{-3}$ (H3) to $5.8x10^{-3}$ (H7) substitutions/site/year. Six of the eight analyzed HA subtypes had estimated median substitution rates above $3.8x10^{-3}$ substitutions/site/year. In contrast, H3 and H4 were estimated to have much slower substitution rates at $2.5x10^{-3}$ (95% HPD $2.2x10^{-3}$ – $2.7x10^{-3}$) and $2.8x10^{-3}$ (95% HPD $2.6x10^{-3}$ – $3.1x10^{-3}$) substitutions/site/year, respectively. The NA subtype with the fastest substitution rate was N7 ($5.0x10^{-3}$ substitutions/site/year; 95% HPD $4.4x10^{-3}$ – $5.6x10^{-3}$).

Internal gene segments were estimated to be sustained by a much larger effective population size as compared to the surface proteins (Supplementary Figure S1C; Supplementary Table S2). 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. Because the genetic diversity of the NS, HA, and NA gene segments was divided between datasets, lower population sizes are needed to explain the observed viral circulation. Variation among HA and NA subtypes was also noted. Most HA subtypes were estimated to have median effective population sizes below 20, but those of H3 and H4 were substantially larger (H3: 99.3, 95% HPD 87.2 – 112.5; H4: 78.0, 95% HPD 68.3 – 88.2). The effective population sizes of the NA subtypes also varied considerably with median sizes ranging from 8.5 (N7) to 85 (N8).

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

146 North America. North American regions were categorized into eight Canadian provinces and
147 territories (Alberta, British Columbia, New Brunswick, Newfoundland and Labrador, Nova
148 Scotia, Ontario, Prince Edward Island, and Quebec), ten United States climate regions (Alaska,
149 Midwest, Northeast, Northwest, Ohio Valley, Northern Rockies and Plains, South, Southeast,
150 Southwest, and West), one Mexican state (Sonora), and Guatemala. Represented host species
151 were of the taxonomic orders Anseriformes (waterfowl) and Charadriiformes (shorebirds). The
152 16 Anseriformes species included American black duck (*Anas rubripes*), bufflehead (*Bucephala*
153 *albeola*), blue-winged teal (*Anas discors*), Canada goose (*Branta canadensis*), cinnamon teal
154 (*Anas cyanoptera*), emperor goose (*Anser canagicus*), gadwall (*Mareca strepera*), greater white-
155 fronted goose (*Anser albifrons*), green-winged teal (*Anas crecca*), mallard (*Anas platyrhynchos*),
156 northern pintail (*Anas acuta*), redhead (*Aythya americana*), ring-necked duck (*Aythya collaris*),
157 northern shoveler (*Anas clypeata*), snow goose (*Anser caerulescens*), and American wigeon
158 (*Anas americana*). Five species of Charadriiformes were represented among the host models:
159 glaucous-winged gull (*Larus glaucescens*), laughing gull (*Leucophaeus atricilla*), red knot
160 (*Calidris canutus*), ruddy turnstone (*Arenaria interpres*), and sanderling (*Calidris alba*).
161 Henceforth, all host species will be referenced using common names.

162 The distribution of hosts and geographic regions were similar among the internal gene
163 segments by both PDA and stratified subsampling strategies (Supplementary Table S3;
164 Supplementary Table S4). Subsampling strategy had an effect on the temporal distribution of
165 host and geographic region variables; proportions were more consistent between 2005 and 2016
166 in the stratified sample compared with the PDA sample (Supplementary Figures S2 & S3).
167 Within both the PDA and stratified samples, Alaska, the Ohio Valley, and the Northeast were the
168 most frequently represented regions between 2005 and 2016 among the internal genes. Regional

distribution varied considerably across HA and NA subtypes. Within the PDA sample, the most frequently represented regions' HA subtypes included Alaska (H3: 30.5%, H4: 20.5%), Northeast (H1: 20.7%, H10: 31.1%, H11: 20.0%), South (H7: 23.7%), and West (H5: 20.5%, H6: 28.1%, H11: 20.0%). The most frequently represented regions among NA subtypes in the PDA sample included Alaska (N3: 18.3%, N6: 25.9%, N8: 24.5%), Midwest (N2: 20.3%, N9: 20.3%), and Northeast (N1: 19.2%, N7: 25.2%). 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). Because the rates of transition between two locations, for instance, are asymmetrical, the transition rate from location A to location B is permitted to differ from the rate in the opposing direction (from location B to A). Across all gene segments, the highest rate of transition between host species was 19.6 transitions/year (95% HPD 16.0 – 23.6) of the PB1 gene segment from mallards to blue-winged teals within the PDA sample (Supplemental Figure S4). In the PDA sample, the N8 transition rate from mallards to blue-winged teals was highest among the NA subtypes at 13.3 transitions/year (95% HPD 9.3 – 17.7). In contrast, among HA subtypes, the reverse transition (that is, blue-winged teals to mallards) within the PDA H4 model had the highest rate (16.1 transitions/year; 95% HPD 10.9 – 21.4). Mallards were supported as the source of AIV across all gene segments and subtypes for green-winged teals and northern shovelers

192 within the stratified sample. These rates were also supported in the PDA sample in all gene
193 segments and subtypes except N6. The rates from mallards to blue-winged teals were also
194 supported among all gene segments and subtypes except H7 in both samples. Within the PDA
195 sample, American black ducks, blue-winged teals, Canada geese, greater white-fronted geese,
196 ring-necked ducks and snow geese were only supported to receive viral diversity from mallards.
197 In contrast, only American black ducks and Canada geese exclusively received virus from
198 mallards in the stratified sample. Similarly, ruddy turnstones were the exclusive source of viral
199 diversity for laughing gulls and sanderlings in both samples, as well as red knots in the stratified
200 sample.

201 A single host diffusion model was also jointly estimated across all internal gene
202 segments, all HA subtypes, and all NA subtypes (Figure 2, Supplementary Figure S5). For each
203 joint host model, the highest transition rate occurred from blue-winged teals to mallards (PDA
204 internal gene model: 40.7 transitions/year, 95% HPD 32.5 – 49.2; PDA HA model: 22.0
205 transitions/year, 95% HPD 16.6 – 27.5; Stratified NA model: 22.6 transitions/year, 95% HPD
206 14.2 – 31.7; Tables S5, S6 & S7). In all three joint host models, all species were supported as
207 receiving virus from at least one other species, except snow geese within the NA models. Not all
208 species acted as a source, however. Cinnamon teals, gadwalls, and red knots were included in all
209 three joint host models, but none were supported to contribute AIV genetic diversity to any other
210 host species. In addition, Canada geese, emperor geese, ring-necked ducks, snow geese, laughing
211 gulls, and sanderlings, which were only included in the internal gene and NA models, were also
212 not supported as viral sources. A marked difference between the PDA and stratified samples can
213 be noted in this regard. Whereas green-winged teals were not supported as a source of virus for
214 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, the highest transition rate was observed from the Ohio Valley to the South within the PA gene segment of the PDA sample at a rate of 23.3 transitions/year (95% HPD 17.6 – 29.0; BF = 73,262) (Figure S6). In the PDA sample, the N6 model showed the highest transition rate of the NA subtypes with a Midwest to South transition rate of 11.4 transitions/year (95% HPD 7.5 – 15.7; BF = 46,210). Within the stratified sample, the highest transition rate among HA subtypes occurred from the Midwest into the Ohio Valley in the H4 model at 14.9 transitions/year (95% HPD 9.9 – 20.1, BF = 46,210). No single transition rate was supported across all gene segments or subtypes. 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 models, the internal gene model has the largest number of decisively supported transition rates between regions (Figure 2, Supplementary Figure S5). 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 S8). 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 S9 & S10). 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 (Figure 3). As expected, the host phylogeny is composed of two main clades, Anseriformes and Charadriiformes, which diverged between 86.6 and 110.2 million years ago. Included Charadriiformes species shared a most recent common ancestor between 46.0 and 64.2 million years ago, whereas Anseriformes species diverged into familial clades of Anatidae (ducks) and Anser (geese) between 10.0 and 31.4 million years ago. All nodes within the

phylogeny were highly supported (posterior probability > 0.98) except for the ancestral node of the green-winged teal and Northern pintail (posterior probability = 0.67). Host habitat distributions tended to have greater percentage of overlap during the nonbreeding winter months versus the breeding summer months. Migration propensity and distance were calculated based on distribution maps and varied widely among included species. On average, 84% of a species' breeding season distribution was considered migratory as opposed to resident. The average distance between North American geographic regions was 2,716 km, and geographic regions overlapped with 47% and 48% of the breeding and nonbreeding distributions, respectively. Regions tended to have greater precipitation, lower humidity, and more lush vegetation during the summer months compared to the winter months. 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 (Figure 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

284 one variable across both host and region GLMs. Nonbreeding distribution overlap and migratory
285 distance of the sink host species, as well as summer distribution overlap of the source region,
286 winter temperature of both source and sink regions, winter precipitation of both source and sink
287 regions, and winter humidity of source regions tended to have lower support among internal gene
288 segment models. Overlap of host breeding distribution was supported across all gene segments
289 and subtypes except H5 in both PDA and stratified samples and N6 in the PDA sample. Regional
290 distance was also frequently supported across the gene segments and subtypes, with support in
291 all but H10, N1, and N7 in both samples as well as H11 in the stratified sample. Summer
292 temperature of the source region was supported in both samples for all internal gene segments as
293 well as subtypes H1, H3, H4, H11, N2, and N3. Host genetic relatedness was supported in all
294 internal gene segments and all but two NA subtypes, N7 and N9, yet there was no support for
295 this variable among HA subtypes.

296 The magnitude and direction of variable effect size differed among the various gene
297 segments although most variables demonstrated the same directional effect across multiple gene
298 segments. Among the variables supported in the models, 22 had the same directional effect
299 (positive or negative) without regard to gene segment or subtype. Host variables which were
300 consistently positively-associated with interspecies transmission included breeding distribution
301 overlap, nonbreeding latitude and migration distance of the source host. Consistent negatively-
302 associated host characteristics included genetic distance between host species and migration
303 propensity of both the source and sink host species. Among North American regional geographic
304 models, the proportion of avian hosts in source regions with summer distribution overlap , the
305 proportion of avian hosts in sink regions with winter distribution overlap , both source and sink
306 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. Several 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 (7,8,14,22,23), 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 (19). 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 (24). 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 (25). 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 (26) and mathematical models (27) 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

353 coerced by immune pressure to constantly accumulate mutations that provide fitness advantages
354 to evade host immune systems.

355 Somewhat surprisingly, summer temperature of the originating region was positively
356 associated with viral dispersal among regions in multiple gene segments. Environmental
357 durability experiments (28,29) and AIV prevalence studies (11,12) have demonstrated evidence
358 that colder temperatures increase risk of AIV infection due to environmental persistence of the
359 virus. In contrast, our geographic model suggests that regions that are warmer on average during
360 the summer are more likely to act as sources of the virus to other regions. It should be noted that
361 causality cannot be established for this association. Proper interpretation of this result is that
362 warmer regions are merely associated with viral dispersal, not that virus is more likely to arise
363 from regions during summer. Summer temperature may be a proxy for other environmental or
364 temporal characteristics. The effect of temperature on AIV dispersal can also be observed in the
365 host models in which latitude of the breeding distribution was negatively associated with viral
366 transitions between host species. In other words, species that breed farther south were more
367 likely to act as sources of AIV diversity to other host species. Similarly, species that overwinter
368 farther north were also more likely to act as sources of the virus. In corroboration with our
369 model, one prevalence study (9) revealed an earlier thaw date of a location to be associated with
370 higher AIV prevalence. Our results may be best explained by timing of breeding and migration
371 rather than environmental persistence alone. Those locations that thaw first (i.e., are warmer in
372 general) become available as breeding habitat sooner than regions farther north. Because
373 breeding marks the influx of new, immunologically naïve juveniles, populations that breed
374 earlier tend to become infected earlier, which may increase their capacity to serve as a source of
375 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 (30). 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. By comparing results between the two datasets, the influence of sampling schemes on the observed results can be approximated. 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

399 genetic diversity, unlike a simple random sample which, by chance, may remove some genetic
400 diversity.

401 Although causality between ecologic factors and AIV diffusion cannot be inferred from
402 this analysis, our results provide further evidence of the association of geographic and host
403 characteristics with AIV diversity and dispersal. Continued AIV surveillance, especially in
404 undersampled regions and hosts, provides valuable information on AIV evolution and diffusion.
405 Furthermore, the inclusion of detailed environmental and host measures within AIV sequence
406 databases will help add granularity to future models.

407 MATERIALS AND METHODS

408 *Sequence data sets*

409 Systematic avian influenza surveillance of wild birds has been performed in Alberta,
410 Canada and Delaware Bay, New Jersey, United States since 1976 and 1985 respectively.
411 Surveillance efforts, viral isolation, and genomic sequencing methods were performed as
412 previously described (7). Newly sequenced genomes from 303 viral isolates were deposited in
413 GenBank (Supplementary Table S11) and were aligned with publicly available AIV sequences
414 from within North America between 1970 and 2016, which were downloaded from the Influenza
415 Research Database (IRD; www.fludb.org) on March 26, 2018. Sequences with vague host (e.g.,
416 “avian,” “bird,” “duck,” etc.) or location (i.e., only country level data for the United States,
417 Canada, or Mexico), more than 10% missing nucleotide sites, or isolated from domestic poultry
418 were removed from the dataset. Internal gene segments PB2, PB1, PA, NP, and MP were aligned
419 separately. Alignments of gene segments NS, HA, and NA were further subdivided: NS by allele
420 group (A and B) and HA and NA by subtype. HA and NA subtypes with sparse representation
421 (<250 sequences between 2005 and 2016) were excluded from the analysis (subtypes H2, H8,

422 H9, H12, H13, H14, H16, N4, N5). Initial maximum likelihood phylogenetic trees were
423 estimated using RAxML v8 (31) with a general time reversible nucleotide substitution model and
424 gamma distribution of sites. TempEst v1.5 (32) was used to identify sequences with a rate of
425 evolutionary divergence out of the expected bounds as compared to the remaining sequences in
426 the dataset. This helps to identify poor quality sequences or viruses under unexpected
427 evolutionary pressure. Eurasian lineages with little continued North American circulation or
428 associated with highly pathogenic avian influenza viruses were removed from the dataset. For
429 each sequence, host of origin was categorized based on host species. Location of origin was
430 categorized based on United States National Oceanic and Atmospheric Administration historical
431 climate region for United States isolates, province and territory for Canadian isolates, state for
432 Mexican isolates, and country for Guatemalan isolates. Categories to be included in the discrete
433 trait models were determined separately for the internal genes, HA subtypes, and NA subtypes.
434 For internal genes, the 20 most common host species and regions (based on average rank among
435 the segments) were chosen to be included in the discrete trait diffusion models. For both HA and
436 NA subtypes, categories with greater than one sequence in a majority of the HA or NA subtypes
437 were included in the models.
438 All sequences collected before 2005 were combined into a single category that was
439 masked in the diffusion models to prevent the influence of inconsistent sampling and to focus
440 diffusion summaries on the most recent years. To mitigate oversampling, two subsampling
441 schemes were used: simple stratified random sampling and phylogenetic diversity-based
442 sampling. In the simple random sample, sequences were stratified by region, host species, and
443 year, and a maximum sample size of three sequences for each stratum were maintained in the
444 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 (33). 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.

Phylogenetic analysis

Using ModelFinder algorithm (34) 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 (35). A general time reversible (GTR) nucleotide substitution model (36–38) with a 4-category gamma distribution of variation among sites and a proportion of invariant sites (39,40) was implemented with a lognormal uncorrelated relaxed molecular clock (41) (mean clock rate prior distribution: uniform 0 – 1, initial value = 0.0033) and a constant coalescent population model (42,43) (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 (44,45). 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. Estimating the posterior distribution of phylogenetic trees based on sequence data alone can be performed separately from the discrete trait diffusion models because the discrete traits represent only two sites (as opposed to the hundreds of nucleotide sites of a genetic sequence) at which the tree likelihood can be calculated. For this reason, the discrete trait model has an insignificant impact on phylogenetic estimation (16). 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. Genetic distance between species was calculated as the average patristic distance, represented as total evolutionary time in million years, across a sample of 1,000 phylogenetic trees estimated under a fossil-calibrated relaxed molecular clock (46). Habitat overlap, migration distance, migration propensity, and latitudinal distribution were summarized from BirdLife species range maps (47) 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 (48) 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 (49). All covariates were log-transformed and standardized before inclusion in the GLM. 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) (50). 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 (51).

DATA AVAILABILITY

Newly sequenced AIV nucleotide sequences have been deposited in GenBank with accession numbers provided in Supplementary Table S11.

ACKNOWLEDGEMENTS

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.

536 REFERENCES

537 1. Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology
538 of influenza A viruses. Microbiol Rev [Internet]. 1992 Mar;56(1):152–79. Available from:
539 http://mmbr.asm.org/content/56/1/152.abstract
540 2. Webby RJ, Webster RG. Emergence of influenza A viruses. Philos Trans R Soc London
541 Ser B Biol Sci [Internet]. 2001 Dec;356(1416):1817–28. Available from:
542 http://rstb.royalsocietypublishing.org/content/356/1416/1817.abstract
543 3. Dhingra MS, Artois J, Dellicour S, Lemey P, Dauphin G, Dobschuetz S Von, et al.
544 Geographical and Historical Patterns in the Emergences of Novel Highly Pathogenic
545 Avian Influenza (HPAI) H5 and H7 Viruses in Poultry. Front Vet Sci [Internet].
546 2018;5:84. Available from: https://www.ncbi.nlm.nih.gov/pubmed/29922681
547 4. Runstadler J, Hill N, Hussein IT, Puryear W, Keogh M. Connecting the study of wild
548 influenza with the potential for pandemic disease. Infect Genet Evol. 2013 Jul;17:162–87.
549 5. Dugan VG, Chen R, Spiro DJ, Sengamalay N, Zaborsky J, Ghedin E, et al. The
550 Evolutionary Genetics and Emergence of Avian Influenza Viruses in Wild Birds. PLoS
551 Pathog [Internet]. 2008 May;4(5):e1000076. Available from:
552 http://www.ncbi.nlm.nih.gov/pubmed/18516303
553 6. Gultyaev AP, Fouchier RAM, Olsthoorn RCL. Influenza Virus RNA Structure: Unique
554 and Common Features. Int Rev Immunol [Internet]. 2010 Nov;29(6):533–56. Available
555 from: https://www.ncbi.nlm.nih.gov/pubmed/20923332
556 7. Bahl J, Krauss S, Kühnert D, Fourment M, Raven G, Pryor SP, et al. Influenza a virus
557 migration and persistence in North American wild birds. PLoS Pathog [Internet].
558 2013;9(8):e1003570. Available from:

559 http://tmclibrary.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwnV07a8MwEBZJ
560 oJCl9N30AZqy2VUky7KHDGloSAsdCn1tQbak1tA4wXGG5tdXZzmh7tjFgyWB0cmnO
561 92n70OIUZ94f3zCQHMVp4aEMQtsHsQNC1IVER2QxMbbHO4Nj0ds_M6fHsRbC21R
562 ZACybEAV_Tz7rOCW9bwWld6abyBvhFdqyAQXdlX2y8Xia7ic
563 8. Spackman E, Stallknecht DE, Slemons RD, Winker K, Suarez DL, Scott M, et al.
564 Phylogenetic analyses of type A influenza genes in natural reservoir species in North
565 America reveals genetic variation. Virus Res. Netherlands; 2005 Dec;114(1–2):89–100.
566 9. Fuller TL, Saatchi SS, Curd EE, Toffelmier E, Thomassen HA, Buermann W, et al.
567 Mapping the risk of avian influenza in wild birds in the US [Internet]. Vol. 10, BMC
568 Infectious Diseases. 2010. p. 187. Available from:
569 https://www.ncbi.nlm.nih.gov/pubmed/20573228
570 10. Belkhiria J, Alkhamis MA, Martínez-López B. Application of Species Distribution
571 Modeling for Avian Influenza surveillance in the United States considering the North
572 America Migratory Flyways [Internet]. Vol. 6, Scientific reports. 2016. p. 33161.
573 Available from: https://www.ncbi.nlm.nih.gov/pubmed/27624404
574 11. Herrick KA, Huettmann F, Lindgren MA. A global model of avian influenza prediction in
575 wild birds: the importance of northern regions [Internet]. Vol. 44, Veterinary research.
576 2013. p. 42. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23763792
577 12. Farnsworth ML, Miller RS, Pedersen K, Lutman MW, Swafford SR, Riggs PD, et al.
578 Environmental and demographic determinants of avian influenza viruses in waterfowl
579 across the contiguous united states [Internet]. Vol. 7, PLoS ONE. 2012. p. e32729.
580 Available from: https://www.ncbi.nlm.nih.gov/pubmed/22427870
581 13. Garamszegi LZ, Møller AP. Prevalence of avian influenza and host ecology. Proc R Soc B

582 Biol Sci [Internet]. 2007 Aug;274(1621):2003–12. Available from:
583 http://rspb.royalsocietypublishing.org/content/274/1621/2003.abstract?cited-
584 by=yes&legid=royprsb;274/1621/2003
585 14. Lam TT, Ip HS, Ghedin E, Wentworth DE, Halpin RA, Stockwell TB, et al. Migratory
586 flyway and geographical distance are barriers to the gene flow of influenza virus among
587 North American birds. Ecol Lett [Internet]. 2012 Jan;15(1):24–33. Available from:
588 https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1461-0248.2011.01703.x
589 15. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral
590 genetics and human transportation data to predict the global transmission dynamics of
591 human influenza H3N2. PLoS Pathog [Internet]. 2014;10(2):e1003932. Available from:
592 http://tmclibrary.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwtV1ba9swFBZpY
593 dCXsfuyG3ov7mRLtuyHPWRhpRuluzRlezOSLbWDxQ65DLpfv3Mk2XFKB9vDSDBB
594 sWVF57POp5NzIYQnRyy6sSYUGOEDul0wU1SxBrVZKM210oxzLVxc7XTCp9_Szx_
595 k19Goq3C4bfuvgoc2ED0G0v6D8PtOoQE-AwTgCCCA41_BADil
596 16. Baele G, A. MS, Rambaut A, Lemey P. Emerging Concepts of Data Integration in
597 Pathogen Phylodynamics. Syst Biol. 2016;66(1):e65.
598 17. Lu L, Brown AJL, Lycett SJ. Quantifying predictors for the spatial diffusion of avian
599 influenza virus in China. BMC Evol Biol [Internet]. 2017;17(1):16. Available from:
600 https://doi.org/10.1186/s12862-016-0845-3
601 18. Wei K, Li Y. Global genetic variation and transmission dynamics of H9N2 avian
602 influenza virus. Transbound Emerg Dis. 2018 Apr;65(2):504–17.
603 19. Faria NR, Suchard MA, Rambaut A, Streicker DG, Lemey P. Simultaneously
604 reconstructing viral cross-species transmission history and identifying the underlying

605 constraints [Internet]. Vol. 368, Philosophical Transactions of the Royal Society of
606 London, Series B: Biological Sciences. 2013. p. 20120196. Available from:
607 https://www.ncbi.nlm.nih.gov/pubmed/23382420
608 20. Breban R, Drake JM, Stallknecht DE, Rohani P. The Role of Environmental Transmission
609 in Recurrent Avian Influenza Epidemics. Fraser C, editor. PLoS Comput Biol [Internet].
610 Public Library of Science; 2009 Apr 10 [cited 2019 Apr 18];5(4):e1000346. Available
611 from: http://dx.plos.org/10.1371/journal.pcbi.1000346
612 21. Roche B, Drake JM, Brown J, Stallknecht DE, Bedford T, Rohani P. Adaptive Evolution
613 and Environmental Durability Jointly Structure Phylodynamic Patterns in Avian Influenza
614 Viruses. Fraser C, editor. PLoS Biol [Internet]. Public Library of Science; 2014 Aug 12
615 [cited 2019 Apr 18];12(8):e1001931. Available from:
616 http://dx.plos.org/10.1371/journal.pbio.1001931
617 22. Bahl J, Vijaykrishna D, Holmes EC, Smith GJD, Guan Y. Gene flow and competitive
618 exclusion of avian influenza A virus in natural reservoir hosts. Virology [Internet].
619 2009;390(2):289–97. Available from:
620 http://utsph.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwpV3JTsMwELVQERI
621 X9qUskj-
622 AtGk2x8cCrZA4IRVVPUW2Y4ugklZpUyhfj8dxWrGIA9wi25GVGWdmbM97g5DvtVz
623 ni03gSntuzl3mEY-
624 HcNsZURkAn35A0kACOHnwEA4HdDT07i00xiZZWk9QWXhju21L28q2Pc0yQPxqX6
625 O9nUthG0GAhDuMYtiO9W-v10hJQ
626 23. Fourment M, Darling AE, Holmes EC. The impact of migratory flyways on the spread of
627 avian influenza virus in North America. BMC Evol Biol. 2017;17(1).

628 24. Woolhouse MEJ. Population biology of emerging and re-emerging pathogens. Trends
629 Microbiol [Internet]. 2002;10(10):s7. Available from:
630 https://www.sciencedirect.com/science/article/pii/S0966842X02024289
631 25. Longdon B, Brockhurst MA, Russell CA, Welch JJ, Jiggins FM. The evolution and
632 genetics of virus host shifts [Internet]. Vol. 10, PLoS pathogens. 2014. p. e1004395.
633 Available from: https://www.ncbi.nlm.nih.gov/pubmed/25375777
634 26. Longdon B, Hadfield JD, Webster CL, Obbard DJ, Jiggins FM. Host phylogeny
635 determines viral persistence and replication in novel hosts [Internet]. Vol. 7, PLoS
636 Pathogens. 2011. p. e1002260. Available from:
637 https://www.ncbi.nlm.nih.gov/pubmed/21966271
638 27. Cuthill JH, Charleston MA. A SIMPLE MODEL EXPLAINS THE DYNAMICS OF
639 PREFERENTIAL HOST SWITCHING AMONG MAMMAL RNA VIRUSES. Evolution
640 (N Y) [Internet]. 2013 Apr;67(4):980–90. Available from:
641 https://www.jstor.org/stable/23463853
642 28. Keeler SP, Dalton MS, Cressler AM, Berghaus RD, Stallknecht DE. Abiotic factors
643 affecting the persistence of avian influenza virus in surface waters of waterfowl habitats
644 [Internet]. Vol. 80, Applied and Environmental Microbiology. 2014. p. 2910–7. Available
645 from: https://www.ncbi.nlm.nih.gov/pubmed/24584247
646 29. Dalziel AE, Delean S, Heinrich S, Cassey P. Persistence of low pathogenic influenza A
647 virus in water: a systematic review and quantitative meta-analysis [Internet]. Vol. 11,
648 PLoS One. 2016. p. e0161929. Available from:
649 https://www.ncbi.nlm.nih.gov/pubmed/27736884
650 30. 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]. Public Library of Science; 2014 Aug 12 [cited 2019 Apr 18];9(8):e104360. Available from: http://dx.plos.org/10.1371/journal.pone.0104360

31. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May;30(9):1312–3.

32. 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

33. 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

34. 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.

35. 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

36. 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

37. Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary

substitution rates. J Mol Evol. 1984;20(1):86–93.

38. 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

39. 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

40. 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

41. 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

42. 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

43. 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_c3TLU

O_fvh6axNHuJFOmSv0fJy15EU3n8-uFrW_

44. 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

45. 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

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

47. 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

48. 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

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

50. 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

51. 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

TABLES

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

FIGURE CAPTIONS

Figure 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.

Figure 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)

Figure 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). 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.

Figure 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.

773 SUPPLEMENTARY MATERIAL CAPTIONS

774 Figure S1. Evolutionary parameter estimation for North American avian influenza viruses of
775 wild birds. Estimated parameters include A) time to most recent common ancestor (TMRCA), B)
776 molecular clock rate, and C) effective population size. Parameters are compared across internal
777 gene segments (blue), hemagglutinin gene subtypes (orange), and neuraminidase gene subtypes
778 (purple) as well as between subsampling strategies, phylogenetic diversity-based sample (left,
779 dark grey) and stratified random sample (right, light grey). Median values (black midline)
780 indicated as well as the 95% highest posterior density (whiskers).

781 Figure S2. Host species temporal distribution of sampled North American avian influenza virus
782 PB2 gene segment sequences, 2005 – 2016. Proportions of represented host species are
783 compared between the phylogenetic diversity-based sample (PDA) and the stratified random
784 sample (stratified).

785 Figure S3. Geographic region temporal distribution of sampled North American avian influenza
786 virus PB2 gene segment sequences, 2005 – 2016. Proportions of represented regions are
787 compared between the phylogenetic diversity-based sample (PDA) and the stratified random
788 sample (stratified).

789 Figure S4. Heat map of supported viral transition rates among host species across avian influenza
790 virus gene segments and subtypes. Colored cells represent the magnitude of the transition rate
791 from the species in the first column (source) to the species in the second column (sink). White
792 cells were transition rates that were not supported (Bayes factor < 100). Results from both
793 subsampling strategies (phylogenetic diversity-based sample (PDA) and stratified random
794 sample (stratified)) are presented for comparison.

795 Figure S5. Discrete trait diffusion models of North American avian influenza using a stratified
796 random sample of genetic sequences. Host models (left) are presented for combined internal
797 gene segment (A), hemagglutinin gene subtype (B), and neuraminidase gene subtype (C) models.
798 Source host species on the left of the chord diagrams contribute viral diversity to sink host
799 species on the right. The magnitude of the viral transition rate is proportional to the width of the
800 band, and statistically supported rates darkened. Bands are colored by the host order of the
801 source species (Charadriiformes – red; Anseriformes – blue). Similarly, geographic models
802 (right) are summarized for combined internal gene segment (D), hemagglutinin gene subtype (E),
803 and neuraminidase gene subtype (F) models. Arrow width is proportional to the magnitude of the
804 transition rate, and only statistically supported rates are displayed. (AK – Alaska, AB – Alberta,
805 BC – British Columbia, GT – Guatemala, MW – Midwest, NB – New Brunswick, NE –
806 Northeast, NL – Newfoundland and Labrador, NRP – Northern Rockies and Plains, NS – Nova
807 Scotia, NW – Northwest, OV – Ohio Valley, PE – Prince Edward Island, QC – Quebec, S –
808 South, SE – Southeast, SON – Sonora, SW – Southwest, W – West)
809 Figure S6. Heat map of supported viral transition rates among geographic regions across avian
810 influenza virus gene segments and subtypes. Colored cells represent the magnitude of the
811 transition rate from the region in the first column (source) to the region in the second column
812 (sink). White cells were transition rates that were not supported (Bayes factor < 100). Results
813 from both subsampling strategies (phylogenetic diversity-based sample (PDA) and stratified
814 random sample (stratified)) are presented for comparison.
815 Table S1. Demographic characteristics of 303 wild bird surveillance samples with newly
816 sequenced avian influenza isolates, 2003 – 2016.

Table S2. 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 S3. Host and regional distribution of phylogenetic diversity-based subsample of influenza virus gene segments isolated from North American wild birds.

Table S4. Host and regional distribution of stratified subsample of influenza virus gene segments isolated from North American wild birds.

Tables S5 – S7. Host species transition rate matrix from combined internal gene model (Table S5), combined hemagglutinin subtype model (Table S6), and combined neuraminidase subtype model (Table S7). 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).

Tables S8 – 10. Geographic region transition rate matrix from combined internal gene model (Table S8), combined hemagglutinin subtype model (Table S9), and combined neuraminidase subtype model (Table S10). 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

840 Labrador, RP – Northern Rockies and Plains, NSC – Nova Scotia, NW – Northwest, OV – Ohio

841 Valley, PEI – Prince Edward Island, QUE – Quebec, S – South, SE – Southeast, SON – Sonora,

842 SW – Southwest, W – West)

843 Table S11. Names and GenBank accession numbers of 303 newly sequenced AIV nucleotide

844 sequences.

Figure 1

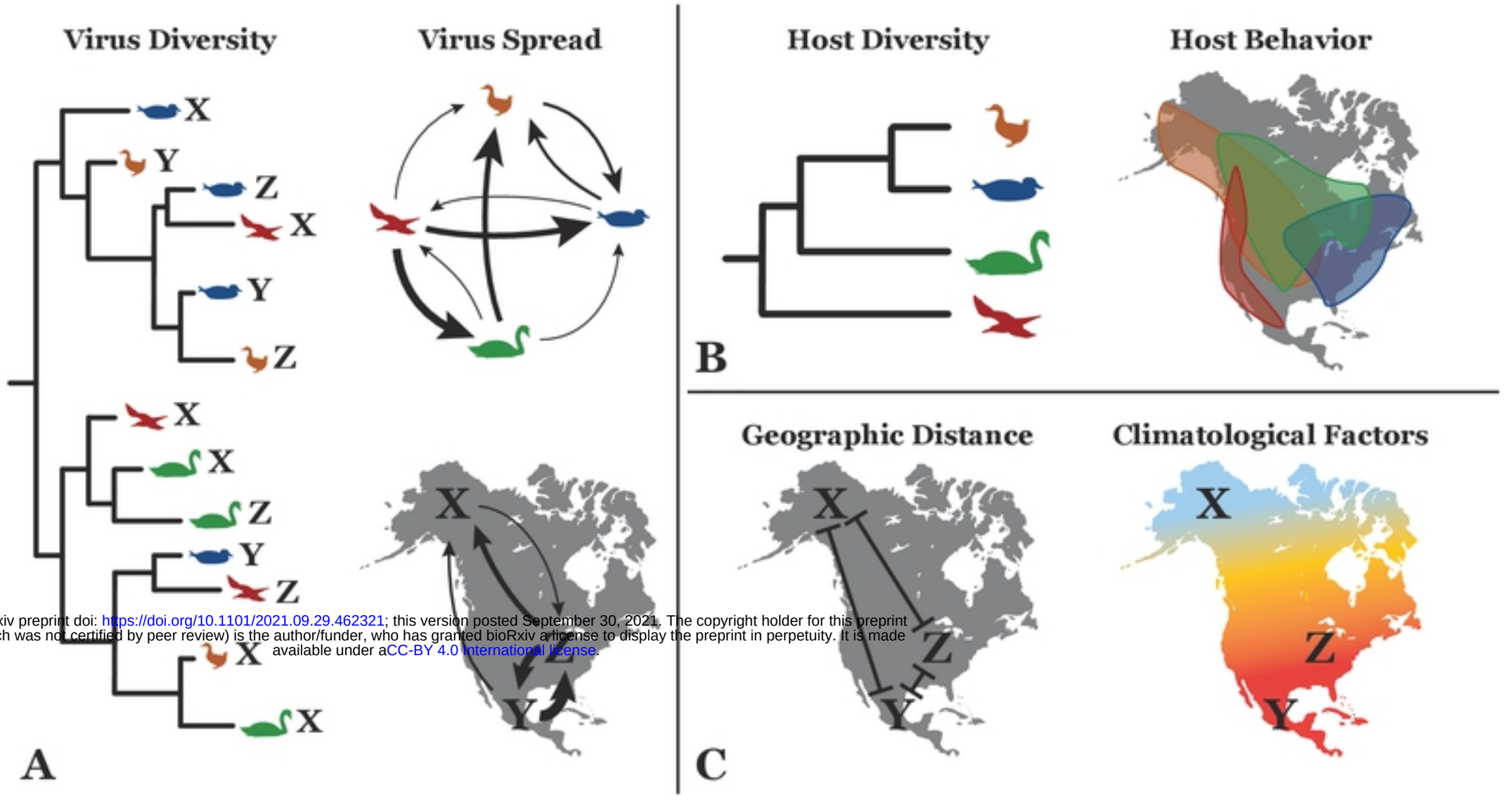

Figure1

Figure 2

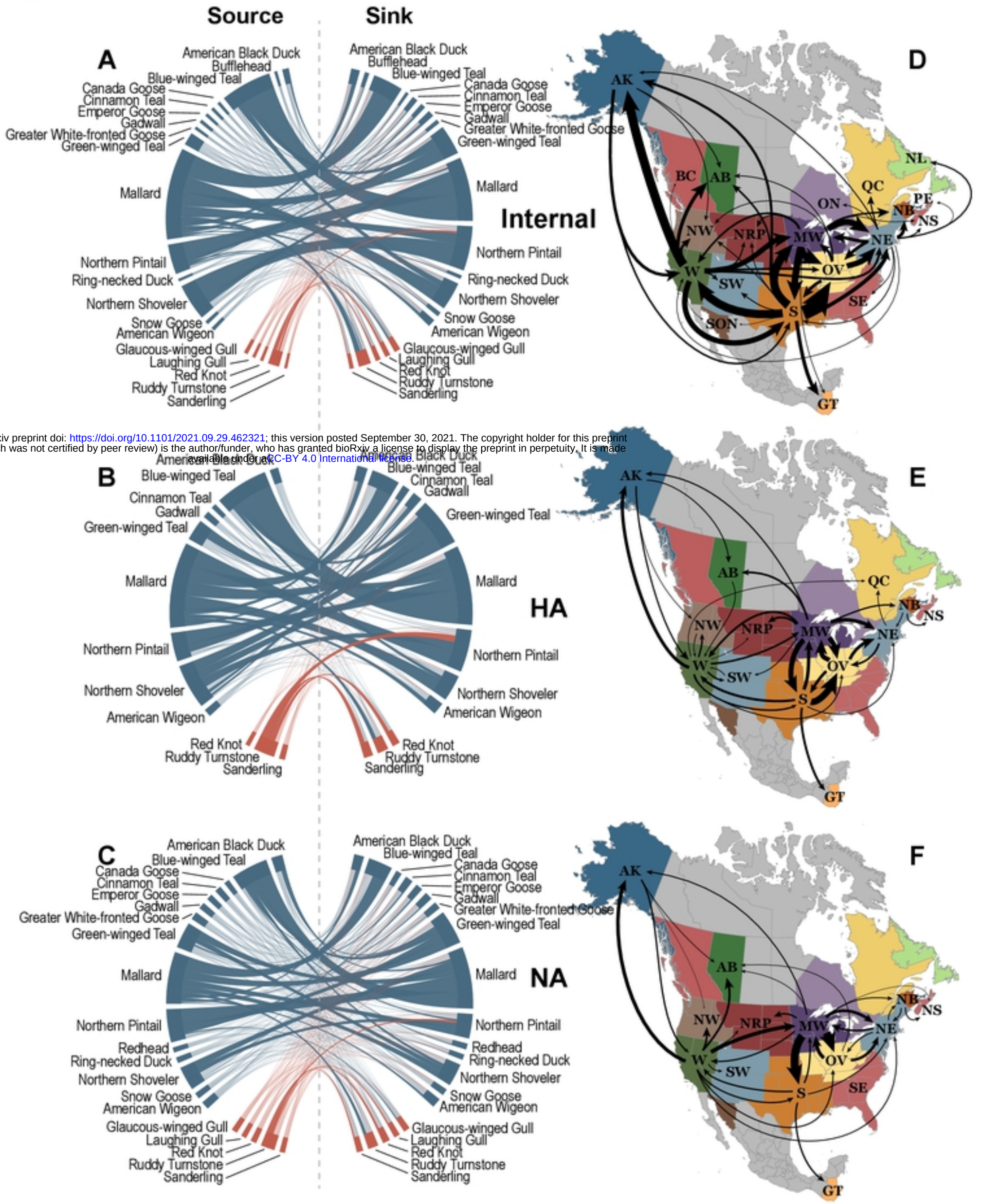

Figure2

Figure 3

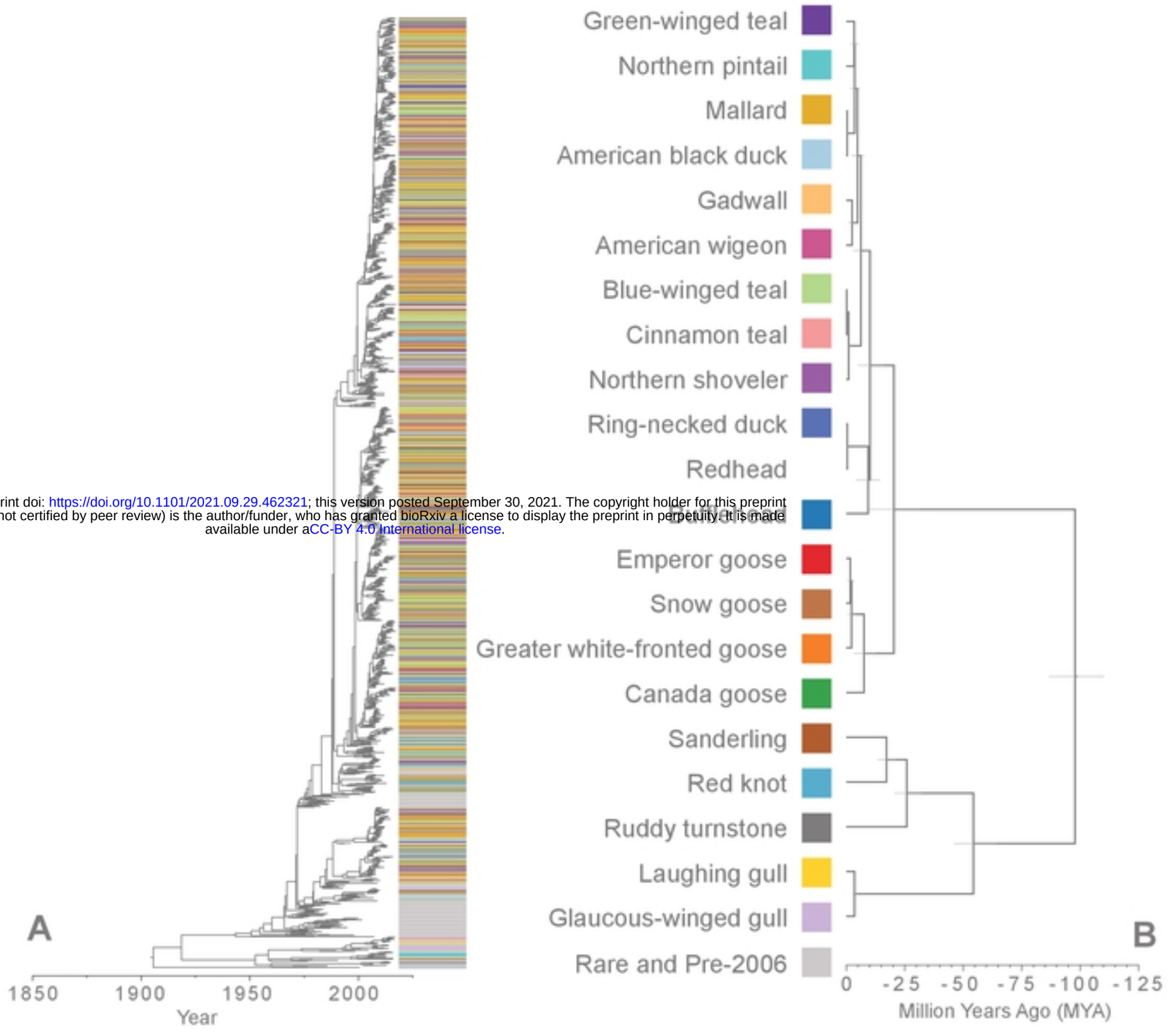

Figure3

Figure 4

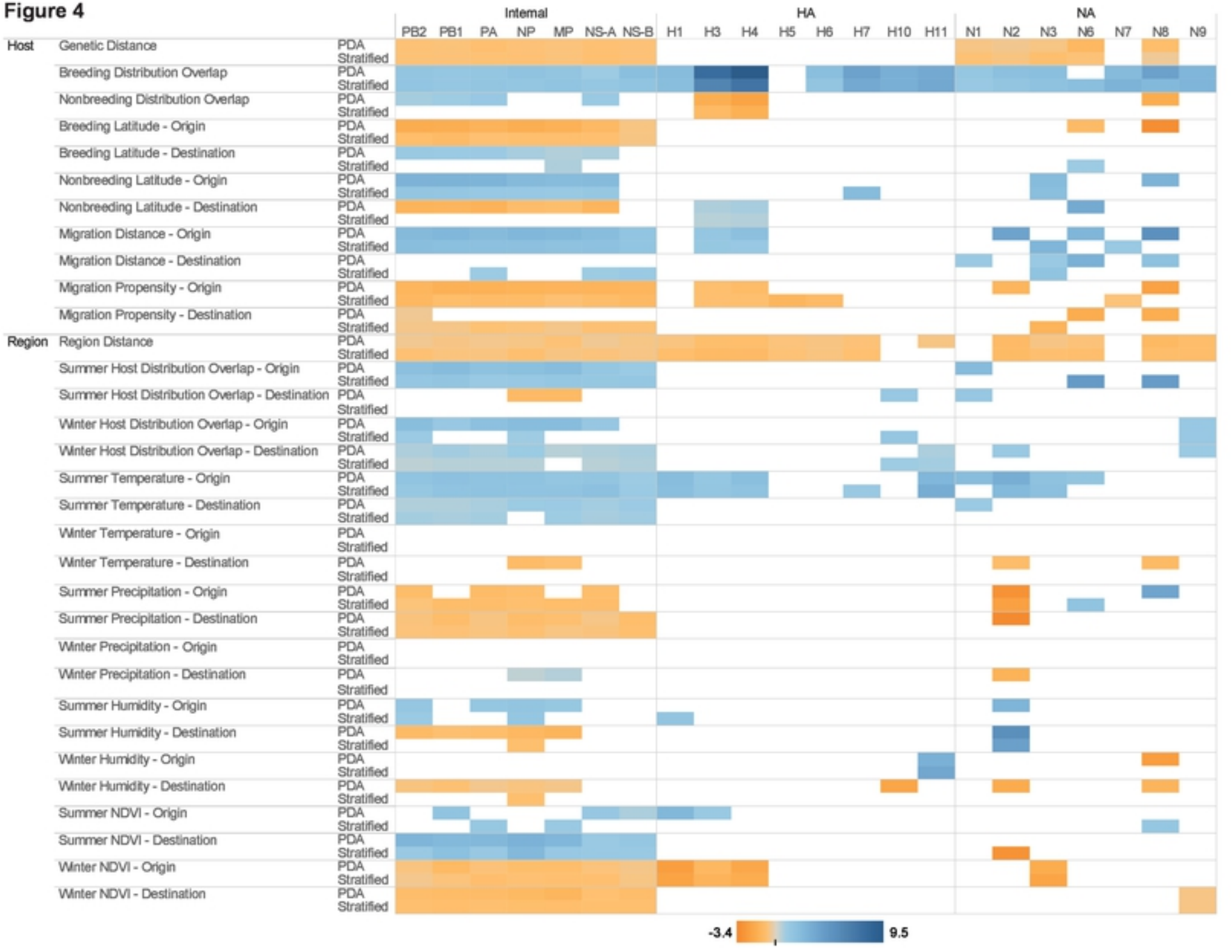

Figure4