VIRUS EVOLUTION

Virus Evolution, 2019, 5(1): vey043

doi: 10.1093/ve/vey043
Research article

Incorporating sampling uncertainty in the geospatial assignment of taxa for virus phylogeography

Matthew Scotch,[1,2,*,†] Tasnia Tahsin,[1] Davy Weissenbacher,[3] Karen O'Connor,[3] Arjun Magge,[1,2] Matteo Vaiente,[1,2] Marc A. Suchard,[4,5,6,‡] and Graciela Gonzalez-Hernandez[3]

[1]College of Health Solutions, Arizona State University, 550 N. 3rd St., Phoenix, AZ 85004, USA, [2]Biodesign Center for Environmental Health Engineering, Arizona State University, 727 E. Tyler St, Tempe, AZ 85287, USA, [3]Department of Biostatistics, Epidemiology, and Informatics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 423 Guardian Drive, Philadelphia, PA 19104, USA, [4]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, 621 Charles E. Young Dr. South, Los Angeles, CA, 90095 USA, [5]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, 695 Charles E. Young Dr. South, Los Angeles, CA 90095, USA and [6]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, 650 Charles E Young Dr. South, Los Angeles, CA 90095, USA

*Corresponding author: E-mail: matthew.scotch@asu.edu

†http://orcid.org/0000-0001-5100-9724

‡http://orcid.org/0000-0001-9818-479X

Abstract

Discrete phylogeography using software such as BEAST considers the sampling location of each taxon as fixed; often to a single location without uncertainty. When studying viruses, this implies that there is no possibility that the location of the infected host for that taxa is somewhere else. Here, we relaxed this strong assumption and allowed for analytic integration of uncertainty for discrete virus phylogeography. We used automatic language processing methods to find and assign uncertainty to alternative potential locations. We considered two influenza case studies: H5N1 in Egypt; H1N1 pdm09 in North America. For each, we implemented scenarios in which 25 per cent of the taxa had different amounts of sampling uncertainty including 10, 30, and 50 per cent uncertainty and varied how it was distributed for each taxon. This includes scenarios that: (i) placed a specific amount of uncertainty on one location while uniformly distributing the remaining amount across all other candidate locations (correspondingly labeled *10*, *30*, and *50*); (ii) assigned the remaining uncertainty to just one other location; thus 'splitting' the uncertainty among two locations (i.e. *10/90*, *30/70*, and *50/50*); and (iii) eliminated uncertainty via two predefined heuristic approaches: assignment to a centroid location (CNTR) or the largest population in the country (POP). We compared all scenarios to a reference standard (RS) in which all taxa had known (absolutely certain) locations. From this, we implemented five random selections of 25 per cent of the taxa and used these for specifying uncertainty. We performed posterior analyses for each scenario, including: (a) virus persistence, (b) migration rates, (c) trunk rewards, and (d) the posterior probability of the root state. The scenarios with sampling uncertainty were closer to the RS than CNTR and POP. For H5N1, the absolute error of virus persistence had a median range of 0.005–0.047 for scenarios with sampling uncertainty—(i) and (ii) above—versus a range of 0.063–0.075 for CNTR and POP. Persistence for the pdm09 case study followed a similar trend as did our analyses of migration rates across scenarios (i) and (ii). When considering the posterior probability of the root state, we found all but one of the H5N1 scenarios with sampling uncertainty had agreement with the RS on the origin of the outbreak whereas both CNTR and POP disagreed. Our results suggest that assigning geospatial uncertainty to taxa benefits estimation of virus phylogeography as compared to *ad-hoc* heuristics. We also found that, in general, there was limited difference in results regardless of how the sampling uncertainty was assigned; uniform distribution or split between two locations did not greatly impact posterior results. This framework is available in BEAST v.1.10. In future work, we will explore viruses beyond influenza. We will also develop a web interface for researchers to use our language processing methods to find and assign uncertainty to alternative potential locations for virus phylogeography.

Key words: phylogeography; influenza A virus; geography.

1. Introduction

The National Center for Biotechnology Information (NCBI), specifically GenBank (Benson et al. 2013), provides an abundance of available viral sequence data for phylogeography. Sequences and their metadata can be downloaded and imported into software applications that run analyses that generate phylogeographic trees. In discrete phylogeography using software such as BEAST (Suchard et al. 2018), virus diffusion is estimated by ancestral state reconstruction of observed geospatial (discrete) locations along a phylogeny (Lemey et al. 2009). In these studies, researchers assign geospatial traits to each taxon often by using metadata from a sequence record. This is different than continuous phylogeography where coordinates in space such as latitude and longitude are utilized and migration is estimated using a random walk (Lemey et al. 2010).

Despite the popularity of the GenBank sequence database, its virus records only contain about 36 per cent precise or sufficient geospatial metadata such as a county, town, or region within a state (Tahsin et al. 2014). For example, locations such as Canada or USA are more often indicated instead of Quebec City, QC or Concord, NH (Scotch et al. 2011). In a GenBank record, this information is most often indicated in the country field (Supplementary Fig. S1, GenBank record (FJ966084) (Garten et al. 2009)), where, despite the label, the researcher is able to indicate the exact location beyond the country such as a county or town. While town or county might be included in the corresponding journal article, this valuable information is not available for immediate use unless it is extracted from the article and then linked back to the appropriate sequence record. This hinders phylogeography based on secondary data sources (such as GenBank) since the researcher is forced to review the corresponding paper (or other primary source document[s]) for additional geospatial metadata and then link this data to the individual record. This also can impact the value of discrete phylogeographic analysis especially for localized studies where second or third level administrative boundaries such as a county or a town is preferable over more generic spatial data such as state or county. For example, town information was used to study the phylogeography of H3N2 in Peru (Pollett et al. 2015). If the authors had access to only country-level information (e.g. Peru), this level of analysis would not have been possible and erroneous assumptions and public health interventions might have been made.

There are different approaches for dealing with imprecise geospatial metadata. The simplest approach could be to discard data. However, this can severely reduce the size of the dataset if there is a large proportion of imprecise records. A second approach could be to select the centroid location (see (Carrel et al. 2010; Hayman et al. 2011; Pybus et al. 2012; Beck et al. 2013; Alkhamis, Moore, and Perez 2015; Lukashev et al. 2016; Wei and Li 2018) for recent examples) by using a resource such as GeoNames.org to identify the middle of a geographic area. Another approach could be to select the area with the greatest population density. Distance between the largest cities in a study area or population size of the destination and origin have been used as a predictor in phylogeographic generalized linear models (GLMs) (Allicock et al. 2012; Dudas et al. 2017; Tian et al. 2017). Population size could also be used to assign locations to taxa with geospatial uncertainty. The use of these heuristics for every taxa with uncertainty in a given dataset will likely produce an oversampling of an incorrect location. Another approach is to incorporate a probability that a more precise location exists for a given taxa. For example, for GenBank record FJ966084 (shown in Supplementary Fig. S1), one could assign a probability that Los Angeles or San Diego is a more precise location than the given state of California. In our prior work, we developed GeoBoost and other automated language processing methods to address the lack of geospatial certainty in sequence databases. *GeoBoost* improves the granularity of the location of the infected host (LOIH) for GenBank records (Tahsin et al. 2018). It scans each record, including strain names such as *A/Boston/YGA_02024/2013*, as well as any full-text article that is linked to the record, full text, tables, and supplementary materials. From these, GeoBoost extracts all geospatial mentions and assigns a probability of the LOIH given the GenBank record, $P(L_i | R_i)$ where L_i represents the unknown location and R_i indicates the linked record information for taxon *i*. The probabilities are currently based on a set of predefined rules that assign higher probabilities to more specific and accurate locations found in papers that can be used jointly with information scanned from the GenBank record (Tahsin et al. 2018). For example, if *Sydney* is found in a table row along with a GenBank accession number then it would be given a high probability for that record. Conversely, if Sydney was found in the free text without any relevant information nearby, it would be assigned a much lower probability for that record. Thus, for each taxon, we obtain a list of location-specific probabilities. In this context, we define *sampling uncertainty* as $1 - P(L_i | R_i)$. Here, when studying viruses, this implies the actual LOIH for that virus (taxa) is not known amongst a set of discrete locations.

In this article, we evaluate the use of sampling uncertainty for virus phylogeography in a Bayesian discrete setting by using the location-specific probabilities. By analyzing posterior metrics of the phylogeographic process, we hypothesize that our work will improve analysis (as compared with centroid and population approaches) in tracking evolutionary changes in viral genomes and their spread. The addition of more precise geospatial metadata in phylogeography has potential public health significance, as it could enable health agencies to better target areas that represent the greatest public health risk, for example.

In addition, by improving geospatial metadata linked to popular sequence databases, we will enrich other sciences beyond phylogeography that utilize this information such as molecular epidemiology, population genetics, and environmental health.

2. Methods

We provide a formal probabilistic framework in which we allow for analytic integration of uncertainty for discrete virus phylogeography. Here the sufficient statistics of the process are location-specific probabilities that each taxon arises from that location. Current discrete phylogeographic analyses consider the sampling location of each taxon as fixed to a single or small number of locations and without observation uncertainty. We relaxed this strong assumption in a prerelease build of BEAST v1.8.4 (r20160319) that allows for analytic integration over the uncertainty to perform virus phylogeography. This framework is also available in BEAST v1.10 (Suchard et al. 2018). We inserted the probabilities generated by GeoBoost (Tahsin et al. 2018) into a BEAST XML. In Supplementary Fig. S2, we show an example for one taxon with two location probabilities ($P(L_i|R_i)$) in Indonesia. The sum of the probabilities for a given taxon should equal 1.0. Here, Jakarta is assigned a probability of 0.45 and Sumatra a probability of 0.55. The sampling uncertainty is 1 – $P(L_i|R_i)$ and thus 0.55 for Jakarta and 0.45 for Sumatra. To incorporate these probabilities, we first assume a uniform distribution over the prior location mass function $P(L_i)$ and employ Bayes theorem to identify that $P(R_i|L_i) = c_i P(L_i|R_i)$, where evaluating the proportionality constant c_i remains unnecessary for inference because our MCMC scheme depends only on ratios of the spatial process likelihood. Without sampling uncertainty $P(R_i|L_i)$ equals an elementary vector, 1 for one value of L_i and 0 for the remaining locations. One then computes the spatial process likelihood via a peeling algorithm (Felsenstein 1981) that integrates out the unobserved locations states for all internal nodes. Here, we replace the elementary vector with $P(R_i|L_i)$ and peel to return the likelihood up to the constant $\prod_{i=0}^{N} c_i$.

For two public health case studies: highly pathogenic avian influenza (HPAI) H5N1 in Egypt and 2009 influenza A (pdm09) H1N1 in North America, we evaluated several scenarios against a *reference standard* (RS) in which all taxa had one sampling location with no uncertainty (which we considered the 'correct' location of the virus) (Supplementary Fig. S3). For some of the scenarios, we included various amounts of sampling uncertainty; for others, we assumed no sampling uncertainty through the assignment of locations to the centroid or to the most populated location in the study area.

From the RS, we randomly selected 25 per cent of the taxa and used these for specifying uncertainty; the remaining 75 per cent we preserved a known location with certainty (Fig. 1). We repeated this experiment five times. We ran multiple scenarios where the only difference was the amount of sampling uncertainty for the taxa with unknown locations in the phylogeny, including:

1. *10*, in which we assigned a taxon a sampling uncertainty of no more than 0.10 for one location. We used the location identified by GeoBoost that had the greatest probability (or least amount of uncertainty) for that taxon. We assigned the remaining amount of sampling uncertainty (∼0.90) uniformly across the other possible locations until each taxon had a sum of sampling uncertainty at or near 1.0. We used the output of GeoBoost to assign sampling uncertainty for each taxon. If, for a given taxa, GeoBoost did not assign a location with a sampling uncertainty at or below 0.10, we randomly assigned the sampling uncertainty.
2. *30*, in which we kept everything from dataset *10* but increased the sampling uncertainty by 0.20 and recalculated the uniform distribution.
3. *50*, in which we kept everything from dataset *30* but increased the sampling uncertainty by 0.20 and recalculated the uniform distribution.
4. *10/90* split, in which we assigned, to each taxon with an unknown location, a sampling uncertainty of ≤0.10 for one location. However, rather than a uniform distribution among all remaining locations, we randomly assigned one other location as having the remaining sampling uncertainty (∼0.90). We used the output of GeoBoost to assign sampling uncertainty for these taxa.
5. *30/70* split, in which we kept everything from dataset *10/90* except we increased the sampling uncertainty by 0.20.
6. *50/50* split, in which we kept everything from dataset *30/70* except we increased the sampling uncertainty by 0.20.
7. *centroid (abbreviated CNTR)*, in which we assigned every taxon in the randomly selected 25 per cent as having a known location (i.e. no sampling uncertainty) in the center of the country. We included this scenario since a researcher, instead of discarding taxa with unknown or uncertain locations, might decide to assign them to the location in the middle of the study area. We used Geonames.org to identify the centroid location for a given country and chose the location in our dataset that was nearest to it.
8. *population* (abbreviates *POP*), in which we assigned every taxon in the randomly selected 25 per cent as having a known location of the most populous part of the country. We included this scenario since a researcher, instead of discarding taxa with unknown or uncertain locations, might decide to assign them a location with the most people. We used Wikipedia to identify the most populated location for a given country.

2.1 Case study 1: H5N1 in Egypt

We selected Egypt since it has the greatest number of human cases of H5N1 than any other country in the world (WHO 2017) and represents an on-going public health threat. There have been several studies of H5N1 phylogeography including in China, Indonesia, and Egypt (Wallace et al. 2007; Wallace and Fitch 2008; Lemey et al. 2009; Fusaro et al. 2010; Lam et al. 2012; Scotch et al. 2013; Rao 2014; Alkhamis, Moore, and Perez 2015; Trovao et al. 2015) highlighting the importance of studying the evolution and spread of this virus.

We used the search feature of ZooPhy (Scotch et al. 2010; Scotch 2018) to obtain metadata and sequences. ZooPhy contains virus sequences from GenBank and can be used as a search engine as well as a pipeline for implementing phylogeography analysis. We specified H5N1 hemagglutinin (HA) gene segments in Egypt from 2005 to 2017 with a minimum segment length of 1,659 nucleotides in length. We identified 727 records but removed 87 of them because they lacked a collection date with at least a month of the year. We ran the remaining 640 GenBank accessions through GeoBoost in order to determine the sampling uncertainty for each location given a GenBank record. Egypt is divided into governorates and we aggregated any location such as a town to its governorate level. In total, 485 records (76%) had one location with sampling uncertainty of ≤ 0.10 (most of these were 0.0 or 0.01). For these 485 records, we changed these negligible levels of sampling uncertainty to zero

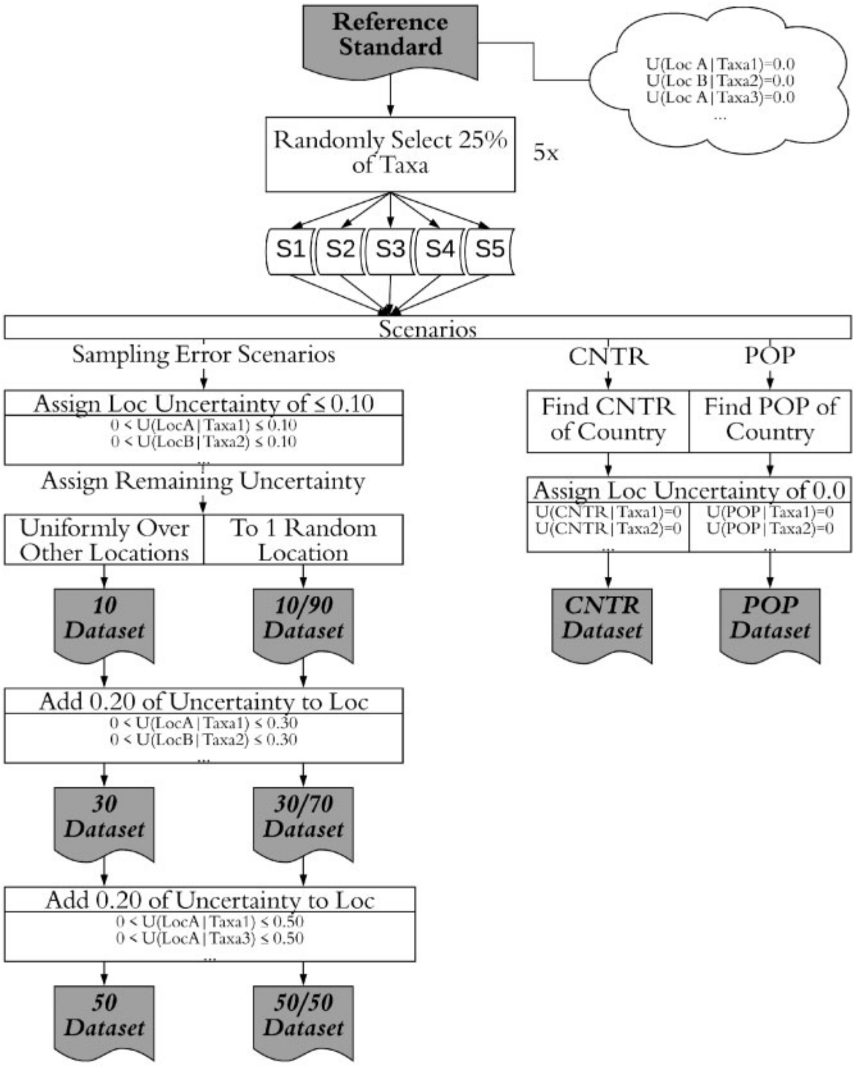

Figure 1. Generation of the individual scenarios from the RS for the evaluation. Abbreviations: S1, sample 1; S2, sample 2, etc.

(thus $P(L_i|R_i) = 1.0$) and discarded the remaining 155 sequences to obtain our RS (Fig. 1). We aligned these 485 sequences in Geneious v6.1.8 (Kearse et al. 2012) using Muscle (Edgar 2004). We used BEAUti v1.8.4 (Drummond et al. 2012) to specify a Hasegowa, Kishino, Yano (HKY) DNA substitution model (Hasegawa, Kishino, and Yano 1985) and a gamma-distributed model of rate heterogeneity (Yang 1994) with codon partitions $1 + 2$ and 3 (Shapiro, Rambaut, and Drummond 2006). We specified an uncorrelated lognormal relaxed molecular clock and a Bayesian Skygrid coalescent tree prior (Gill et al. 2013). In this initial run, we did not specify a geographic state partition. We set the length of the Markov chain Monte Carlo (MCMC) to 100M and a sub sampling rate of 1K steps. We used BEAST v1.8.4 to perform the analyses (Drummond et al. 2012). We initiated three independent runs and used Tracer v.1.6.0 to check for convergence (Rambaut et al. 2018). We used LogCombiner v1.8.4 (Drummond et al. 2012) to combine the log files and specified a 10 per cent burn-in.

We modified the BEAST XML file to specify a symmetric Bayesian stochastic search variable selection (BSSVS) for continuous-time Markov chain (CTMC) ancestral state reconstruction (Lemey et al. 2009) along a target sample of 1K trees from the posterior distribution from our previous MCMC. Here, we commented out parameters and blocks related to DNA evolution. We ran two independent analyses for 10M steps subsampling every 1K steps and used Tracer to check for convergence. We used LogCombiner v1.8.4 to combine the logs and specified 10 per cent burn-in. We used Tracer to check final ESS values (most parameters were well above 200 across all scenarios). We used TreeAnnotator v1.8.4 (Drummond et al. 2012) to create a maximum clade credibility (MCC) tree and specified *Common Ancestor Heights* (*Median Heights* produced negative branch lengths in some instances).

2.2 Case study 2: pdm09 in North America

For the second case study, we studied 2009 influenza A (pdm09) H1N1 in North America during its outbreak year in 2009. It continues to circulate as a seasonal virus and competes with the other influenza A subtype, H3N2. There have been phylogeographic studies on this virus (Lemey, Suchard, and Rambaut 2009; Holmes et al. 2011; Nelson et al. 2011; Su et al. 2015) which suggest that it originated in North America and likely in Mexico (Lemey, Suchard, and Rambaut 2009; Nelson et al. 2011; Su et al. 2015).

We specified pdm09 HA gene segments from infected humans in the USA and Mexico in 2009 with a minimum segment length of 1,659 nucleotides in length. We identified 2,081 records but removed 14 of them because they lacked a collection date with at least a day of the year. We ran the remaining 2,067 GenBank accessions through GeoBoost in order to determine the sampling uncertainty estimates for each location given a GenBank record. In order to reduce bias due to the large amount of US sequences compared with Mexico, we only included sequences from the ten US states with the most sampling uncertainty of ≤0.01. We randomly down-sampled to 300 US sequences across these 10 states. We did not down sample the 143 sequences from Mexico (Supplementary Fig. S3). We aligned our dataset of 443 sequences in Geneious v6.1.8 (Kearse et al. 2012) using Muscle (Edgar 2004). We used BEAUti v1.8.4 (Drummond et al. 2012) to specify a Hasegowa, Kishino, Yano (HKY) DNA substitution model (Hasegawa, Kishino, and Yano 1985) and a gamma-distributed model of rate heterogeneity (Yang 1994) with codon partitions 1 + 2 and 3. We specified an exponential coalescent tree prior as it is often used to study pdm09 during its outbreak year (see (Rambaut and Holmes 2009; Baillie et al. 2012; Lycett et al. 2012; Su et al. 2015; Gachara et al. 2016)). We considered both an uncorrelated lognormal relaxed molecular clock and a strict clock and selected the latter after examination of posterior log files, marginal likelihoods, and the regression of the root-to-tip genetic distance in TempEst v1.5 (Rambaut et al. 2016). In this initial run, we did not specify a geographic state partition. We set the length of the MCMC to 200M and a subsampling rate of 1K steps. We used BEAST v1.8.4 to perform the analyses (Drummond et al. 2012). We initiated two independent runs and used Tracer v.1.6.0 to check for convergence (Rambaut et al. 2018). We combined the posterior log and trees files using LogCombiner v1.8.4 (Drummond et al. 2012) to generate a set of 1K empirical trees.

Since our number of total discrete states was much less than our H5N1 example (17 vs. 24) we were able to specify an asymmetric BSSVS (thus enabling for directionality of transmission routes) along a set of 1K trees for our phylogeographic reconstruction. We used TreeAnnotator to generate MCC trees using *Median Heights* for tree nodes.

2.3 Posterior analysis

For both case studies, we compared the different scenarios to their corresponding RS by performing posterior analysis. Here, we used the program PACT (Bedford 2011) to examine relevant phylogeographic phenomena including persistence (the amount of years that a given virus remains in its geographic origin (Bedford et al. 2015)) and migration rate (the number of lineage-specific migration events (Bedford et al. 2015)). For both persistence and migration, we computed an ‘absolute error’ by taking the absolute value of the difference between the RS and each scenario.

We also calculated Markov rewards on the trunk (i.e. backbone) of the tree as prior work (Su et al. 2015) has shown its value for understanding spatial dynamics of pathogens. Here, the rewards represent the time between the jumps (e.g. state transitions or changes) between geographic locations (Su et al. 2015). For Egypt, we defined the trunk as the basal clade of the tree and the earlier of the two major clades. We modified the XML file by adding Markov jump and reward blocks. We specified a chain length of 10M steps and subsampled every 1K steps. We used TimeSlicer v1.8.4 (Drummond et al. 2012), a program within the BEAST software package to analyze the rewards by time intervals.

For additional analysis of the geospatial process, we used the software program SpreaD3 (Bielejec et al. 2016) to calculate the Bayes factor (BF) for nonzero pairwise migration rates and considered a BF > 100 as a threshold for decisive support per Jeffreys (1998) and Liang and Xiong (2013). We repeated these analyses for all of the scenarios previously described. We also used FigTree (Rambaut 2018) to observe the most likely root state as identified with the greatest posterior probability.

3. Results

3.1 H5N1 in Egypt

In Fig. 2, we show the MCC for the RS. We dated the age of the root as 2005.8 or 9.26 years. This is consistent with prior phylogeographic and epidemiologic findings that suggest that the outbreak began in the second half of 2005 or early 2006 (Cattoli et al. 2011; Arafa et al. 2016; Naguib, Abdelwhab, and Harder 2016). We identify Monufia as the outbreak location which aligns with prior work that suggest the virus originated in the Nile Delta region (Scotch et al. 2013; Arafa et al. 2016). In our online data repository (see *Data availability*), we provide the individual MCCs for each scenario. In Supplementary Fig. S4, we show the average root state posterior probabilities and their corresponding 95 per cent Bayesian highest posterior density (HPD) for the different scenarios (across the five samples). We considered the origin of each MCC as the governorate with the highest posterior probability in the oldest node. We counted the root for each MCC and considered the most frequent location (across the five samples) as the overall origin for the scenarios. In Supplementary Fig. S4, we color the bars to match the branch colors represented in the MCCs for that governorate. Here, we see that five out of the six scenarios with sampling uncertainty agree with the RS on the location of the root state (with the exception of 30 where 2 Qalyubia had a slight edge over Monufia). The CNTR and POP scenarios, both disagreed with the RS as New Valley and Cairo (respectively) were identified as the origin. We also see that the posterior probabilities increase as the amount of sampling uncertainty increases although there is overlap of the 95 per cent Bayesian HPD. It is not surprising that both the CNTR and POP scenarios have high posterior probabilities as the assigned location (either by centroid or through population) dominates selection at these nodes. As an additional posterior metric, we also show the estimate of the likelihood of the tree given the governorates (Supplementary Fig. S5).

In Fig. 3, we show a violin plot of the absolute error of the estimated persistence times compared with the RS for each scenario. We see that the scenarios with sampling uncertainty are much closer to the RS (thickness around zero) than the CNTR or POP scenarios with increasing absolute error as the amount of sampling uncertainty increases (and whether with uniform or split distributions). This implies that scenarios with less sampling uncertainty are closer to the RS in regard to the time that an H5N1 virus is circulating in its original location; yet the deviations are not as severe as they are with the population and centroid scenarios, CNTR and POP. Examination within the types of sampling uncertainty indicate that the scenarios that ‘split’ sampling uncertainty (10/90, 30/70, 50/50) were closer to the RS than the ones in which sampling uncertainty was uniformly distributed among the remaining taxa. In Fig. 4, we show a heatmap of the absolute

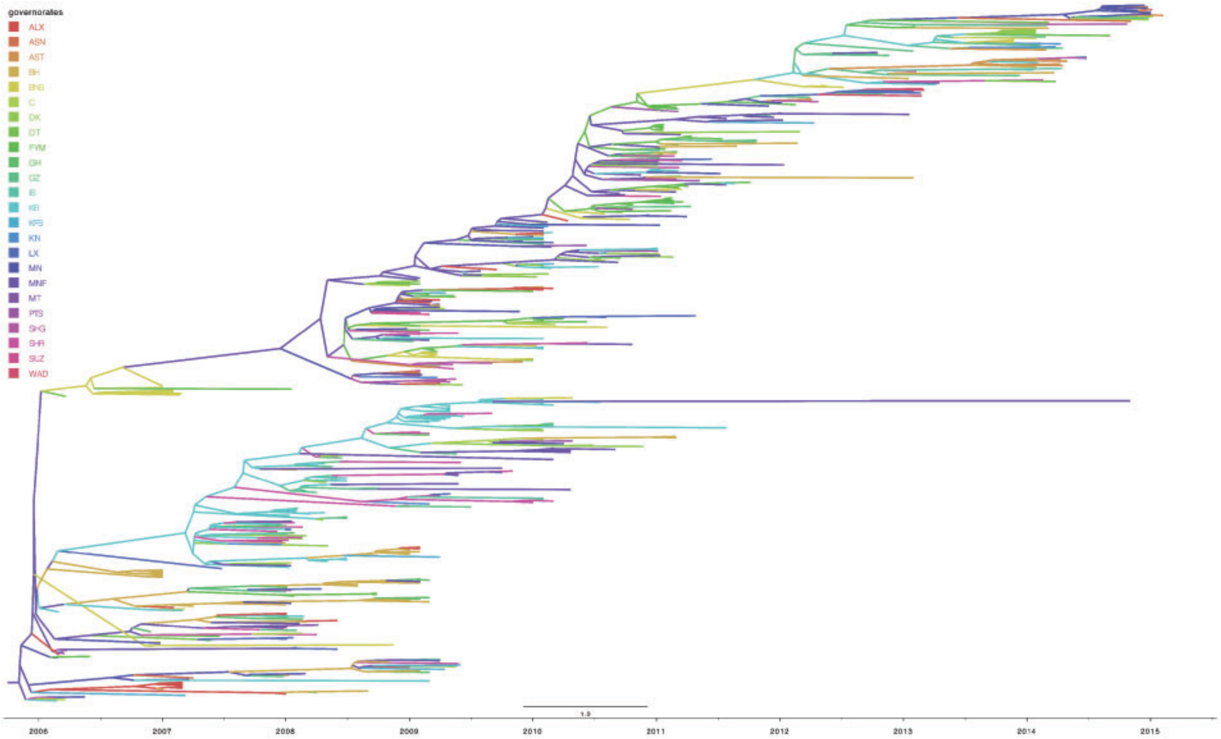

Figure 2. MCC tree of 485 H5N1 HA genes in Egypt with no sampling uncertainty (RS). We color the branches based on governorates with the greatest posterior probability. Abbreviations: ALX, Alexandria; ASN, Aswan; AST, Asyut; BH, Beheira; BNS, Beni Suweif; C, Cairo; DK, Dakahlia; DT, Damietta; FYM, Faiyum; GH, Gharbia; GZ, Giza; IS, Ismailia; KB, Qalyubia; KFS, Kafr el-Sheikh; KN, Qena; LX, Luxor; MN, Minya; MNF, Monufia; MT, Matruh; PTS, Port Said; SHG, Sohag; SHR, Sharqia; SUZ, Suez; WAD New Valley.

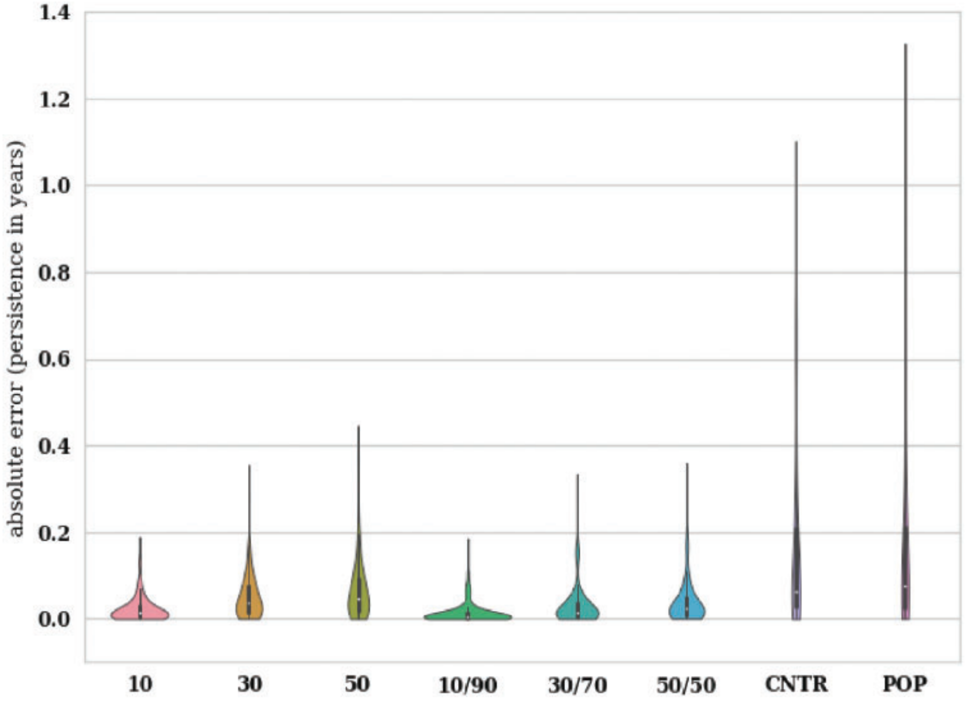

Figure 3. Absolute error of the estimated persistence times compared with the RS for each H5N1 scenarios. Smaller errors are characterized by greater thickness of the violins close to zero. Here, absolute error is measured by $|G - X|$ where X and G are estimated and RS persistence times, respectively. We included all five samples for each scenario.

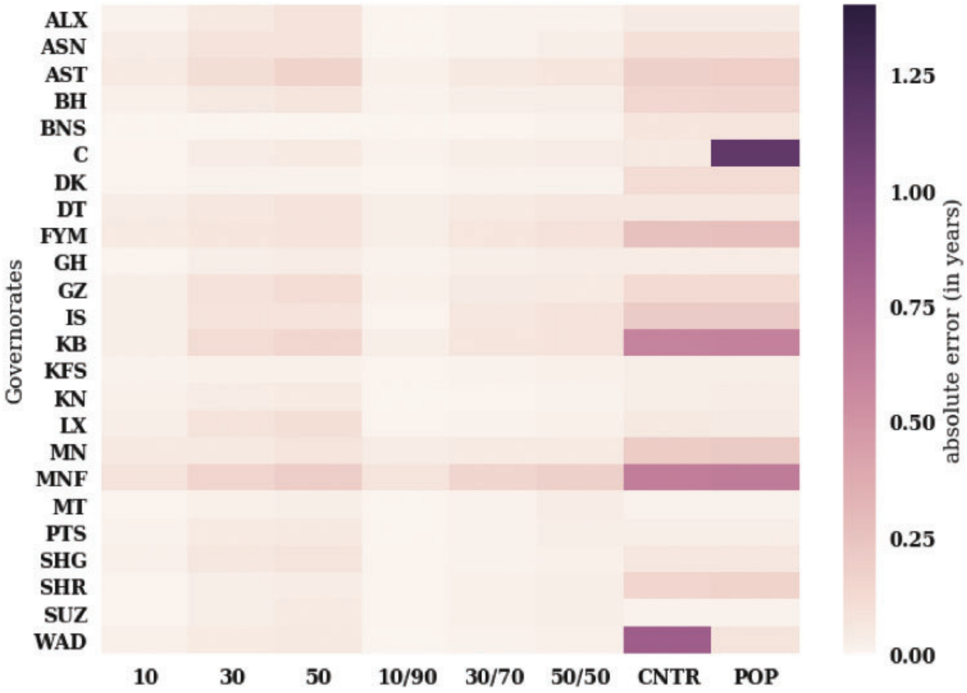

Figure 4. Absolute error of the estimated persistence times compared with the reference for each Egyptian governorate and H5N1 scenario. Smaller errors are characterized by lighter colors. Here, we see the differences are greater for the POP and CNTR scenario including governorates in the Nile Delta such as: Cairo (POP scenario); Qalyubia (both CNTR and POP); and Monufia (both CNTR and POP).

error (in years) for persistence across each governorate. Here, we see great differences in virus persistence with the POP and CNTR scenarios; especially for governorates in the Nile Delta region where H5N1 has shown to propagate such as Cairo, Qalyubia, and Monufia.

In Fig. 5, we show a violin plot of the absolute error of the estimated migration rates compared with the respective RS values for each scenario. As we saw with the persistence results, the scenarios with less sampling uncertainty are closer to the RS estimates of the number of migration events per lineage per year; yet any amount of sampling uncertainty is closer than either CNTR and POP. The same is true when examining individual governorates (Supplementary Fig. S6) including in the Nile Delta region when viruses are originating from Monufia, Minya, or Qalyubia.

In Fig. 6, we show the trunk rewards for each scenario. Here, we see the dominance of the Qalyubia Governorate (abbreviated *KB*) across the RS and all scenarios with sampling uncertainty until 2010 when more equity in trunk time exists among the governorates. The CNTR and POP scenarios have a much different trend as the virus harbors mostly in the location selected by the scenario (New Valley for CNTR or Cairo for POP). In Supplementary Table S1, we display the mean reward share over the time slices with differences from the standard deviation of the RS highlighted in red. CNTR *and* POP contain by far the most absolute differences from RS (shaded in red) with fourteen and fifteen governorates respectively (out of 24). The scenarios with sampling uncertainty ranged from 5 to 8 differences across the 24 governorates.

We show additional posterior estimates related to virus spread including the number of nonzero rates (Supplementary Fig. S7) and the complete list of routes with BFs > 100 (Supplementary Table S2).

3.2 pdm09 in North America

In Fig. 7, we show the MCC for the RS. The age of the root is dated to be 2008.7 or 1.29 years. This is a little earlier than prior studies which have suggested the virus originated in the early part of 2009 (Lemey, Suchard, and Rambaut 2009; Rambaut and Holmes 2009; Mena et al. 2016). In our online data repository (see *Data availability*), we provide the individual MCCs for each scenario. In Supplementary Fig. S8, we show the average root state posterior probabilities and their corresponding 95 per cent Bayesian HPD for the different scenarios (across the five samples). We color the bars to match the branch colors represented in the MCCs for that governorate. Here, we see that all scenarios agreed with the RS that Mexico City was the likely origin of the outbreak. The high average posterior probability for the POP scenario is reflective of the over-assignment of Mexico City to taxa in this scenario since it is the most heavily populated state in Mexico.

As an additional posterior metric, we also show the estimate of the likelihood of the tree given the states (Supplementary Fig. S9).

In Fig. 8, we show a violin plot of the absolute error of the estimated persistence times compared with the respective RS for each scenario. Like the H5N1 results, we see that the scenarios with sampling uncertainty are much closer to the RS (thickness around zero) than CNTR or POP scenarios with increasing absolute error as the amount of sampling uncertainty increases whether with uniform or split distributions. This implies that scenarios with less sampling uncertainty are closer to the RS in regard to the time that a pdm09 virus is circulating in its original location; yet the deviations are not as severe as the CNTR and POP scenarios. As we saw with the H5N1 results, examination within the types of sampling uncertainty indicate that the 'split' scenarios (e.g. 10/90. 30/70, 50/50) were slightly closer to the RS than the scenarios in which sampling uncertainty was uniformly distributed among the remaining taxa. In Fig. 9, we show

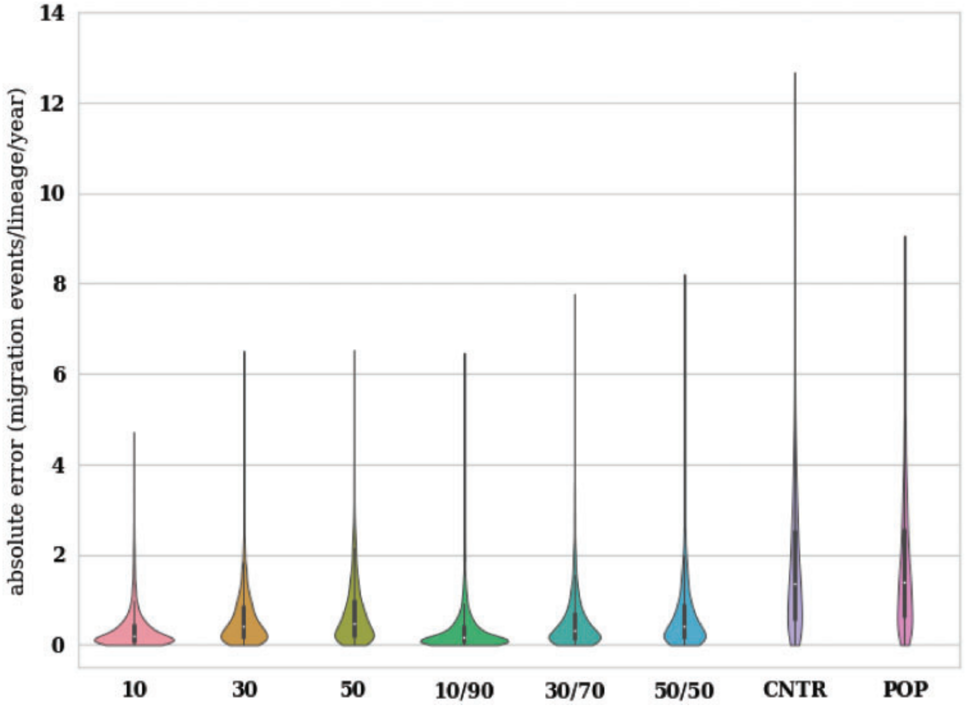

Figure 5. The absolute error of the estimated migration rates compared with the respective RS values for H5N1 in Egypt. Here, smaller errors are characterized by greater thickness of the violins close to zero. Absolute error here is measured by |G − X| where X and G are log transformed estimated and RS migration rates respectively. We included all five samples for each scenario.

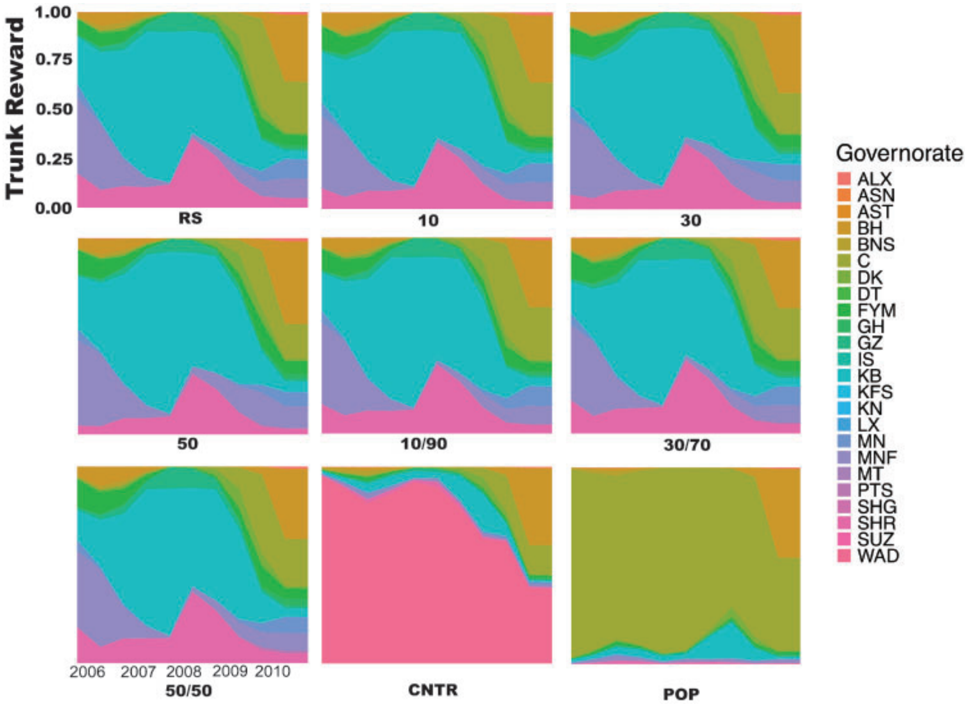

Figure 6. Trunk rewards for the RS (top left) and the eight scenarios for H5N1 in Egypt.

a heatmap of the absolute error (in years) for persistence across each state. Here, we see great differences in the POP and CNTR scenarios for heavily populated US states such as California, New York, and Texas. There are also some differences in persistence via the CNTR and POP scenarios with Mexico City.

In Fig. 10, we show a violin plot of the absolute error of the estimated migration rates compared with the respective RS values for each pdm09 scenario. We observe closer differences here than with H5N1. This is likely due to the choice of Mexico City as the state with the highest population. We

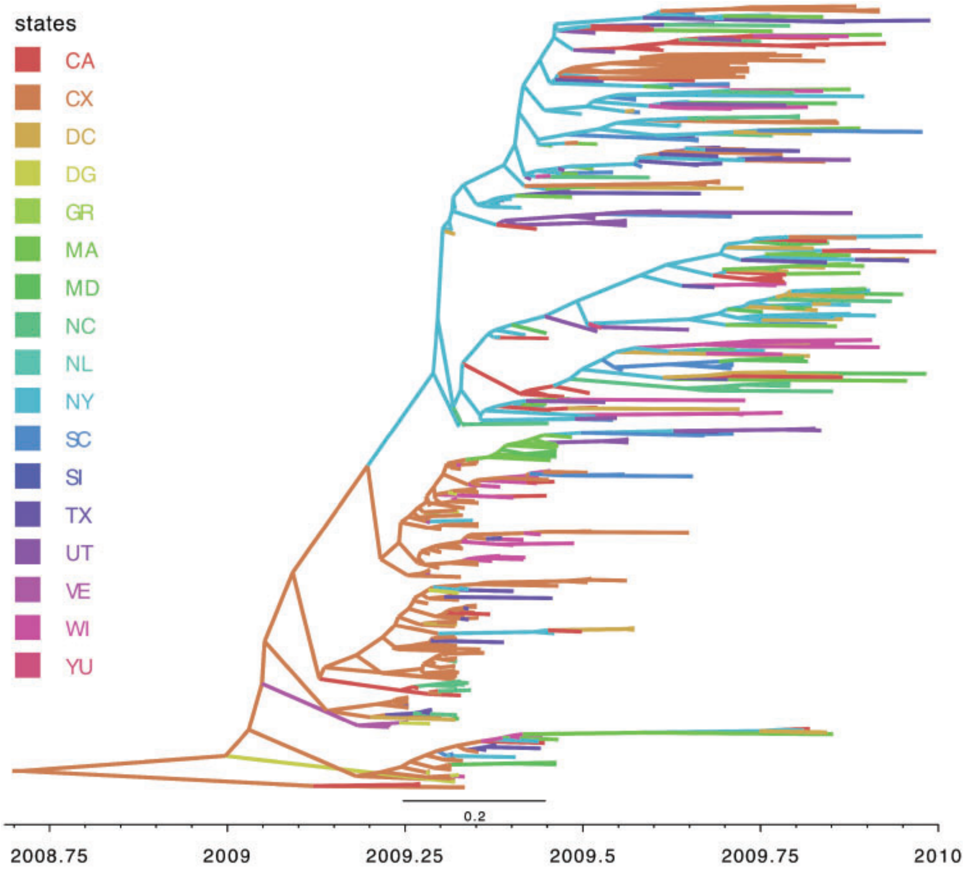

Figure 7. MCC tree of 443 pdm09 HA genes in Mexico and the United States with no sampling uncertainty (RS). We color the branches based on states with the greatest posterior probability. Abbreviations: CA, California; CX, Mexico City; DC, District of Columbia; DG, Durango; GR, Guerrero; MA, Massachusetts; MD, Maryland; NC, North Carolina; NL, Nuevo León; NY, New York; SC, South Carolina; SI, Sinaloa; TX, Texas; UT, Utah; VE, Veracruz; WI, Wisconsin; YU, Yucatán.

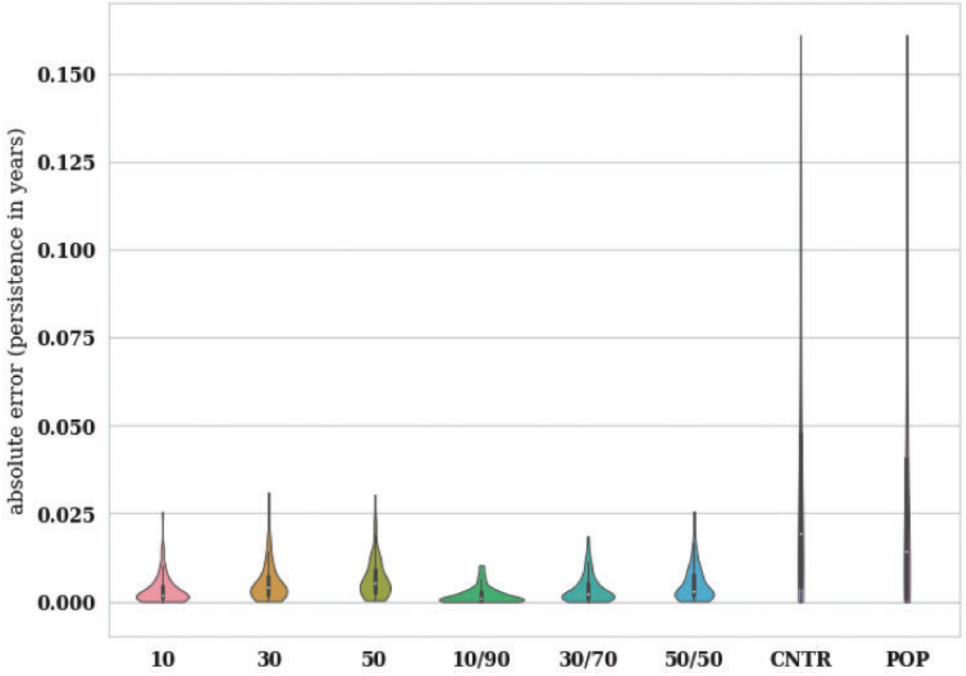

Figure 8. Absolute error of the estimated persistence times compared with the RS for each pdm09 scenario. Smaller errors are characterized by greater thickness of the violins close to zero. We included all five samples for each scenario.

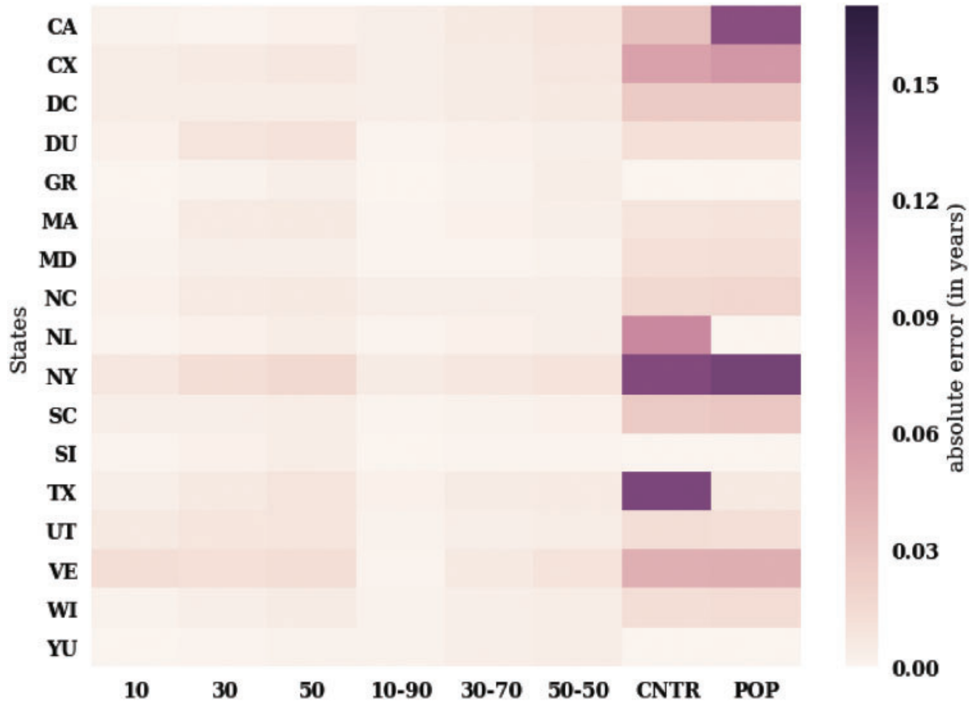

Figure 9. Absolute error of the estimated persistence times compared to the RS for each state and pdm09 scenario. Smaller errors are characterized by lighter colors. Here, we see the differences are greater for the CNTR and POP scenarios including US states California (POP scenario), New York (both CNTR and POP), and Texas (CNTR).

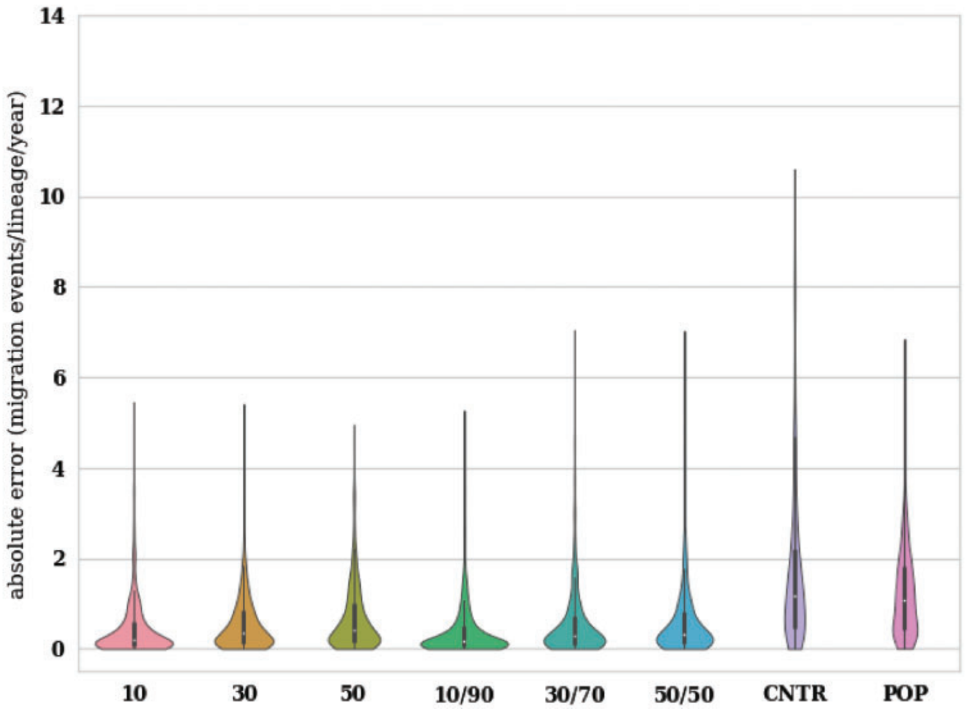

Figure 10. The absolute error of the estimated migration rates (per lineage per year) compared with the respective RS values for pdm09. We included all five samples for each pdm09 scenario.

identified Mexico City as a critical region for early circulation of new podm09 viruses and the eventual dispersion to the USA (Fig. 7). Thus, overrepresentation in the POP scenario enabled it to more closely align with the RS. However, the median absolute error of POP is 1.3, which is still larger than the other scenarios (range 0.2–0.48) except for CNTR. The same is true when examining individual states (Supplementary Fig. S10).

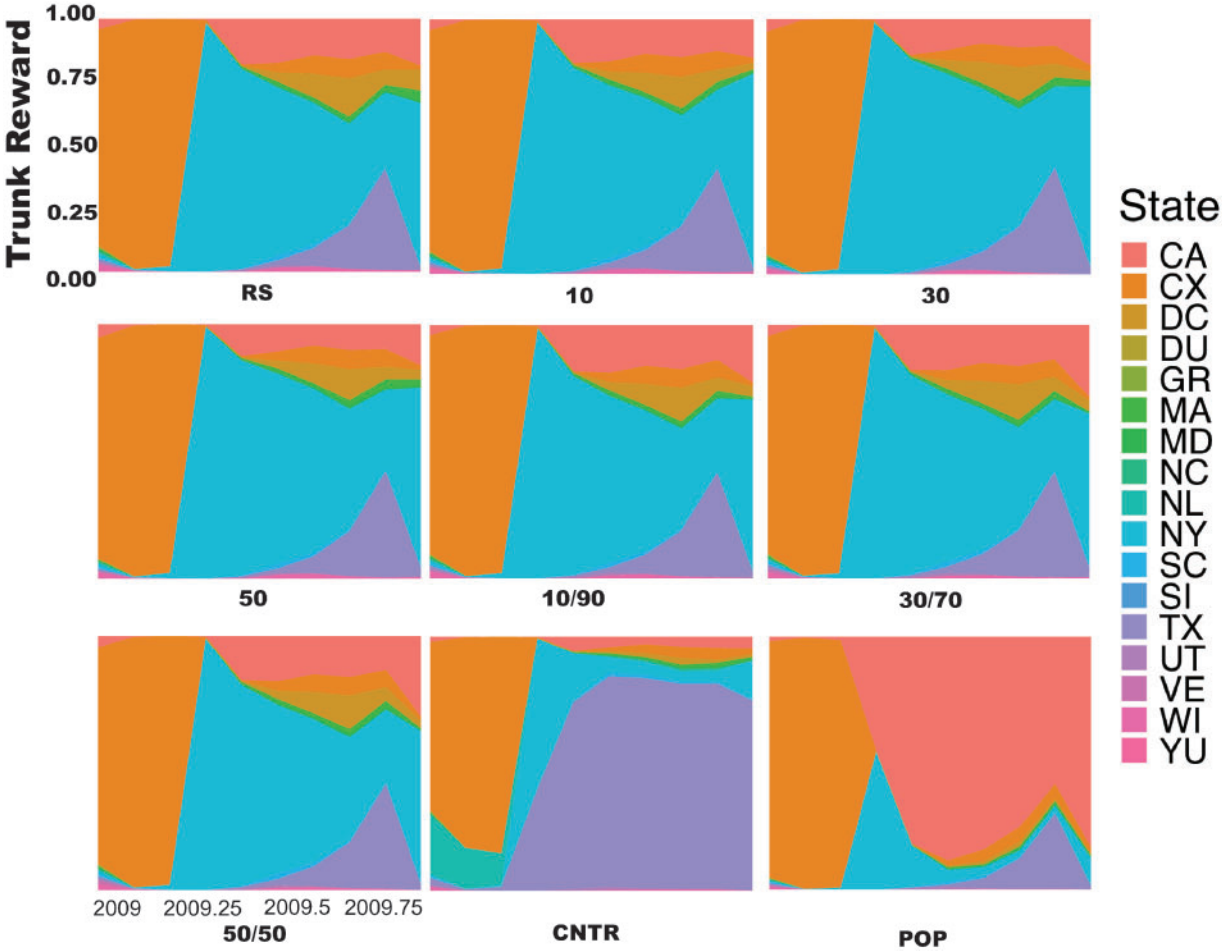

Figure 11. Trunk rewards for the reference (top left) and the eight scenarios for pdm09 in Mexico and the USA.

In Fig. 11, we show the trunk rewards for each scenario. Here, we see the dominance split between Mexico City at the beginning of the epidemic followed by New York as the virus spread to the USA. We do not observe this theme in the CNTR or POP scenarios which replace New York with Texas and California, respectively; suggesting that once dominance exists in the USA, the virus harbors mostly in the location selected by the scenario (either centroid or population). In Supplementary Table S3, we display the mean reward share over the time slices with differences from the standard deviation of the reference highlighted in red. CNTR and POP each contain the most differences from RS with seven and eight, respectively (although not as severe as the H5N1 example). In addition, these two scenarios had the largest differences with the RS for a given state including nearly 42 per cent for POP when considering California and 45 per cent for CNTR when considering Texas. All scenarios except for CNTR were within the RS's standard deviation of trunk reward share over time for Mexico City.

We show additional posterior estimates related to virus spread including the number of nonzero rates (Supplementary Fig. S11) and the complete list of routes with BFs > 100 (Supplementary Table S4).

4. Discussion

We evaluated the impact of sampling uncertainty in the Bayesian discrete phylogeography setting. When examining the posterior metrics, we found that scenarios with sampling uncertainty were closer to the RS than those that use scenarios either through the assignment of unknown locations to the centroid (CNTR) or the greatest population (POP). This included accuracy in determining the location of origin (root), local persistence, migration events, and trunk rewards of geographic states. This suggests that analysis with sampling uncertainty benefits phylogeographic estimates over ones that use a constant heuristic for unknown sampling locations. This might be particularly important for researchers wishing to study localized spread between cities or regions but do not have finite geospatial data. Several recent phylogeography or phylogenetic studies have explored virus evolution at this level (see (Holmes et al. 2011; Dibia et al. 2015; Pollett et al. 2015; Cerutti et al. 2016; Trewby et al. 2017)).

In our analysis, we considered distribution of the remaining sampling uncertainty either uniformly to candidate locations (scenarios *10*, *30*, *50*) or distributed to a single location (scenarios *10/90*, *30/70*, *50/50*). We found limited differences between these two techniques; thus, the choice by the researcher might depend on the number of realistic locations for a given record. For example, if there are only a few locations that are likely, the researcher might decide to use the split approach.

We recognize several limitations with our work. We utilized a discrete phylogeography approach in BEAST and thus we did not evaluate our approach on a continuous landscape. Prior work has highlighted the potential of continuous phylogeography under a relaxed random walk (Lemey et al. 2010) and a robust framework would need to consider the benefits for these types of scenarios. In addition, we implemented phylogeographic analysis via a BSSVS and did not explore the impact of different combinations of priors on the outcome of our statistics.

In addition, we note that our approach will produce BEAST analyses that include a larger number of K discrete locations. This could result in a failure of the Eigen decomposition because of the large rate matrix (Lemey 2014). In this environment, dimensionality increases by $O(K^{\circ}2)$. In general, issues normally start around $K > 20$–50 discrete traits depending on the data and other complexity issues. To circumvent this problem of larger discrete state spaces, one could estimate the log rates through a GLM where the regression part of the GLM is a function of a small (0–10 or so) number P of potential predictors. Here the number of estimable parameters reduces from $O(K^2)$ to $O(P)$.

In summary, phylogeography considers the sampling location of each taxon as fixed; often to a single discrete location. In this work, we relaxed this strong assumption and allowed for analytic integration of the uncertainty to evaluate the likelihood of the spatial process. The framework is now available in the new version of BEAST v.1.10 (Suchard et al. 2018). We used our Natural Language Processing (NLP) tool, GeoBoost (Tahsin et al. 2018), to extract geospatial locations found in the corresponding PubMed Central (PMC) article of a virus sequence record in GenBank and assign probabilities (the inverse of which is sampling uncertainty) for each candidate location (Tahsin et al. 2018).

We evaluated scenarios with sampling uncertainty to ones that assign a location to taxa using a predefined heuristic, either to the centroid of the study area or the most populated place. We compared all scenarios to a RS in which we are certain of all of the locations for all of taxa. Our scenarios with sampling uncertainty were closer to the RS across different posterior analysis than the centroid or population scenarios; suggesting that assigning uncertainty to taxa location benefits estimation of virus diffusion as compared to *ad-hoc* heuristics. We also found that there was limited difference between how the sampling uncertainty was assigned; that uniformly or split between two locations did not greatly impact posterior results. In future work, we will explore viruses beyond influenza A. We will also develop a web interface for researchers to use our language processing methods to find and assign uncertainty to alternative potential locations for virus phylogeography.

Data availability

Data available at doi: https://doi.org/10.5061/dryad.7jr85rj

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

The authors thank Research Computing at Arizona State University for access to the Saguaro high-performance computer as well as the NVIDIA Corporation for a GPU Grant. In addition, we thank Dr Justin Bahl and Dr Daniel Magee for their help with the trunk rewards analysis.

Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01AI117011 to M.S. and G.G. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: None declared.

References

Alkhamis, M. A., Moore, B. R., and Perez, A. M. (2015) ‘Phylodynamics of H5N1 Highly Pathogenic Avian Influenza in Europe, 2005–2010: Potential for Molecular Surveillance of New Outbreaks’, *Viruses*, 7: 3310–28.

Allicock, O. M. et al. (2012) ‘Phylogeography and Population Dynamics of Dengue Viruses in the Americas’, *Molecular Biology and Evolution*, 29: 1533–43.

Arafa, A. et al. (2016) ‘Phylodynamics of Avian Influenza Clade 2.2.1 H5N1 Viruses in Egypt’, *Virology Journal*, 13: 49.

Baillie, G. J. et al. (2012) ‘Evolutionary Dynamics of Local Pandemic H1N1/2009 Influenza Virus Lineages Revealed by Whole-Genome Analysis’, *Journal Virology*, 86: 11–8.

Beck, A. et al. (2013) ‘Phylogeographic Reconstruction of African Yellow Fever Virus Isolates Indicates Recent Simultaneous Dispersal into East and West Africa’, *PLoS Neglected Tropical Diseases*, 7: e1910.

Bedford, T. (2011) PACT <http://bedford.io/projects/PACT/> accessed Jul 14.

—— et al. (2015) ‘Global Circulation Patterns of Seasonal Influenza Viruses Vary with Antigenic Drift’, *Nature*, 523: 217–20.

Benson, D. A. et al. (2013) ‘GenBank’, *Nucleic Acids Research*, 41: D36–42.

Bielejec, F. et al. (2016) ‘SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes’, *Molecular Biology and Evolution*, 33: 2167–9.

Carrel, M. A. et al. (2010) ‘Spatiotemporal Structure of Molecular Evolution of H5N1 Highly Pathogenic Avian Influenza Viruses in Vietnam’, *PLoS One*, 5: e8631.

Cattoli, G. et al. (2011) ‘Evidence for Differing Evolutionary Dynamics of a/H5N1 Viruses among Countries Applying or Not Applying Avian Influenza Vaccination in Poultry’, *Vaccine*, 29: 9368–75.

Cerutti, F. et al. (2016) ‘Phylogeography, Phylodynamics and Transmission Chains of Bovine Viral Diarrhea Virus Subtype 1f in Northern Italy’, *Infection Genetics and Evolution*, 45: 262–7.

Dibia, I. N. et al. (2015) ‘Phylogeography of the Current Rabies Viruses in Indonesia’, *Journal of Veterinary Science*, 16: 459–66.

Drummond, A. J. et al. (2012) ‘Bayesian Phylogenetics with BEAUti and the BEAST 1.7’, *Molecular Biology and Evolution*, 29: 1969–73.

Dudas, G. et al. (2017) ‘Virus Genomes Reveal Factors That Spread and Sustained the Ebola Epidemic’, *Nature*, 544: 309–15.

Edgar, R. C. (2004) ‘MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput’, *Nucleic Acids Research*, 32: 1792–7.

Felsenstein, J. (1981) ‘Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach’, *Journal of Molecular Evolution*, 17: 368–76.

Fusaro, A. et al. (2010) ‘Evolutionary Dynamics of Multiple Sublineages of H5N1 Influenza Viruses in Nigeria from 2006 to 2008’, *Journal of Virology*, 84: 3239–47.

Gachara, G. et al. (2016) ‘Whole Genome Characterization of Human Influenza A(H1N1)pdm09 Viruses Isolated from Kenya during the 2009 Pandemic’, *Infection Genetics and Evolution*, 40: 98–103.

Garten, R. J. et al. (2009) *Influenza A Virus (A/California/04/2009(H1N1)) Segment 6 Neuraminidase (NA) Gene, Complete cds GenBank* <http://www.ncbi.nlm.nih.gov/nuccore/FJ966084.1> accessed Dec 15.

Gill, M. S. et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

Hasegawa, M., Kishino, H., and Yano, T. (1985) 'Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA', *Journal of Molecular Evolution*, 22: 160–74.

Hayman, D. T. et al. (2011) 'Evolutionary History of Rabies in Ghana', *PLoS Neglected Tropical Diseases*, 5: e1001.

Holmes, E. C. et al. (2011) 'Extensive Geographical Mixing of 2009 Human H1N1 Influenza a Virus in a Single University Community', *Journal of Virology*, 85: 6923–9.

Jeffreys, H. (1998) *Theory of Probability*, 3rd edn., Oxford Classic Texts in the Physical Sciences. Oxford & New York: Clarendon Press & Oxford University Press, x, 459 pp.

Kearse, M. et al. (2012) 'Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data', *Bioinformatics*, 28: 1647–9.

Lam, T. T. et al. (2012) 'Phylodynamics of H5N1 Avian Influenza Virus in Indonesia', *Molecular Ecology*, 21: 3062–77.

Lemey, P. (2014), 'Number of Discrete States in Phylogeo Model', in M. Scotch (ed.).

—— et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

—— et al. (2010) 'Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time', *Molecular Biology and Evolution*, 27: 1877–85.

——, Suchard, M., and Rambaut, A. (2009) 'Reconstructing the Initial Global Spread of a Human Influenza Pandemic: A Bayesian Spatial-temporal Model for the Global Spread of H1N1pdm', *PLoS Currents*, 1: RRN1031.

Liang, F., and Xiong, M. (2013) 'Bayesian Detection of Causal Rare Variants under Posterior Consistency', *PLoS One*, 8: e69633.

Lukashev, A. N. et al. (2016) 'Phylogeography of Crimean Congo Hemorrhagic Fever Virus', *PLoS One*, 11: e0166744.

Lycett, S. et al. (2012) 'Origin and Fate of a/H1N1 Influenza in Scotland during 2009', *The Journal of General Virology*, 93: 1253–60.

Mena, I. et al. (2016) 'Origins of the 2009 H1N1 Influenza Pandemic in Swine in Mexico', *Elife*, 5: e16777.

Naguib, M. M., Abdelwhab, E. M., and Harder, T. C. (2016) 'Evolutionary Features of Influenza a/H5N1 Virus Populations in Egypt: Poultry and Human Health Implications', *Archives of Virology*, 161: 1963–7.

Nelson, M. I. et al. (2011) 'Phylogeography of the Spring and Fall Waves of the H1N1/09 Pandemic Influenza Virus in the United States', *Journal of Virology*, 85: 828–34.

Pollett, S. et al. (2015) 'Phylogeography of Influenza A(H3N2) Virus in Peru, 2010–2012', *Emerging Infectious Diseases*, 21: 1330–8.

Pybus, O. G. et al. (2012) 'Unifying the Spatial Epidemiology and Molecular Evolution of Emerging Epidemics', *Proceedings of the National Academy of Sciences of the United States of America*, 109: 15066–71.

Rambaut, A. (2018) *FigTree* <http://tree.bio.ed.ac.uk/software/figtree/> accessed Jul 14.

——, and Holmes, E. (2009) 'The Early Molecular Epidemiology of the Swine-origin a/H1N1 Human Influenza Pandemic', *PLoS Currents*, 1: RRN1003.

—— et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (formerly Path-O-Gen)', *Virus Evolution*, 2: vew007.

Rambaut, A., et al. (2018), 'Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7', *Systematic Biology*, 67: 901-04.

Rao, D. M. (2014) 'Enhancing Epidemiological Analysis of Intercontinental Dispersion of H5N1 Viral Strains by Migratory Waterfowl Using Phylogeography', *BMC Proceedings*, 8: S1.

Scotch, M. (2018) *ZooPhy Online Portal* <https://zodo.asu.edu/zoophy/> accessed Aug 14.

—— et al. (2010) 'At the Intersection of Public-Health Informatics and Bioinformatics: Using Advanced Web Technologies for Phylogeography', *Epidemiology*, 21: 764–8.

—— et al. (2011) 'Enhancing Phylogeography by Improving Geographical Information from GenBank', *Journal of Biomedical Informatics*, 44: S44–7.

—— et al. (2013) 'Phylogeography of Influenza a H5N1 Clade 2.2.1.1 in Egypt', *BMC Genomics*, 14: 871.

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

Su, Y. C. et al. (2015) 'Phylodynamics of H1N1/2009 Influenza Reveals the Transition from Host Adaptation to Immune-Driven Selection', *Nature Communications*, 6: 7952.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4: vey016.

Tahsin, T. et al. (2014) 'Natural Language Processing Methods for Enhancing Geographic Metadata for Phylogeography of Zoonotic Viruses', *AMIA Joint Summits on Translational Science Proceedings*, 2014: 102–11.

—— et al. (2018) 'GeoBoost: Accelerating Research Involving the Geospatial Metadata of Virus GenBank Records', *Bioinformatics*, 34: 1606–8.

Tian, H. et al. (2017) 'Increasing Airline Travel May Facilitate Co-Circulation of Multiple Dengue Virus Serotypes in Asia', *PLoS Neglected Tropical Diseases*, 11: e0005694.

Trewby, H. et al. (2017) 'Processes Underlying Rabies Virus Incursions across US-Canada Border as Revealed by Whole-Genome Phylogeography', *Emerging Infectious Diseases*, 23: 1454–61.

Trovao, N. S. et al. (2015) 'Bayesian Inference Reveals Host-Specific Contributions to the Epidemic Expansion of Influenza a H5N1', *Molecular Biology and Evolution*, 32: 3264–75.

Wallace, R. G., and Fitch, W. M. (2008) 'Influenza a H5N1 Immigration Is Filtered out at Some International Borders', *PLoS One*, 3: e1697.

—— et al. (2007) 'A Statistical Phylogeography of Influenza a H5N1', *Proceedings of the National Academy of Sciences of the United States of America*, 104: 4473–8.

Wei, K., and Li, Y. (2018) 'Global Genetic Variation and Transmission Dynamics of H9N2 Avian Influenza Virus', *Transboundary and Emerging Diseases*, 65: 504–17.

WHO (2017), *Cumulative Number of Confirmed Human Cases for Avian Influenza A(H5N1) Reported to WHO, 2003–2017*, updated Oct 30 <http://www.who.int/influenza/human_animal_interface/2017_10_30_tableH5N1.pdf?ua=1> accessed Dec 15.

Yang, Z. (1994) 'Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods', *Journal of Molecular Evolution*, 39: 306–14.