Phylogenetics

Incorporating heterogeneous sampling probabilities in continuous phylogeographic inference – application to H5N1 spread in the Mekong region

Simon Dellicour[1,2,*], Philippe Lemey[1], Jean Artois[2], Tommy T. Lam[3], Alice Fusaro[4], Isabella Monne[4], Giovanni Cattoli[4,5], Dmitry Kuznetsov[6], Ioannis Xenarios[7], Gwenaelle Dauphin[8], Wantanee Kalpravidh[9], Sophie Von Dobschuetz[10], Filip Claes[9], Scott H. Newman[11], Marc A. Suchard[12,13,14], Guy Baele[1,†], Marius Gilbert[2,†]

[1] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[2] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP160/12 50, av. FD Roosevelt, 1050 Bruxelles, Belgium.

[3] State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Hong Kong SAR, China.

[4] Department of Comparative Biomedical Sciences, Istituto Zooprofilattico Sperimentale delle Venezie (IZSVe), Viale dell'Università 10, Legnaro, Italy.

[5] Animal Production and Health Laboratory, Joint FAO/IAEA Division, 2444 Seibersdorf, Austria.

[6] SIB, Swiss Institute of Bioinformatics, Lausanne, Switzerland.

[7] Center for Integrative Genomics, University of Lausanne, 1005 Lausanne, Switzerland.

[8] Ceva Santé Animale, 10 Avenue de la Ballastière, 33500 Libourne, France.

[9] Food and Agriculture Organization of the United Nations, Regional Office for Asia and the Pacific, Emergency Center of the Transboundary Animal Diseases, Bangkok 10200, Thailand.

[10] Food and Agriculture Organization of the United Nations, Headquarters, Rome, Italy.

[11] Food and Agriculture Organization of the United Nations, Regional Office for Africa, Accra, Ghana.

[12] Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.

[13] Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA, USA.

[14] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.

[†] These authors contributed equally to this work

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The potentially low precision associated with the geographic origin of sampled sequences represents an important limitation for spatially-explicit (i.e. continuous) phylogeographic inference of fast-evolving pathogens such as RNA viruses. A substantial proportion of publicly available sequences

are geo-referenced at broad spatial scale such as, for example, the administrative unit of origin rather than more exact locations (e.g. GPS coordinates). Most frequently, such sequences are either discarded prior to continuous phylogeographic inference or arbitrarily assigned to the geographic coordinates of the centroid of their administrative area of origin for lack of a better possibility.

Results: We here implement and describe a new approach that allows to incorporate heterogeneous prior sampling probabilities over a geographic area. External data, such as outbreak locations, are used to specify these prior sampling probabilities over a collection of sub-polygons. We apply this new method to the analysis of highly pathogenic avian influenza (HPAI) H5N1 clade data in the Mekong region. Our method allows to properly include, in continuous phylogeographic analyses, H5N1 sequences that are only associated with large administrative areas of origin and assign them with more accurate locations. Finally, we use continuous phylogeographic reconstructions to analyse the dispersal dynamics of different H5N1 clades and investigate the impact of environmental factors on lineage dispersal velocities.

Availability: Our new method allowing heterogeneous sampling priors for continuous phylogeographic inference is implemented in the open-source multi-platform software package BEAST 1.10.

Contact: simon.dellicour@ulb.ac.be

Supplementary information: Supplementary data are available at Bioinformatics online and on figshare.com.

1 Introduction

The lack of submission/publication of precise sampling locations associated with viral sequences limits the application of spatially-explicit phylogeographic analyses of fast evolving pathogens such as RNA viruses. Continuous phylogeographic inference (Lemey *et al*. 2010) requires sampled sequences to be associated with relatively precise geographic coordinates and as such, viral isolates/sequences with low sampling accuracy pose a concrete problem. Indeed, when studying the spread of a pathogen, a potentially large proportion of publicly available sequences is only associated with very large administrative units or even countries, as more precise data are either not traceable, lost or not deemed important at the time of sample collection or submission (Claes *et al*. 2014). If precise information exists, it may require contacting the original submitters or searching independent databases and/or as supplementary materials related to publications (Tahsin *et al*. 2017). Therefore, it is difficult to quantify the exact proportions of sequences submitted with more or less precise sampling origin such as country, administrative area, city, village or even precise geographic coordinates. However, for an important proportion - or even the majority - of viral sequences deposited in a database like GenBank, the level of precision of the sampling location does not go beyond the country or the first national administrative subdivision. In the present study, we use a classification by levels that refer to such administrative subdivisions. This classification is for instance used by the Database of Global Administrative Areas (GADM; gadm.org) and avoids using terminology that is country-specific (e.g. state, province, chiefdom, etc). Administrative subdivision of levels 1 and 2, hereafter referred as "admin-1" and "admin-2", are here respectively defined as the first and second administrative entities below the country level ("admin-0"), which would e.g. reflect states and counties in the USA or provinces and districts in Thailand.

Common practice entails that sequences associated with an imprecise sampling location, such as a country or a broad administrative area, are either discarded during the data collection step of a continuous phylogeographic inference (e.g. Holden *et al*. 2013) or assigned to the geographic coordinates of the centroid points of their administrative polygons of origin (e.g. Biek *et al*. 2007, Pybus *et al*. 2012). The former may lead to a non-negligible loss of (genetic) data whereas the latter may prove to be completely irrelevant given the very low or even unlikely probability that the sequence has been collected at that centroid location. Alternatively, it may be of interest to use polygons to define a prior range of sampling coordinates (Nylinder *et al*. 2014). This approach remains restricted to uniform sampling probabilities within polygons and may therefore only be relevant for relatively small administrative areas (e.g. admin-2 level). However, integrating unknown sampling coordinates over relatively broad administrative areas (e.g. of admin-1 or -0 level) in a uniform fashion may introduce undue uncertainty.

Geographic spread of avian influenza viruses has frequently been the focus of continuous phylogeographic analyses (e.g. Jin *et al*. 2014, Lu *et al*. 2014, Trovão *et al*. 2015). Among these viruses, the highly pathogenic avian influenza (HPAI) H5N1 virus of the Goose/Guangdong/96 lineage caused a panzootic of unprecedented proportions in poultry, in terms of number of outbreaks and animals affected, socio-economic impact and geographical range (Li *et al*. 2004, Kilpatrick *et al*. 2006, Gilbert *et al*. 2008). The HPAI H5N1 virus responsible for the panzootic was reported for the first time in live bird markets in the Chinese province of Guangdong in 1996 (Sims *et al*. 2005). The virus started spreading worldwide in 2003 and by 2008, it had extensively spread across the Eurasian and African continents. H5N1 currently persists in many countries such as Egypt, Indonesia, Vietnam and China (Domenech *et al*. 2009), but also regularly re-emerges in non-endemic countries. As for many other studies, an important proportion of the geo-referenced viral sequences of HPAI H5N1 available for the Mekong region are associated with imprecise sampling locations, which impeded applications of high-resolution continuous phylogeographic analysis to gain insights into the virus dissemination pattern and its drivers (Lemey *et al*. 2010). Therefore, the spatiotemporal reconstruction of H5N1 spread in the Mekong region represents an interesting example for investigating new approaches aiming to reduce the uncertainty related to sampling origin in the context of phylogeographic analyses.

In the present study, we aim to present and apply a new approach to address sampling location uncertainty. This method is based on the specification of heterogeneous and potentially null sampling probabilities associated with a series of sub-polygons and informed by external spatial data. For the specific case of HPAI H5N1 in the Mekong region, we have identified two data sets to inform the priors of the sampling location of viral

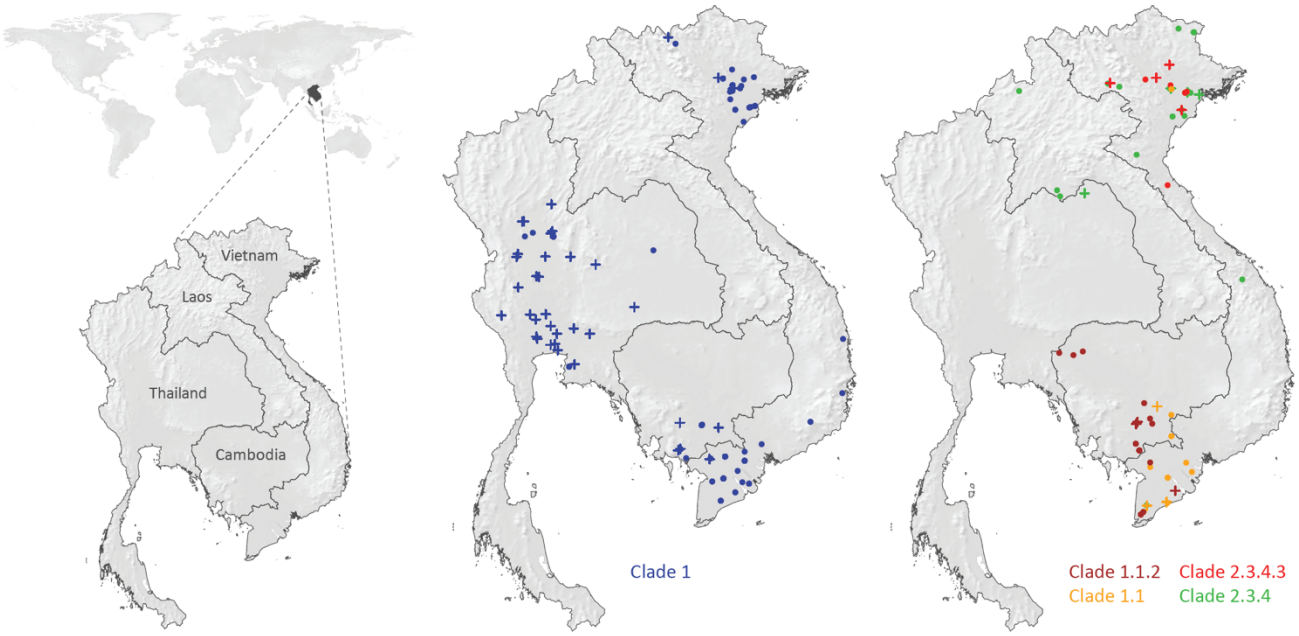

Fig. 1. Sampling maps of H5N1 clades in the Mekong region. Crosses and dots refer to sampling locations of H5N1 sequences assigned to an admin-1 or admin-2 polygon, respectively.

sequences: the distribution of HPAI H5N1 outbreaks from the EMPRES-i database (a global database of reported outbreaks in animals) and host (chickens, ducks) incidence data. We here focused on a data set of sequences from the Mekong region in the period 2003-2012. Poultry production and trade systems are very heterogeneously distributed in South-East Asia, resulting in a fairly heterogeneous distribution of H5N1 outbreak records (Gilbert *et al*. 2008, Pfeiffer *et al.* 2013). Therefore, this area is a very good study case to unravel the dispersal history of the virus using a continuous phylogeographic inference, as the results could help understanding the spatial dynamics of virus dispersal.

2 Methods

2.1 Compilation of H5N1 sequences data sets

We extracted HPAI H5N1 sequences (HA gene) from GenBank that originated from the Mekong region (Cambodia, Laos, Thailand, Vietnam) and that belong to some of the main clades circulating in these countries: clades 1, 1.1, 1.1.2, 2.3.4 and 2.3.4.3. Clade 1 and its subclades (e.g. clades 1.1 and 1.1.2) were identified in 2003 and were predominant in the Mekong region until 2012 (Cuong *et al*. 2016). As for clade 2.3.4 (and its subclade 2.3.4.3), it superseded clade 1 in northern Vietnam from 2005 to 2009 (Nguyen *et al*. 2008, Artois *et al*. 2016). The selection of these clades was based on the availability of associated metadata like sampling year and geographic origin, which are required for spatiotemporal reconstructions using continuous phylogeography. Further, we only extracted sequences for which at least an admin-1 area of origin was known, which was directly retrieved from publications, or from either the OpenFlu database managed by the Swiss Institute of Bioinformatics (openflu.vital-it.ch) or the EMPRES-i database developed and managed by the Food and Agriculture Organization of the United Nations (FAO, empres-i.fao.org). This led to the compilation of a data set consisting of 214 sequences from clade 1, 25 from clade 1.1, 30 from clade 1.1.2, 29 from clade 2.3.4, and 22 from clade 2.3.4.3 (Figure 1). Among these 320 sequences, 138 were associated with an admin-1 polygon of origin (hereafter referred to as "admin-1 sequences") and 182 were associated with an admin-2 polygon of origin (hereafter referred to as "admin-2 sequences"; Figure 1).

2.2 Constraining sampling uncertainty

BEAST (Bayesian Evolutionary Analysis by Sampling Trees) is an open-source multi-platform software package to perform Bayesian phylodynamic inference while accommodating phylogenetic uncertainty (Suchard *et al*. 2018). Now at version 1.10, BEAST allows to specify, for a given sequence, a polygon defining a uniform prior range of sampling coordinates (hereafter referred as the "uniform prior approach"; Nylinder *et al*. 2014). This prior can be specified using a polygon defined in a Keyhole Markup Language (KML) file, which is a file format used to display geographic information system (GIS) data in software packages such as Google Earth (www.google.com/earth).

To allow assigning non-uniform prior sampling probabilities, we extended this feature by allowing the specification of several non-overlapping sub-polygons, where each sub-polygon can be associated with a different sampling probability. Any given sequence can be associated with an external KML file, which can specify several sub-polygons linked to specific sampling probabilities, the sum of which is constrained to be equal to 1. We implemented this novel heterogeneous sampling prior approach (hereafter referred as the "heterogeneous prior approach") in BEAST 1.10 (Suchard *et al*. 2018) and an example of the related XML settings as well as edited KML files can be found as Supplementary Files along with Appendix S1.

In practice, each admin-2 sequence was associated with a single polygon of origin with a sampling probability equal to 1. Indeed, the size of admin-2 polygons typically remains relatively small at the scale of an entire study area on which we aim to infer the lineages dispersal history (Figure 2). At this geographical scale and for these sequences, there is little to be gained by further specifying areas associated with different sampling probabilities within an already small admin-2 polygon. Our proposed heterogeneous prior approach was only used for admin-1 sequences (Figure 2). For admin-1 sequences, simply considering the entire polygon of origin to define a uniform prior range of sampling coordinates would involve an important uncertainty on the sampling location. Overall, we here aimed to use more informative/constrained priors by assigning sampling probabilities to the admin-2 polygons nested within the admin-1 polygon of origin.

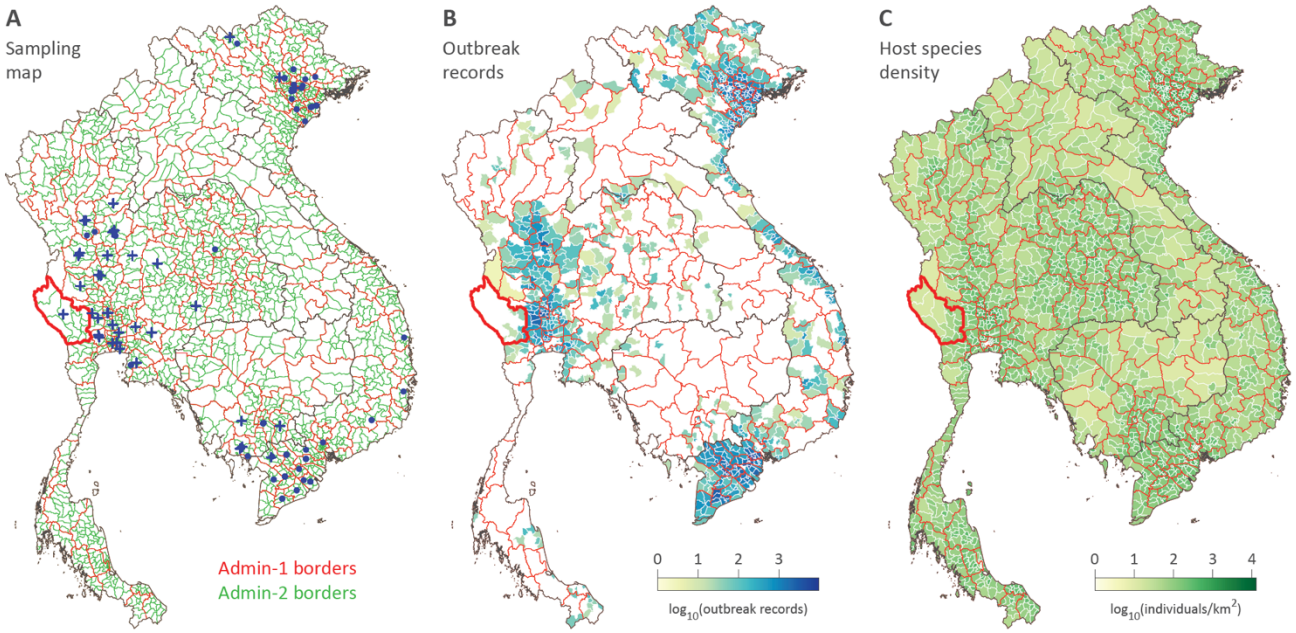

Fig. 2. External data considered to assign sampling probabilities to admin-2 polygons. A: sampling map of H5N1 clade 1 sequences. As in Figure 1, crosses and dots refer to sampling locations of H5N1 sequences assigned to an admin-1 or admin-2 polygon, respectively. **B**: map of outbreak records gathering data cumulated from 2004 to 2012 for admin-2 polygons (see also Figure S1 for annual maps). **C**: host species density map displaying, for each admin-2 polygon, log-transformed densities for both host species (chickens, ducks). While admin-2 polygons are coloured by hosts density for visual clarity, it is the actual total number of hosts that was considered to compute and assign a sampling probability to each of these admin-2 polygons. For illustrative purpose, the admin-1 area highlighted by the red contour is displayed with a larger size in Figure 3.

As detailed below, we used external data to define a sampling probability assigned to each admin-2 polygon.

2.3 Generating sub-polygons with a sampling probability

We employed our novel heterogeneous prior approach that avoids having to either specify large prior ranges of sampling coordinates or discard admin-1 sequences. In our application to HPAI H5N1, we used external data, such as annual outbreak records, to constrain the initial broad prior range of sampling coordinates. For a given admin-1 sequence, we first analysed the map of outbreak records corresponding to the sampling year of that sequence. Such annual maps gather the number of outbreak records per admin-2 polygon (Supplementary Figure S1). For each admin-2 polygon nested in the admin-1 polygon of origin, the assigned sampling probability was then estimated as the number of outbreak records in that particular admin-2 polygon divided by the total number of records in the overall admin-1 polygon. With this approach, we explicitly assumed that the probability that the sequence originated from a given admin-2 area is linearly proportional to the number of outbreaks recorded in that area. It is important to note (i) that the admin-2 sampling probabilities always sum to one, and (ii) that some admin-2 polygons can be associated with a zero-sampling probability (Figure 3). The rationale behind the former is that it is highly unlikely that the admin-1 sequence originates from an admin-2 area for which there is no outbreak record during its sampling year. In the absence of outbreak records, we could have used host species incidence data (for chickens and ducks; Figure 2) to define the sampling probabilities to assign to admin-2 polygons. However, this was not the case since we had outbreak records within all admin-1 areas for which we obtained sampled sequences. The detailed procedure is available in Appendix S1.

HPAI H5N1 outbreak records in the Mekong region were previously compiled and used in Artois *et al*. (2016). These records were extracted from the database of the Global Animal Health Information System of the FAO (EMPRES-i; http://empres-i.fao.org). The final data set is made up of 6762 outbreak records complemented with 338 records provided by the

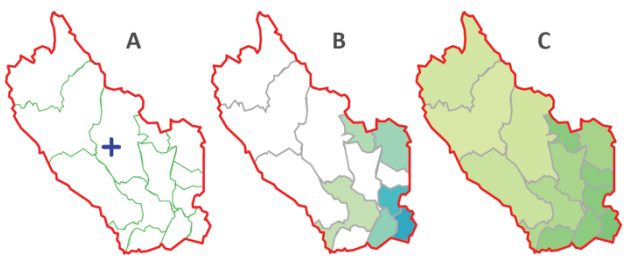

Fig. 3. Zoom on a given admin-1 polygon for which a sampled sequence is assigned. Admin-2 polygons are delimited by green borders (A) and further coloured by log-transformed outbreak records (B) and log-transformed host species densities (C, see Figure 2 for the respective colour scales). In this example, only 6 out of 13 admin-2 polygons are associated with a non-null number of outbreak records. If none of these admin-2 were associated with outbreak records, the host incidence data would have been used to assign a sampling probability to each admin-2 polygon.

Department of Animal Health (Hanoi, Vietnam). 98% of these outbreaks were geo-referenced at the administrative level 3 (commune), 0.5% at the administrative level 2 (district) and 1.5% at the administrative level 1 (province). The definition of an outbreak may have varied over time and by country, but in most cases, outbreaks were assumed to represent a farm, or a group of farms where HPAI H5N1 was observed at least once at a given point in time (Artois *et al*. 2016).

2.4 Continuous phylogeographic analyses

Continuous phylogeographic analyses were performed using the relaxed random walk (RRW) diffusion model implemented in BEAST (Lemey *et al*. 2010, Pybus *et al*. 2012), applied to each clade separately and using a Cauchy distribution to model among-branch heterogeneity in diffusion velocity. We also used the BEAGLE 3 library (Ayres *et al*. 2019) to improve computational performance. In addition, and for each analysis, we specified a non-parametric coalescent model as the tree topology prior (Gill *et al*. 2013), we modelled the substitution process according to the SRD06 parametrisation (Shapiro *et al*. 2006), and we specified a relaxed clock

model with rates drawn from an underlying lognormal distribution (Drummond *et al*. 2006). MCMC chains were run for 500 million iterations while sampling every 100,000 generations, and discarding the first 10% of the samples in each chain as burn-in. Finally, maximum clade credibility (MCC) trees were obtained with TreeAnnotator 1.10 (Suchard *et al*. 2018) and convergence and mixing properties were inspected using Tracer 1.7 (Rambaut *et al*. 2018). For comparative purposes, we performed each continuous phylogeographic analysis with the uniform prior as well as the new heterogeneous prior approach. In the former approach, uniform prior ranges of sampling locations were defined for all the sampled sequences, no matter if they were associated with an admin-1 or an admin-2 polygon of origin (Nylinder *et al*. 2014). For the latter approach, and as described above, prior ranges assigned to admin-1 sequences were defined with a collection of admin-2 polygons each associated with a distinct sampling probability (see the Appendix S1 for the practical details).

For each clade-specific phylogeographic analysis, we used the R package "seraphim" (Dellicour *et al*. 2016a, 2016b) to extract the spatio-temporal information embedded in 100 trees sampled from the posterior distribution (after burn-in had been removed). After this extraction step, each branch in the phylogeny can be treated as a distinct movement vector (Pybus *et al*. 2012) and we further used "seraphim" to estimate the mean branch dispersal velocity.

2.5 Comparing the uniform and heterogeneous prior approaches

To further test and compare the uniform and heterogeneous prior approaches, we performed additional phylogeographic analyses solely based on admin-2 sequences from clade 1 and considered different levels of sampling precision. Specifically, we performed three distinct phylogeographic analyses: (i) analyses using the uniform prior approach while considering only admin-1 level sampling locations for all sequences, (ii) analyses using the heterogeneous prior approach also considering only the admin-1 origin of each sequence, and (iii) analyses using the uniform prior approach but this time based on the original admin-2 sampling locations. In order to compare the spatial uncertainty associated with each phylogeographic inference, we reported the area of time-sliced polygons representing the 95% highest posterior density (HPD) regions computed for each successive year. In addition, we also used these analyses to compare the gain in accuracy, i.e. the increase in probability to position a given sequence in its actual admin-2 polygon of origin, when using the heterogeneous instead of the homogeneous prior approach.

Given that the accuracy gain assessment is only possible when relatively precise sampling locations are available, we also proposed an exploratory investigation of the relevance of the external data used to inform sampling priors. Specifically, we simulated 1,000 sets of sampling coordinates for the admin-2 sequences of clade 1 (but it could have been for any set of sampling points randomly distributed in sampled admin-1 polygons) following both the uniform and heterogeneous priors. We then compared the maximum and average gain in sampling location accuracy when considering heterogeneous rather than uniform priors of sampling coordinates. This assessment was performed once considering the past outbreak records and once considering the host incidence data to inform the heterogeneous priors.

2.6 Investigating the impact of environmental factors

Finally, we used the package "seraphim" to exploit phylogenetically-informed movement data for studying the association between particular environmental factors and the dispersal velocity of the H5N1 virus lineages in the Mekong region. Specifically, we investigated the impact of several factors: the main land cover variables for the study ("croplands", "forests", "savannas"), the elevation, the inaccessibility to major cities (quantified as the time it takes to travel to the nearest major city of >50,000 inhabitants), as well as human population and host species (chicken and duck) densities. All environmental factors were tested as potential conductance factors (i.e. factors facilitating movement) and as potential resistance factors (i.e. factors impeding movement). Correlations between phylogenetic branch durations and environmental distances were quantified as a statistic Q, which is computed as the difference between (i) the coefficient of determination obtained when branch durations are regressed against environmentally-scaled distances and (ii) the coefficient of determination obtained when branch durations are regressed against distances computed on a "null" raster, i.e. a raster with a value of "1" assigned to every cell. An environmental factor was only considered as potentially explanatory if both its distribution of regression coefficients and its associated distribution of Q values were positive (Jacquot *et al*. 2017). In a positive distribution of estimated Q values (i.e. with at least 90% of positive values), statistical support was then evaluated against a null distribution generated by a randomisation procedure and formalised using a Bayes factor (BF) value (Dellicour *et al*. 2017). The full procedure is detailed in Appendix S2.

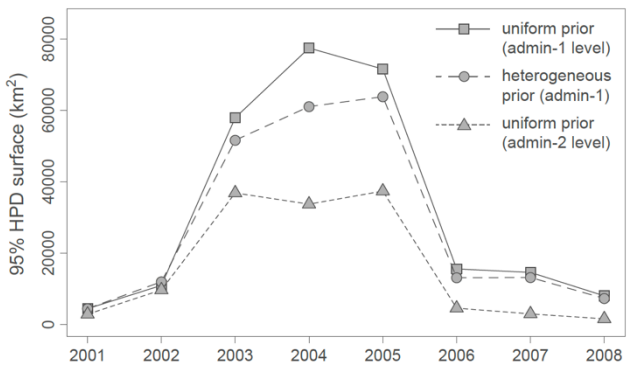

Fig. 4. Comparison of the phylogeographic uncertainty between the uniform and heterogeneous prior approaches. This comparison is based on three additional phylogeographic analyses solely considering admin-2 sequences from clade 1: (i) analyses using the uniform prior approach while considering only less precise admin-1 level locations for all sequences, (ii) analyses using the heterogeneous prior approach also considering the admin-1 origin of each sequence, and (iii) analyses using the uniform prior approach but this time based on the original admin-2 areas of origin. For each analysis, we report the area of time-sliced polygons representing the 95% highest posterior density (HPD) regions computed for each successive year.

3 Results

As detailed above, we used the continuous diffusion model implemented in BEAST (Lemey *et al*. 2010) to infer the dispersal history of the five H5N1 clades selected in this study. We performed a separate continuous phylogeographic inference for each clade. For the largest clade (i.e. clade 1, 214 sequences), we also performed these analyses both (i) using the uniform prior approach for all the sequences (no matter the precision level of their geographic origin) and (ii) using a combination of the uniform and heterogeneous prior approaches, respectively for the admin-2 and admin-1 sequences. The comparison of prior ranges of sampling coordinates generated in both cases is displayed in Supplementary Figure S2. This visual comparison shows that our heterogeneous prior approach leads naturally to a more heterogeneous sampling probability across admin-1 polygons. In addition, we also compared the uncertainty related to the continuous phylogeographic inference. In particular, we compared the area of the

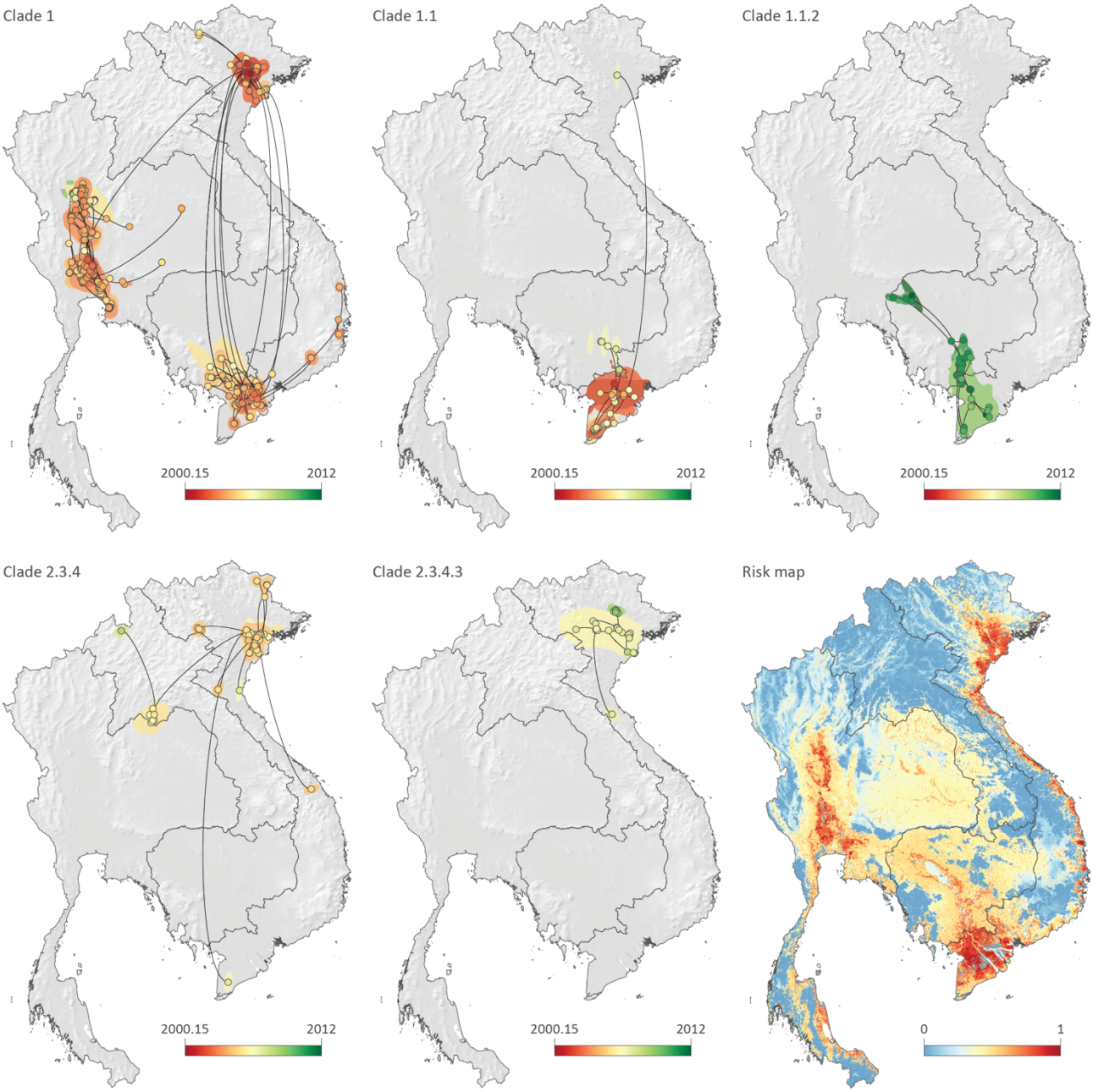

Fig. 5. Risk map and reconstructed spatiotemporal diffusion for each H5N1 clade considered in this study: mapped maximum clade credibility (MCC) trees and 95%highest posterior density (HPD) regions. MCC trees and 95% HPD regions are based on 100 trees subsampled from each post burn-in posterior distribution. Nodes of MCC trees are coloured according to their time of occurrence. 95% HPD regions were computed for successive time layers and then superimposed using the same colour scale reflecting time. The risk map comes from Gilbert *et al*. (2008).

time-sliced polygons representing the 95% HPD (highest posterior density) regions computed for each successive year (Supplementary Figure S3). Although this difference is not necessarily obvious when visually comparing both overall phylogeographic reconstructions, this comparison confirmed that the 95% HPD polygons tend to be smaller when using the heterogeneous prior approach (Supplementary Figure S3).

To further compare the uniform and heterogeneous prior approaches, we have also performed additional phylogeographic analyses only based on admin-2 sequences from clade 1. These analyses confirm that the heterogeneous approach results in a decrease in estimated spatial uncertainty relative to the uniform approach (Figure 4). Indeed, when admin-2 sequences are downscaled at the sampling precision of the admin-1 level, the heterogeneous approach allows reaching phylogeographic uncertainty in between the uncertainties obtained with the uniform prior approach at the admin-1 and admin-2 levels (Figure 4). Furthermore, we have also compared the gain in accuracy, i.e. to what extent the heterogeneous prior outperforms the uniform prior approach in estimating sampling coordinates in the actual admin-2 polygons of origin. This analysis reveals that using the heterogeneous instead of the uniform prior approach leads to an increase of up to 0.33 in the posterior probability to accurately position a given sequence in its actual admin-2 polygon of origin (on average, an increase in posterior probability equal to 0.11; 95% HPD = [0.00-0.22]). Although relatively moderate, this analysis formally demonstrates the gain

in sampling accuracy offered by the heterogeneous prior approach. Taken together, these evaluations of precision and accuracy gain confirm the utility of the heterogeneous prior approach when some sequences are associated with a large sampling area: it allows including such sequences while increasing the probability of accuracy and also avoiding integrating too much uncertainty in the continuous phylogeographic reconstruction.

In parallel, we also illustrate how to explore the potential interest of using a specific external data set to inform heterogeneous sampling priors. As detailed in the Methods section, we have simulated sampling coordinates following uniform as well as heterogeneous priors informed by the past outbreak record data or, alternatively, by the host incidence data. Again, we then compared the number of sampling coordinates accurately estimated in the actual admin-2 polygon of origin. When based on the past outbreak records, this analysis confirms the gain in accuracy obtained when using heterogeneous as compared to uniform priors, i.e. a similar increase up to 0.39 in the probability to accurately position a given sequence in its actual admin-2 polygon of origin (averaged increase of 0.13, 95% HPD = [0.00-0.31]). When based on host species incidence data, using the heterogeneous priors only leads to an average probability increase of 0.09 (95% HPD = [0.04-0.16]), which actually illustrates that this alternative data source results in a lower increase in sampling accuracy. These results thus further demonstrate the utility of using such a heterogeneous prior, as well as the importance of considering the appropriate external data to inform these priors.

Phylogeographic reconstructions based on the combination of the uniform prior and heterogeneous prior approaches reveal different dispersal histories among the five H5N1 clades considered in this study. While clades 1, 2.3.4 and 2.3.4.3 were primarily introduced in northern Vietnam, clades 1.1 and 1.1.2 rather appeared in the South of the Mekong region (in southern Vietnam and Cambodia, respectively). Additionally, phylogeographic reconstructions reveal a few long-distance lineage dispersal events for clades 1, 1.1 and 2.3.4. When compared to an H5N1 risk map built for the region (Figure 5), these long-distance dispersal events appear to have occurred through areas associated with lower risk of H5N1 occurrence. Such a risk map was previously generated by Gilbert *et al*. (2008) using an ecological niche modelling approach to statistically analyse the association between the recorded HPAI H5N1 virus presence and environmental factors. The comparison of mean branch dispersal velocities estimated for each clade further reveals differences in terms of dispersal dynamics: while we estimate a similar mean branch dispersal velocity for clades 1 (559 km/year 95% HPD [352, 1359]) and 2.3.4.2 (508 km/year 95% HPD [230, 1381]), a lower value is estimated for clade 1.1 (162 km/year 95% HPD [65, 473]), and higher values are estimated for clades 1.1.2 (901 km/year 95% HPD [502, 2793]) and 2.3.4 (1501 km/year 95% HPD [689, 5408]). Overall, these results highlight a dispersal velocity heterogeneity at the clade level.

Finally, we investigated whether heterogeneity in lineage dispersal velocity (among and within clades) could be related to a specific environmental factor. For this purpose, we used an analytical workflow already described in previous studies (e.g. Laenen *et al*. 2016, Dellicour *et al*. 2017) and implemented in the R package "seraphim" (Dellicour *et al*. 2016a, 2016b). Here, we investigated the impact of the following environmental factors (Supplementary Figure S4): elevation, land cover variables ("croplands", "forests", "savannas"), inaccessibility (quantified as the time it takes to travel to the nearest major city of >50,000 inhabitants), as well as human, chicken and duck population densities. All factors were tested as single potential conductance factors (i.e. factors facilitating movement) and as single potential resistance factors (i.e. factors impeding movement). Our analyses do not reveal any strong support (BF >20) for environmental factor acting as resistance or as conductance factors, i.e. significantly explaining the overall heterogeneity measured for the lineage dispersal velocity. However, a few environmental factors are associated with both a positive Q distribution and an approximated BF just above 3, which would correspond to a "positive" support according to the scale of interpretation of Kass & Raftery (1995; Supplementary Figure S5; see also Appendix S2 for the details of this analysis).

4 Discussion

The continuous phylogeographic reconstructions performed in this study reveal different dispersal histories between clades, but also heterogeneous dispersal velocities within clades. Yet, our assessment of environmental factors that might have impacted lineage dispersal velocity have not highlighted any factor that is significantly more correlated with dispersal durations than geographic distance alone. This overall result indicates that lineage dispersal velocity for H5N1 does not tend to uniformly vary according to particular environmental conditions. This result is also in line with the fact that poultry trade structure is known to be impacted by international borders (Pfeiffer *et al*. 2007, 2013). As we can see in Figure 5, continuous phylogeographic inferences reveal several long-distance dispersal events of lineages crossing the Mekong region. However, most of these events have occurred between the distant northern and southern Vietnam areas rather than between different countries. This result further emphasizes the importance of national and human-mediated movement of infected poultry within the trading network. The velocity of HPAI H5N1 virus spread does not seem to be notably associated with continuous environmental conditions that would impact the virus dispersal in wild populations.

The clade-specific continuous phylogeographic inferences performed in this study were based on equivalent proportions of admin-2 and admin-1 sequences. This illustrates that, for the empirical example we have analysed here, our heterogeneous prior approach allows to roughly double the amount of sequences associated with an informative prior range of sampling coordinates. As illustrated in Figure 3, assigning sampled sequences to the centroid point of a broad (e.g. admin-1) polygon can be irrelevant. In this example, the centroid point of the admin-1 polygon does not fall in an admin-2 polygon for which H5N1 outbreaks have been recorded at all. This provides a good illustration of a situation that motivates the development of the heterogeneous prior approach presented here. The second motivation was to avoid having to potentially discard an important number of valuable admin-1 sequences from the present analysis, which amount to 43% of our sequences data set. In the context of spatially-explicit phylogeographic analyses, there is a trade-off between the geographic precision of the study and the amount of genetic data that can be involved in the analysis. With the heterogeneous prior approach we propose here, we aim to increase the number of sequences that can be properly included in spatially continuous inferences by employing a prior range on the geographic origin as constrained as possible.

Acknowledgements

SD was supported by the *Fonds Wetenschappelijk Onderzoek* (FWO, Belgium) and is currently funded by the *Fonds National de la Recherche Scientifique* (FNRS, Belgium). JA and MG are supported by study grant from the National Institutes of Health (1R01AI101028-02A1). PL and MAS acknowledge funding from the Euro-

pean Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS) and from the Wellcome Trust Collaborative Award, 206298/Z/17/Z. PL acknowledges support by the Special Research Fund, KU Leuven (*Bijzonder Onderzoeksfonds*, KU Leuven, OT/14/115), and the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek - Vlaanderen*, G066215N, G0D5117N and G0B9317N). TL was supported by Theme-based Research Scheme (T11-705/14-N) from University Grants Committee of the HKSAR. MS acknowledges support from NIH R01 grant AI135995, NIH R01 grant AI117011 and NSF DMS grant 1264153. GB acknowledges support from the *Interne Fondsen* KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government - department EWI.

Conflict of Interest: none declared.

References

Artois J, Newman SH, Dhingra MS, Chaiban C, Linard C, Cattoli G, Monne I, Fusaro A, Xenarios I, Engler R, Liechti R, Kuznetsov D, Pham TL, Nguyen T, Pham VD, Castellan D, Von Dobschuetz S, Claes F, Dauphin G, Inui K, Gilbert M (2016). Clade-level spatial modelling of HPAI H5N1 dynamics in the Mekong region reveals new patterns and associations with agro-ecological factors. *Scientific Reports* **6**: 30316.

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA (2019). BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic Biology*, doi:10.1093/sysbio/syz020.

Biek R, Henderson JC, Waller LA, Rupprecht CE, Real LA (2007). A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 7993-7998.

Claes F, Kuznetsov D, Liechti R, Von Dobschuetz S, Truong BD, Gleizes A, Conversa D, Colonna A, Demaio E, Ramazzotto S, Larfaoui F, Pinto J, Le Mercier P, Xenarios I, Dauphin G (2014). The EMPRES-i genetic module: a novel tool linking epidemiological outbreak information and genetic characteristics of influenza viruses. *Database* **2014**: bau008.

Cuong NV, Truc VNT, Nhung NT, Thanh TT, Chieu TTB, Hieu TQ, Men NT, Mai HH, Chi HT, Boni MF, van Doorn HR, Thwaites GE, Carrique-Mas JJ, Hoa NT (2016). Highly pathogenic avian influenza virus A/H5N1 infection in vaccinated meat duck flocks in the Mekong Delta of Vietnam. *Transboundary & Emerging Diseases* **63**: 127-135.

Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG (2016). SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**: 3204-3206.

Dellicour S, Rose R, Faria NR, Vieira LFP, Bourhy H, Gilbert M, Lemey P, Pybus OG (2017). Using viral gene sequences to compare and explain the heterogeneous spatial dynamics of virus epidemics. *Molecular Biology & Evolution* **34**: 2563-2571.

Dellicour S, Rose R, Pybus OG (2016). Explaining the geographic spread of emerging epidemics: A framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics* **17**: 1-12.

Domenech J, Dauphin G, Rushton J, McGrane J, Lubroth J, Tripodi A, Gilbert J, Sims LD (2009). Experiences with vaccination in countries endemically infected with highly pathogenic avian influenza: The Food and Agriculture Organization perspective. *OIE Revue Scientifique et Technique* **28**: 293-305.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**: 699-710.

Gilbert M, Xiao X, Pfeiffer DU, Epprecht M, Boles S, Czarnecki C, Chaitaweesub P, Kalpravidh W, Minh PQ, Otte MJ, Martin V, Slingenbergh J (2008). Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 4769-4774.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA (2013). Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology & Evolution* **30**: 713-724.

Holden MTG, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, De Lencastre H, Skov R, Westh H, Žemličková H, Coombs G, Kearns AM, Hill RLR, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramírez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, Nübel U (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. *Genome Research* **23**: 653-664.

Jacquot M, Nomikou K, Palmarini M, Mertens P, Biek R (2017). Bluetongue virus spread in Europe is a consequence of climatic, landscape and vertebrate host factors as revealed by phylogeographic inference. *Proceedings of the Royal Society B: Biological Sciences* **284**: 20170919.

Jin Y, Yu D, Ren H, Yin Z, Huang Z, Hu M, Li B, Zhou W, Yue J, Liang L (2014). Phylogeography of avian influenza A H9N2 in China. *BMC Genomics* **15**: 1110-1110.

Kilpatrick AM, Chmura AA, Gibbons DW, Fleischer RC, Marra PP, Daszak P (2006). Predicting the global spread of H5N1 avian influenza. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 19368-19373.

Laenen L, Dellicour S, Vergote V, Nauwelaers I, De Coster S, Verbeeck I, Vanmechelen B, Lemey P, Maes P (2016). Spatio-temporal analysis of Nova virus, a divergent hantavirus circulating in the European mole in Belgium. *Molecular Ecology* **25**: 5994-6008.

Lemey P, Rambaut A, Welch JJ, Suchard MA (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology & Evolution* **27**: 1877-1885.

Li KS, Guan Y, Wang J, Smith GJD, Xu KM, Duan L, Rahardjo AP, Puthavathana P, Buranathai C, Nguyen TD, Estoepangestie ATS, Chaisingh A, Auewarakul P, Long HT, Hanh NTH, Webby RJ, Poon LLM, Chen H, Shortridge KF, Yuen KY, Webster RG, Peiris JSM (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**: 209-213.

Lu L, Lycett SJ, Leigh Brown AJ (2014). Determining the phylogenetic and phylogeographic origin of highly pathogenic avian influenza (H7N3) in Mexico. *PLoS One* **9**: e107330.

Nguyen TD, The Vinh N, Vijaykrishna D, Webster RG, Guan Y, Peiris JSM, Smith GJD (2008). Multiple sublineages of influenza A virus (H5N1), Vietnam, 2005−2007. *Emerging Infectious Diseases* **14**: 632-636.

Nylinder S, Lemey P, De Bruyn M, Suchard MA, Pfeil BE, Walsh N, Anderberg AA (2014). On the biogeography of *Centipeda*: a species-tree diffusion approach. *Systematic Biology* **63**: 178-191.

Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ (2007). An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. *The Veterinary Journal* **174**: 302-309.

Pfeiffer DU, Otte MJ, Roland-Holst D, Zilberman D (2013). A one health perspective on HPAI H5N1 in the Greater Mekong sub-region. *Comparative Immunology, Microbiology & Infectious Diseases* **36**: 309-319.

Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, Gray RR, Arinaminpathy N, Stramer SL, Busch MP, Delwart EL (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**: 15066-15071.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* **67**: 901-904.

Shapiro B, Rambaut A, Drummond AJ (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology & Evolution* **23**: 7-9.

Sims LD, Domenech J, Benigno C, Kahn S, Kamata A, Lubroth J, Martin V, Roeder P (2005). Origin and evolution of highly pathogenic H5N1 avian influenza in Asia. *Veterinary Record* **157**: 159-164.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**: vey016.

Tahsin T, Weissenbacher D, Jones-Shargani D, Magee D, Vaiente M, Gonzalez G, Scotch M (2017). Named entity linking of geospatial and host metadata in GenBank for advancing biomedical research. *Database* **2017**: bax093.

Trovão NS, Suchard MA, Baele G, Gilbert M, Lemey P (2015). Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1. *Molecular Biology & Evolution* **32**: 3264-3275.