Supplementary Information

Bayesian inference reveals host-specific contributions to the epidemic expansion of Influenza A H5N1

Nídia Sequeira Trovão*[,1], Marc A. Suchard[2,3], Guy Baele[1], Marius Gilbert[4,5], Philippe Lemey,[1]

[1]Department of Microbiology and Immunology, Rega Institute, KU Leuven. Leuven, Belgium

[2]Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California. Los Angeles, California, USA

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

[4]Biological Control and Spatial Ecology, Université Libre de Bruxelles, Brussels, Belgium

[5]Fonds National de la Recherche Scientifique, Brussels, Belgium.

∗ E-mail: Nidia.SequeiraTrovao@rega.kuleuven.be.

1 Supplementary Methods

1.1 Data collection

We applied two different subsampling procedures to the sequence data in order to mitigate sampling bias: 1) by randomly selecting 50 isolates for regions with a higher number of available sequences and 2) by selecting the same number of sequences based on phylogenetic diversity in those regions. For the latter, we used the Phylogenetic Diversity Analyzer tool (`www.cibiv.at/software/pda`) in order to select a subset that comprised the maximal phylogenetic diversity [1, 2]. Both down-sampling procedures were performed on the HA sequence set, resulting in data set of 806 sequences (Supplementary Table 1), and the corresponding NA subset was created by collecting the NA sequences from the same isolates included in the HA data set. In this supplementary information, we refer to the randomly downsampled data sets as HA_R and NA_R and the data sets subsampled by diversity as HA_D and NA_D for hemagglutinin and neuraminidase respectively.

We attempted to collect information about the wild or domestic status of the *Anatidae* hosts, but this was only available for about 20% of the sequences based on our database and literature search. By contacting the original authors, we were able to collect additional information for about 13% of the *Anatidae* sequences, but this remains insufficient to take this into account in our analysis.

1.2 Bayesian inference of sequences and traits

1.2.1 Sequence evolution

We performed Bayesian genealogical inference of time-measured trees using Markov chain Monte Carlo (MCMC) sampling implemented in the Bayesian Evolutionary Analysis Sampling Trees (BEAST) software [3]. We partitioned the coding genes into first+second and third codon positions and applied a separate Hasegawa-Kishino-Yano 85 (HKY85 [4]) substitution model with gamma-distributed rate variation among sites to both partitions [5]. We used an uncorrelated lognormal relaxed molecular clock to account for evolutionary rate variation among lineages [6] and specified a flexible Bayesian Skyride coalescent tree prior [7].

We ran and combined three independent MCMC analyses for 100 million generations, sampling every 10000th generation and removed 10% as chain burn-in. Stationarity and mixing was investigated using Tracer version 1.5, making sure that effective sample sizes for the continuous parameters were greater than 200.

Supplementary Table 3 compares posterior estimates for phylogenetic divergence times, evolutionary rates and statistics for the Bayesian discrete diffusion analysis (see below) of the different data sets. While the evolutionary rates are all close to 5 x 10^{-3} substitutions per site per year for the different data sets, the estimates for HA_D and NA_D are slightly higher compared to the randomly down-sampled data sets. This may be explained by the fact that the most diverse subset is selected by searching for the subset of taxa in a tree that maximises the sum of the branch lengths of the corresponding subtree [1, 2]. This will prefer longer terminal branches for samples drawn in the same year and consequently yield slightly faster divergence rates. The coefficient of variation for the evolutionary rate indicates somewhat higher substitution rate variability among lineages for NA as compared to HA. The time estimates for the most recent common ancestor are all close to the earliest sample in our data set (from 1996). The uncertainty for these dates will be slightly underestimated as only the sampling year - and not the exact sampling date, which was not available for all the sequences analysed here - was accommodated in our inference.

1.2.2 Discrete geography

To compare how the different data sets (HA_R, NA_R, HA_D and NA_D) inform our discrete state inference, we followed Lemey *et al.* (2009) [8] and calculated the Kullback-Leibler (KL) divergence [9] between a distribution with equal probabilities for each state and the posterior state distribution for each node, and sum this quantity across all internal nodes. A larger KL divergence value implies that the model extracts more information from the data. We also investigate how this compares to the extent of phylogenetic structure for a trait using the association index (AI) [10, 8]. This metric quantifies the degree to which the same traits (e.g. location or host) tend to cluster together relative to the expectation for randomised trait assignments. AI values close to 0 reflect strong phylogeny-trait correlation whereas AI values close to 1 reflect the absence of phylogenetic structure for the trait [10, 8]. Supplementary Table 3 lists these metrics or both the location and host traits in the different data sets. The HA data sets appear to yield higher KL divergences and therefore less uncertainty for the location estimates at the internal nodes than the NA data sets, and this is associated with a higher degree of phylogenetic clustering by trait as indicated by lower association indices. For HA, there is also less uncertainty and higher phylogeny-trait association for the randomly down-sampled data set. Interestingly, the same trend is reproduced by the host state estimates. Based on these results, we mainly restrict ourselves to reporting the HA_R and NA_R results in the main manuscript, but largely mirror them here with estimates for the other dataset sets.

In order to identify long-distance dispersal dynamics that significantly stand out from a more regular, distance-based diffusion process, we extended the recently developed generalized linear model (GLM) approach [11]. To this purpose, we introduce random effects (ϵ) in the GLM model such that every instantaneous movement rate Λ_{ij} for $i \neq j$ is parametrised as:

$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{ij1} + \ldots + \beta_P \delta_P x_{ijP} + \epsilon_{ij}, \tag{1}$$

where each predictor set $\mathbf{x}_p = \{x_{ijp}\}$ for $p = 1, \ldots, P$ potentially influences the rate from location i to j. We further group all effect sizes for the predictors into $\beta = (\beta_1, \ldots, \beta_P)'$, quantifying their contribution to Λ, and encode $(\delta_1, \ldots, \delta_P)$ as (0,1)-indicator variables that determine the inclusion or exclusion of the P predictors in the model. Importantly, random effects ϵ_{ij} for each entry in the instantaneous rate matrix account for the unexplained variability in the diffusion process when restricting its description to a number of potential predictors. We use this GLM diffusion approach extended here to accommodate random effects in order to identify exceptions to distance-based diffusion as a possible baseline spatial dispersal process.

To this aim, we incorporated log-transformed great-circle distances between each pair of locations as a predictor ($\mathbf{x}$) in our analyses. If geographic distances suffice in providing an adequate description of the diffusion process, all posterior estimates for the random effects are expected to be close to zero. We specify a 50% prior probability on the inclusion of the great-circle distance predictor and a normal prior distribution with a mean of zero and a standard deviation of 2 on its coefficient in log space (β). We specify a hierarchical normal prior distribution over the random effects (ϵ) with a mean of zero and an estimable precision, which we assume to be gamma-distributed with a shape and a rate of 0.001. This enables us to build a highly effective Gibbs sampler [12, 13] over the joint space of these random variables. Markov chain Monte Carlo (MCMC) transition kernels for the standard GLM-diffusion model parameters (β and δ) are described elsewhere [11].

We note that the introduction of random effects confronts us with the challenge to estimate a large amount of additional parameters, in our case $19 \times 18 = 342$. As is the case for the set of instantaneous transition rates in a standard discrete continuous-time Markov chain (CTMC) model, it is difficult to adequately inform these random effect parameters based on a single trait observation, and as a consequence, they will be subject to high variances. Therefore, in our posterior summary, we aim to focus on a limited number of random effects that are consistently high on an absolute (log) scale. We do this through the use of a posterior rank statistic that, for each effect, summarises the probability that the effect is the highest among all random effects:

$$\Pr(|\epsilon_{ij}| = \max_{0 \leq i,j \leq K} |\epsilon_{ij}|), \tag{2}$$

where $i \neq j$.

Using random effect estimates from the GLM model to identify exceptions to distance-based diffusion requires distances for each pair of location states to adequately represent geographic distances between all samples associated with these states. In other words, the approach will be most useful when location states represent a spatially coherent set of samples, which is not necessarily the case for location discretisation based on administrative borders we used for sub-sampling (Supplementary Table 1). In order to arrive at a spatially more coherent sampling for the same number of locations, we adjusted the discretisation according to a K-means identification of 19 clusters based on the geographic coordinates for all the sequences [14]. The K-means clustering was performed using R [15] and the resulting discretisation is represented in Supplementary Figure 1 and Supplementary Table 2.

1.3 Grid-based visualisation of continuous spatial diffusion

We developed a novel visualisation approach of continuous phylogeographic dispersal on a two-dimensional grid. To this purpose, we specify a grid composed of arbitrarily-sized cells covering the area of interest. For each branch in the posterior tree sample, we conditionally simulate its Brownian bridge representation of the random walk process at several time points along the branch. We identify which cells the process visits using Bresenham's line algorithm [16] to interpolate between points. Repeating this simulation over the entire sample approximates the posterior mass of occupancy with each grid cell that we visualise using a colour gradient or opacity. For the combined discrete host and continuous location diffusion, we summarise host-specific densities based on the host associated with the branch at the time a cell is visited as estimated by the complete Markov jump history. We also obtain similar visualisations for the diffusion rate by summarising cell-specific mean diffusion rates for the rates along the branches that visit the cells and representing them using heat map colouring.

2 Supplementary Figures

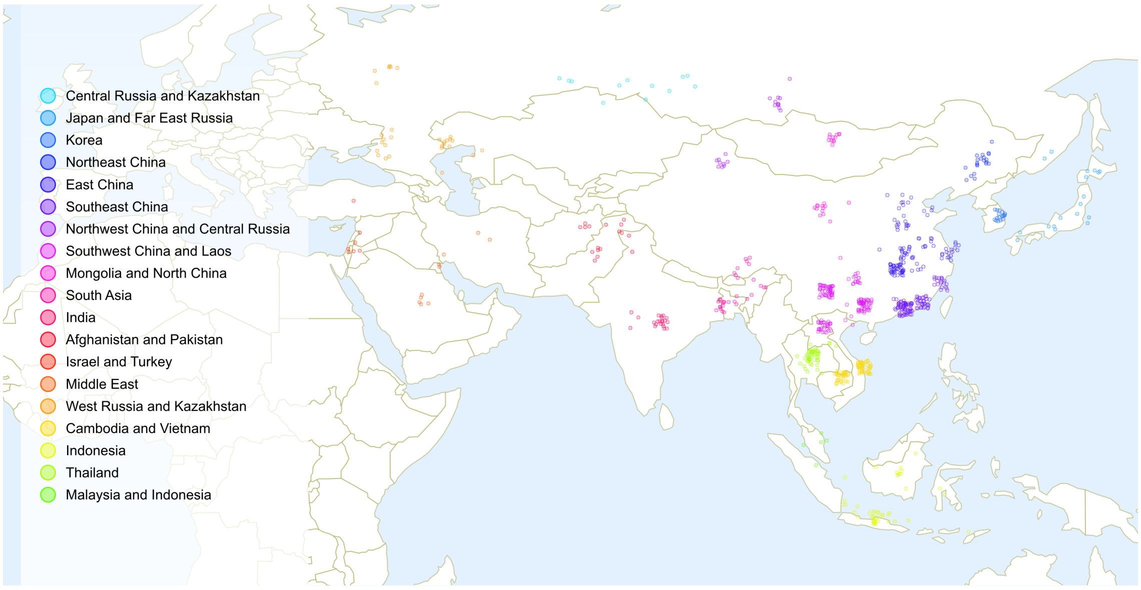


Figure 1: Sampling locations across Russia and Asia for the HPAIV H5N1 sequences we study here. The colours and the legend represent the 19 locations specified in the discrete phylogeographic analysis. The maps were taken from Natural Earth (www.naturalearthdata.com) and visualised using Cartographica (www.macgis.com).

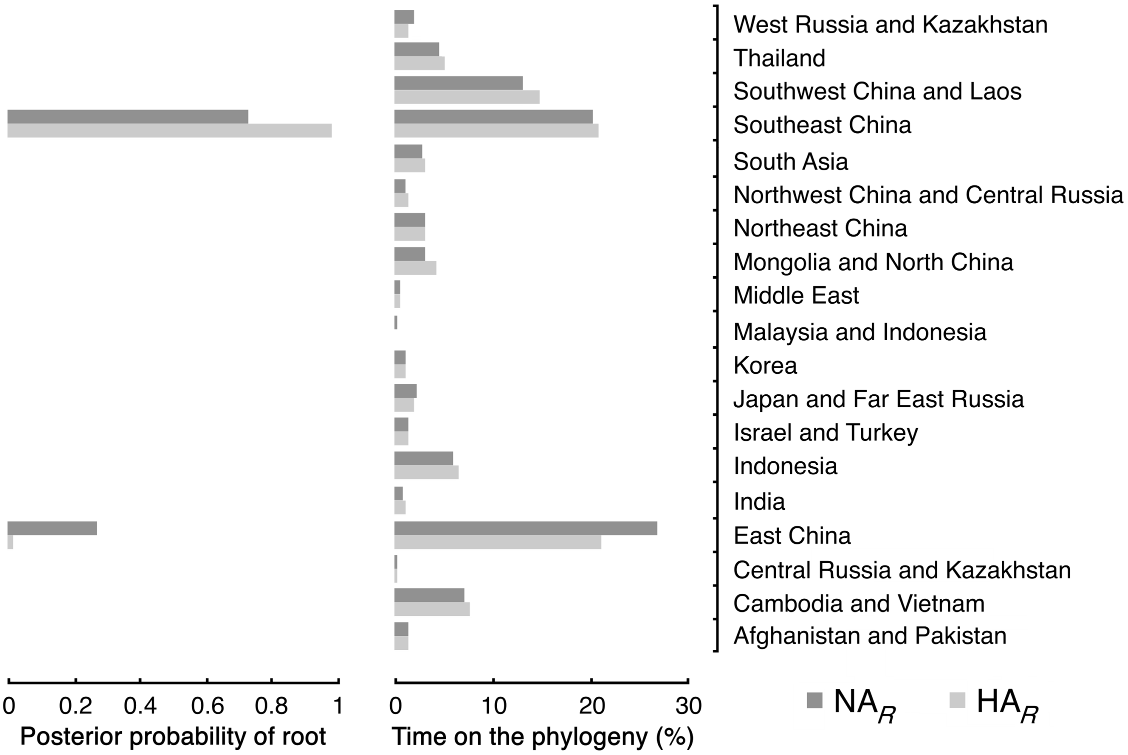


Figure 2: Posterior probabilities for location states at the root and Markov reward times in the discrete non-reversible phylogeographic analysis. Southeast China has ≈ 98% and 72% probability of being the origin location of HPAIV H5N1 for HA$_R$ and NA$_R$ (left bar chart). The amount of time spent in a particular location state throughout the HA$_R$ and NA$_R$ evolutionary histories is summarised from the Markov rewards (right bar chart).

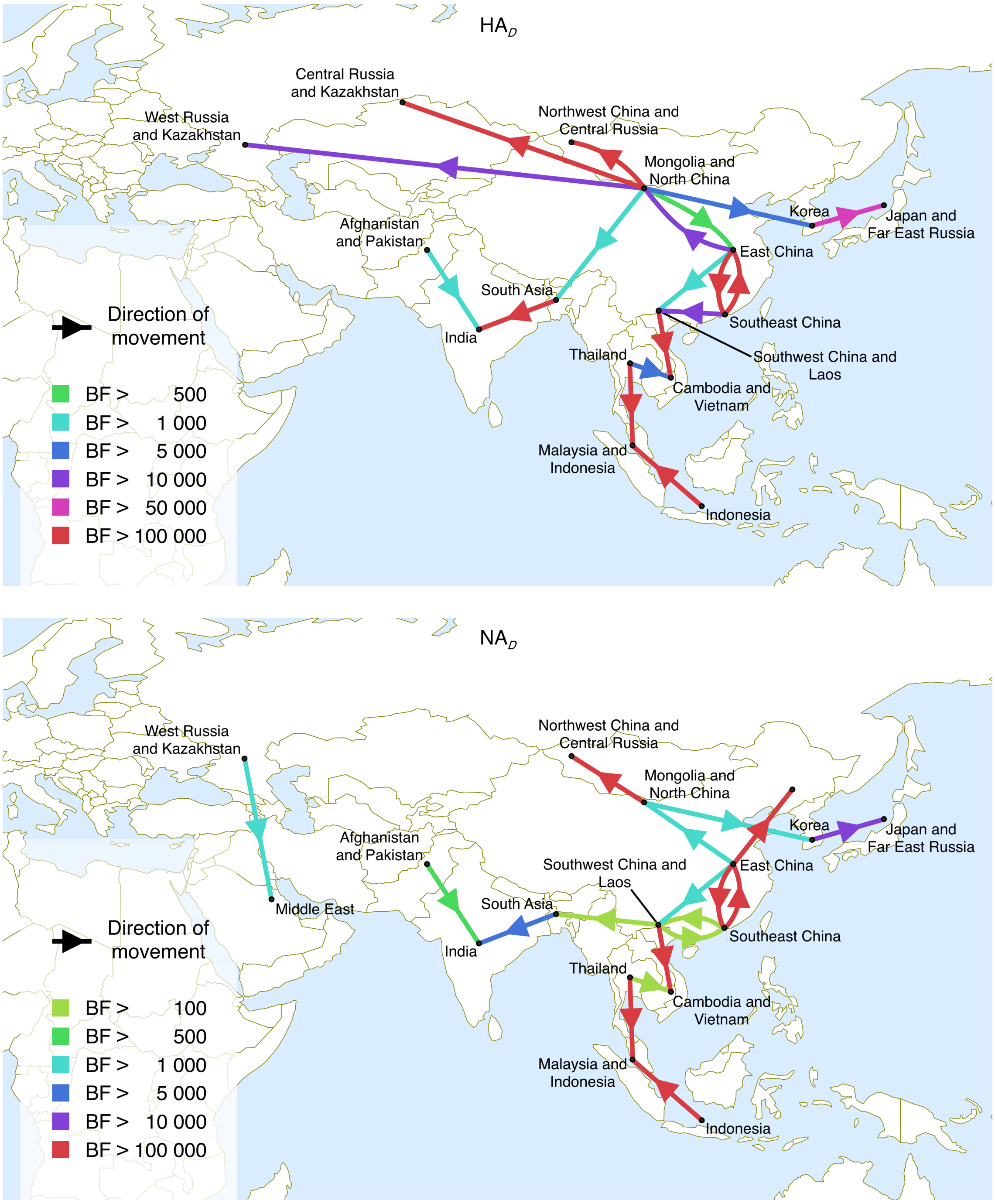
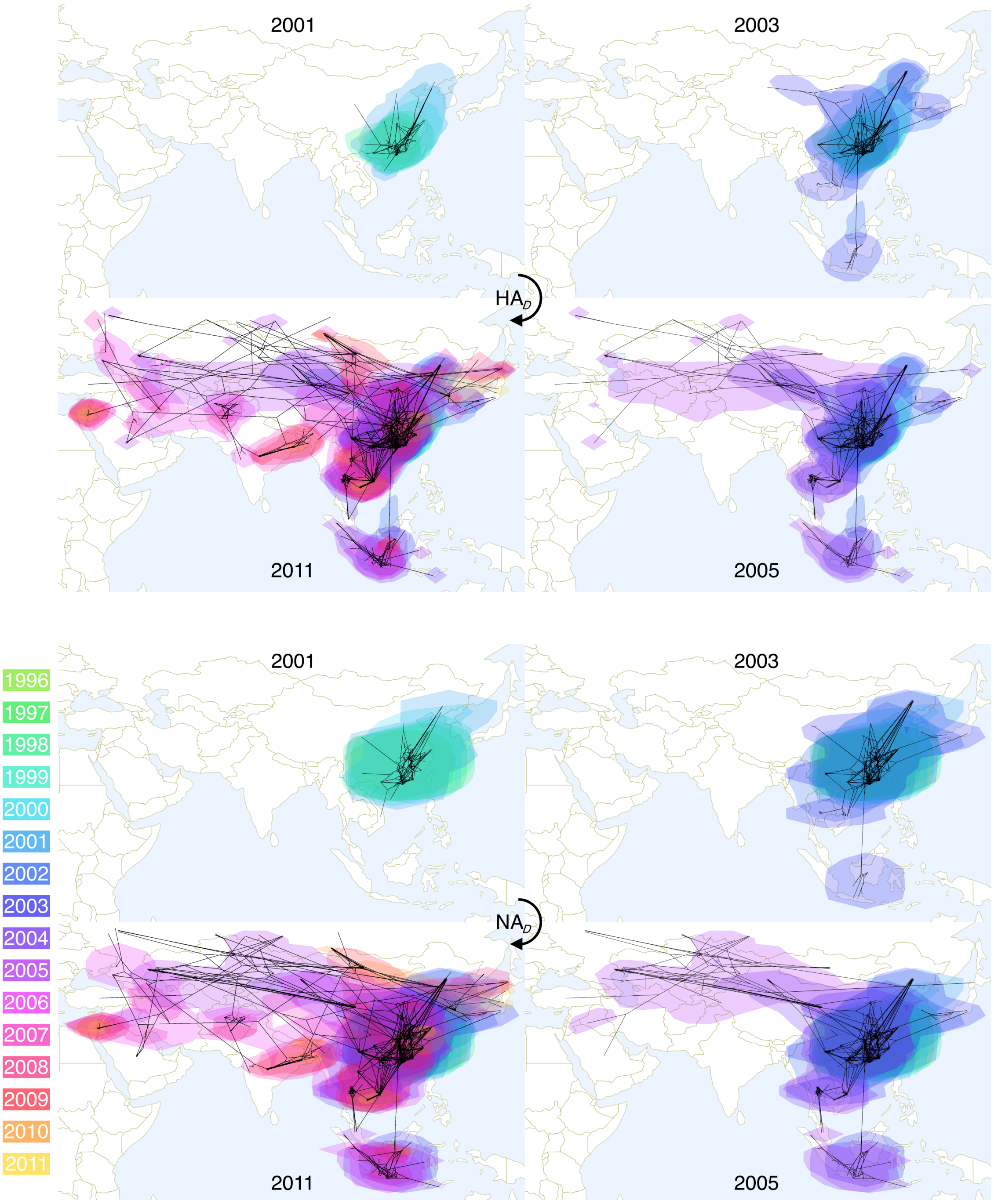


Figure 3: Bayes factor (BF) test for significant non-zero rates in HPAIV H5N1 HA$_D$ (top) and NA$_D$ (bottom). Only rates supported by a BF>100 are plotted. The line colour represents the relative strength by which the rates are supported: green lines and red lines suggest relatively weak and strong support respectively.


2001
2003
HA_D
2011
2005
2001
2003
NA_D
2011
2005
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011

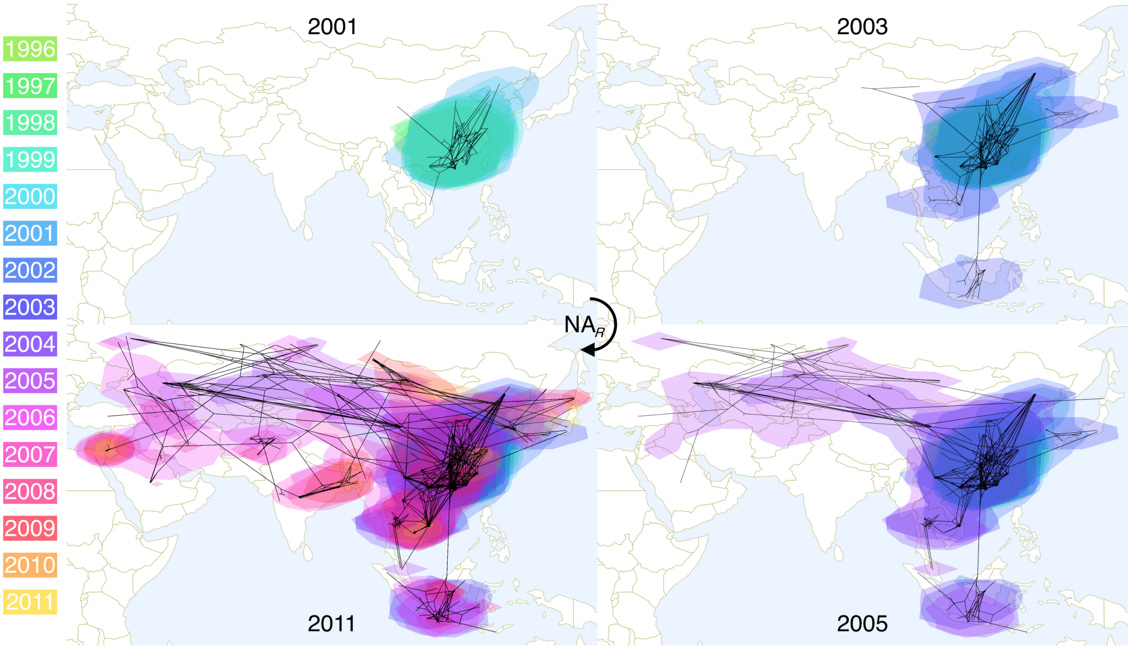


Figure 5: Reconstruction of the spatiotemporal dispersal of HPAIV H5N1 for HA$_D$ (top), NA$_D$ (middle) and NA$_R$ (bottom) throughout Russia and Asia, shown since 2001 onwards at intervals where major dispersal events occur: invasion of southeast Asia (2003), dispersal towards west Asia and invasion of Russia (2005) and spread to southwest Asia (2011). Black lines show a spatial projection of the representative phylogeny. Coloured clouds represent statistical uncertainty in the estimated locations of HPAIV H5N1 internal nodes (95% HPD intervals).

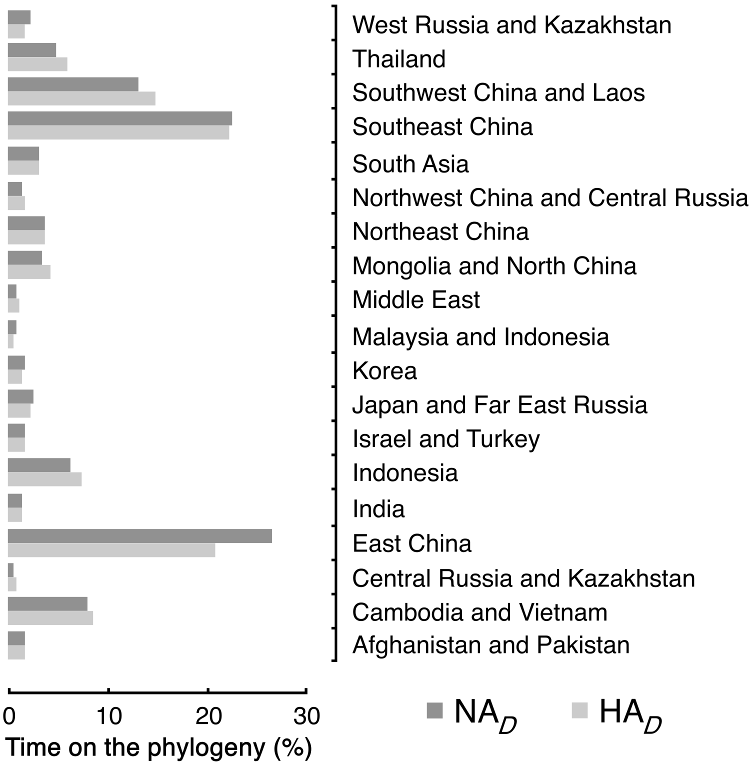


Figure 6: Percentage of total amount of time rewarded for each location state along the phylogenies of HA_D and NA_D.

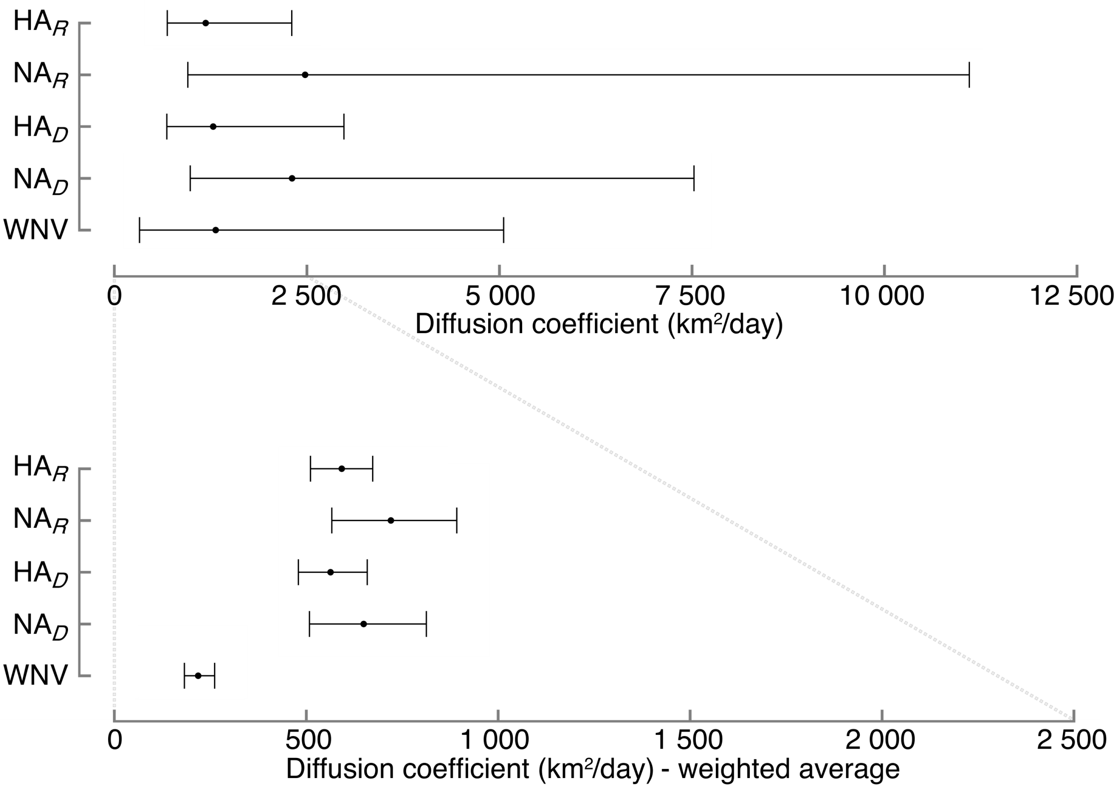


Figure 7: Standard diffusion coefficient (top) and weighted average diffusion coefficient (bottom) for both HPAIV H5N1 HA$_R$, NA$_R$, HA$_D$, NA$_D$ and West Nile virus in North America [17] (posterior mean and 95% HPD interval). Using a weighted average estimate for the diffusion coefficient, we arrive at consistently lower estimates, but importantly, much lower variance estimates as compared to the original diffusion coefficient statistic.

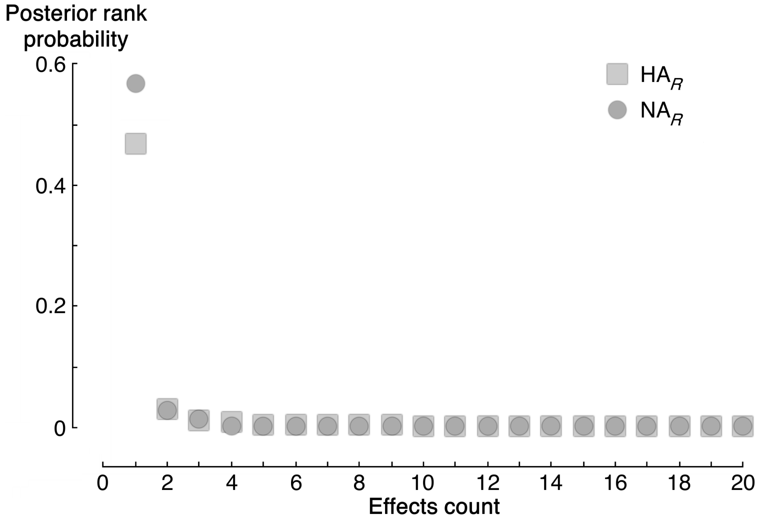


Figure 8: Posterior rank distribution for HA$_R$ and NA$_R$ for first 20 effects. One particular effect stands out as being more consistently ranked as the largest effect size on an absolute scale.

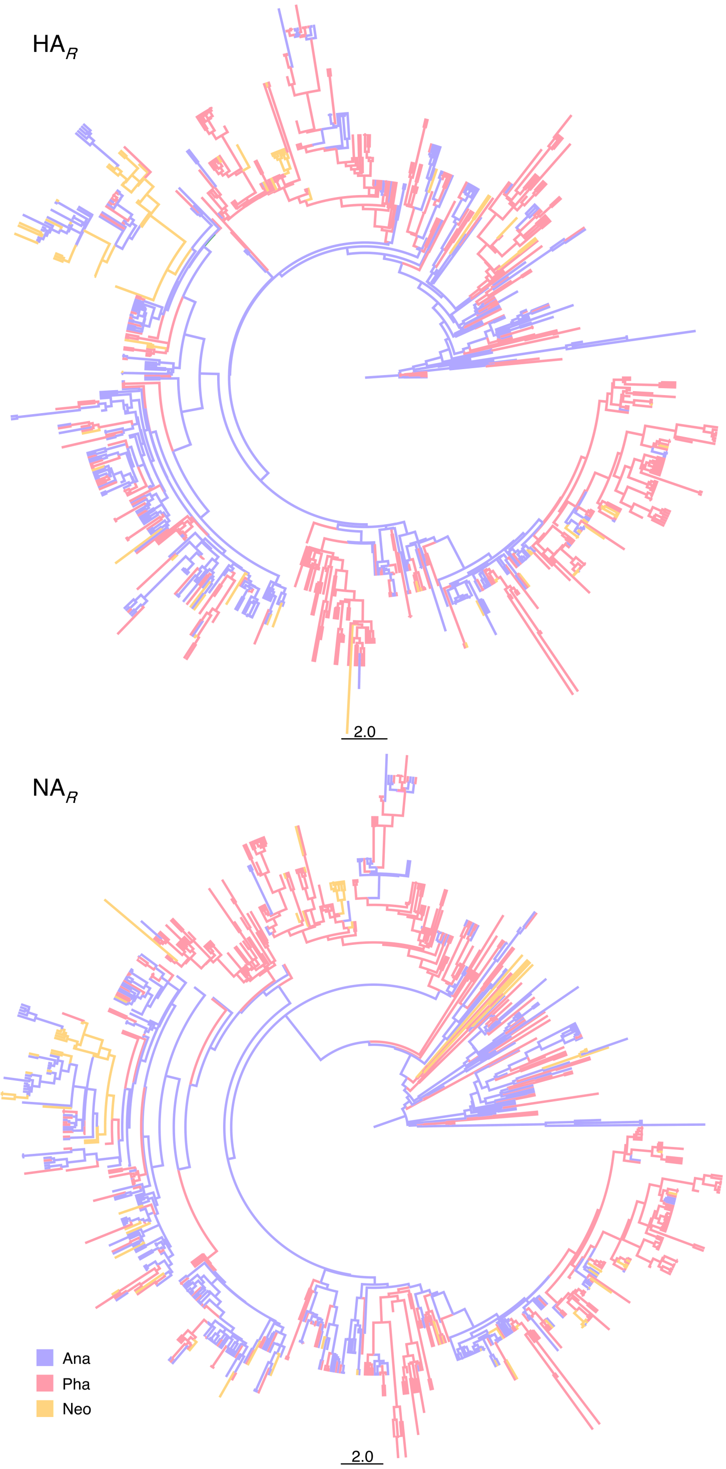


Figure 9: Time-calibrated maximum clade credibility tree inferred for 806 HA$_R$ and NA $_R$sequences sampled from 3 avian orders. Branches were coloured according to most probable host order, indicated in the coloured legend. Ana - *Anatidae*, Pha - *Phasianidae*, Neo - *Neoaves*.

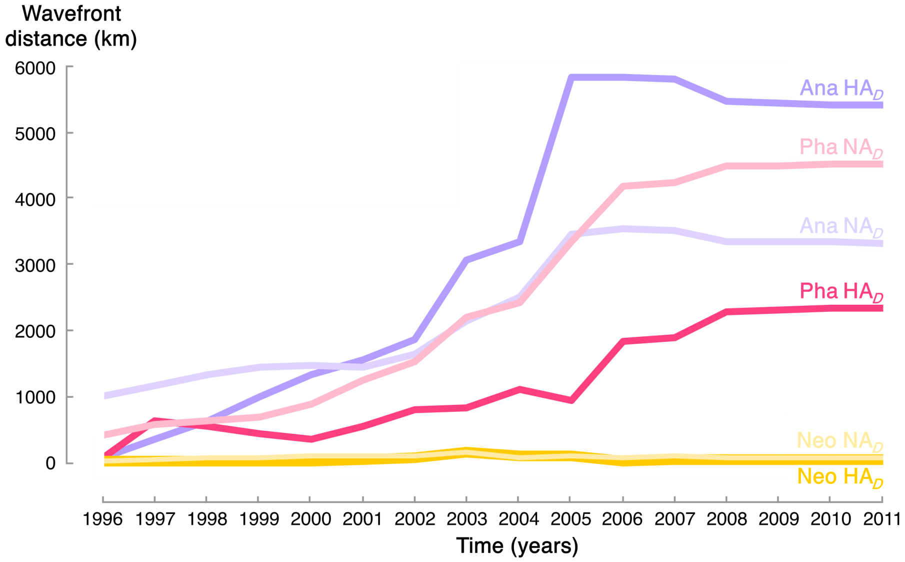


Figure 10: Host-specific wavefront distance estimates for HA_D and NA_D. These estimates summarise, for each host (*Anatidae* - Ana, *Phasianidae* - Pha and *Neoaves* - Neo), the fraction of estimated amount of great circle distance from the phylogeographic origin to the wavefront that can be associated with that host according to the host ancestral reconstruction.

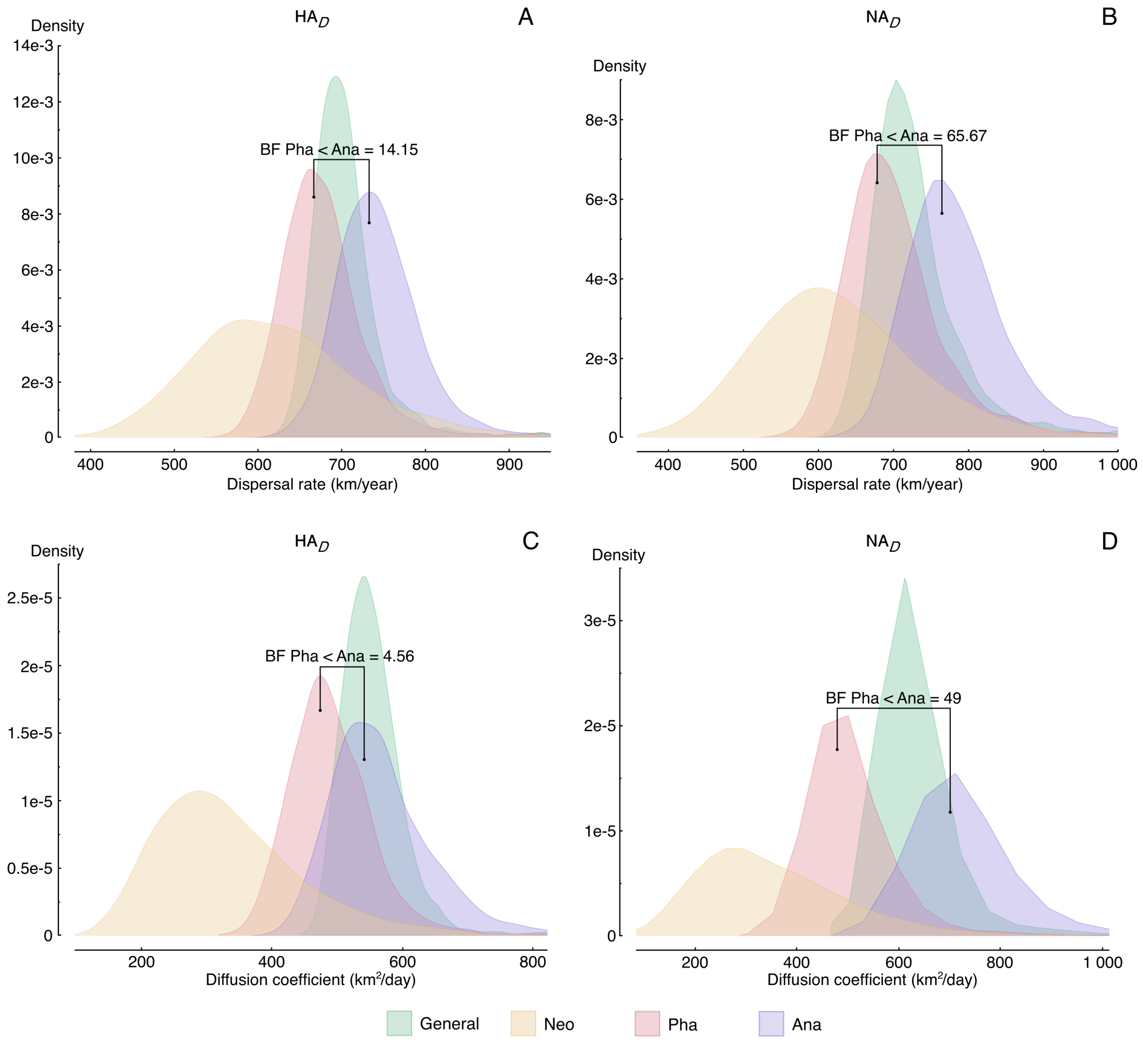


Figure 11: Posterior dispersal rate (top) and diffusion coefficient (bottom) distributions for each host (blue - *Anatidae*, red - *Phasianidae*, yellow - *Neoaves* and green - general representing the joint host analysis) in the analysis of HA_D (left) and NA_D (right)

3 Supplementary Tables

Table 1: Number of sequences by administrative region

Location	Number of sequences before/after down-sampling
Central-West Asia	38
Central China	36
Cambodia	25
West Russia	27
Guangdong	143 / 50
Guangxi	146 / 50
Hong Kong	252 / 50
Hunan	44
Indonesia	216 / 50
Japan	16
Korea	17
Laos	30
Mid-East China	26
Mid-West China	42
Mongolia	12
Northeast China	29
Southeast China	32
South Asia	49
Thailand	312 / 50
Vietnam	490 / 50
East Russia	23
Yunnan	106 / 50

Table 2: Number of sequences by K-Means location clustering

Location cluster	City	Administrative subdivision	Country	*Anatidae*	*Phasianidae*	*Neoaves*
Afghanistan and Pakistan	6	1	17	1	15	1
Cambodia and Vietnam	0	0	78	23	55	0
Central Russia and Kazakhstan	12	0	12	6	4	2
East China	10	111	121	65	48	7
India	3	0	22	1	21	0
Indonesia	18	21	46	1	44	1
Israel and Turkey	1	1	7	0	7	0
Japan and Far East Russia	6	11	18	6	5	6
Korea	0	0	17	8	9	0
Malaysia and Indonesia	2	0	5	0	5	0
Middle East	0	0	10	1	4	5
Mongolia and North China	0	13	25	21	2	2
Northeast China	0	21	21	2	17	2
Northwest China and Central Russia	1	19	20	2	10	8
South Asia	0	24	31	3	25	3
Southeast China	37	90	127	44	62	21
Southwest China and Laos	15	105	150	97	52	1
Thailand	1	26	50	4	32	14
West Russia and Kazakhstan	28	1	29	15	11	3

Table 3: Parameter estimates and posterior statistics for the Bayesian sequence and discrete trait phylogenetic inference applied to the HA and NA HPAIV H5N1 data sets.

	HA_R	NA_R	HA_D	NA_D
Date for the MRCA[1] (year)	1995.7 [1995.45 - 1995.98]	1994.69 [1994.24 - 1995.14]	1995.98 [1995.70 - 1996.28]	1994.64 [1994.14 - 1995.14]
Evolutionary rate (μ) (substitutions/ site/year)	5.13x10^{-3} [4.69x10^{-3} - 5.61x10^{-3}]	5.09x10^{-3} [4.56x10^{-3} - 5.61x10^{-3}]	5.33x10^{-3} [4.92x10^{-3} - 5.76x10^{-3}]	5.25x10^{-3} [4.72x10^{-3} - 5.82x10^{-3}]
Coefficient of variation for μ	0.78 [0.70 - 0.86]	1.14 [0.99 - 1.30]	0.71 [0.64 - 0.79]	1.18 [1.04 - 1.35]
Internal node location KL[2] divergence	2263.65	2222.71	2251.05	2223.04
Location association index	0.24 [0.22 - 0.26]	0.26 [0.24 - 0.28]	0.27 [0.25 - 0.29]	0.28 [0.25 - 0.30]
Internal node host KL divergence	765.55	736.15	759.04	724.73
Host association index	0.51 [0.47 - 0.54]	0.56 [0.52 - 0.61]	0.60 [0.55 - 0.64]	0.65 [0.60 - 0.70]

Values in between brackets represent 95% highest posterior density (HPD) intervals

[1] MRCA: most recent common ancestor

[2] KL: Kullback-Leibler

Table 4: Log marginal likelihoods estimated by stepping stone sampling for strict Brownian and different relaxed random walk (RRW) models

	HA_R	NA_R	HA_D	NA_D
Homogeneous (strict Brownian)	-5491.83	-5851.49	-5576.16	-5881.30
Cauchy RRW	-5123.00	-5395.54	-5199.13	-5428.73
Gamma RRW	-5376.62	-5753.08	-5454.14	-5784.51
Lognormal RRW	-5398.32	-5764.56	-5468.42	-5791.90

Table 5: Log random effect sizes with the highest posterior rank probability (first three) for HA$_R$, NA$_R$, HA$_D$ and NA$_D$

	Effects	Mean of effect	Posterior rank probability
HA$_R$	Mongolia and North China to West Russia and Kazakhstan	8.15 [2.93 - 13.53]	0.47
	Japan and Far East Russia to West Russia and Kazakhstan	2.84 [-5.31 - 9.86]	0.03
	Mongolia and North China to Middle East	1.21 [-5.76 - 8.65]	0.01
NA$_R$	Mongolia and North China to West Russia and Kazakhstan	9.47 [2.05 - 14.39]	0.57
	West Russia and Kazakhstan to Mongolia and North China	0.68 [-6.12 - 9.49]	0.03
	Japan and Far East Russia to West Russia and Kazakhstan	2.89 [-5.28 - 11.74]	0.01
HA$_D$	Mongolia and North China to West Russia and Kazakhstan	8.11 [5.80 - 10.81]	0.53
	Mongolia and North China to Middle East	1.34 [-4.46 - 8.69]	0.02
	Japan and Far East Russia to West Russia and Kazakhstan	2.51 [-4.06 - 9.14]	0.02
NA$_D$	Mongolia and North China to West Russia and Kazakhstan	9.75 [2.39 - 14.87]	0.56
	West Russia and Kazakhstan to Mongolia and North China	0.69 [-6.55 - 10.03]	0.03
	Japan and Far East Russia to West Russia and Kazakhstan	2.91 [-5.78 - 11.50]	0.01

Values in between brackets represent 95% HPD intervals

Table 6: Sequences classified by avian host

Host		Wild	Domestic	Unknown
Anatidae		61	36	203
Phasianidae		6	397	25
Neoaves	*Accipitriformes*	1	0	0
	Charadriiformes	4	0	0
	Ciconiiformes	12	0	0
	Columbiformes	0	0	3
	Falconiformes	12	0	0
	Gruiformes	3	0	0
	Passeriformes	25	0	0
	Pelecaniformes	8	0	0
	Podicipediformes	9	0	0

Two sequences are not listed in this table, one sampled from a *Diptera* host and the other from a *Struthioniformes* host, which were treated as unknown hosts.

Table 7: Relative host transition rates estimated from an asymmetric continuous-time Markov chain model

	HA_R				NA_R		
	Ana	Pha	Neo		Ana	Pha	Neo
Ana	———	3.59 [1.21 - 6.91]	0.65 [0.19 - 1.40]	Ana	———	3.81 [1.42 - 7.45]	0.75 [0.23 - 1.57]
Pha	1.18 [0.36 - 2.39]	———	0.69 [0.21 - 1.46]	Pha	1.39 [0.44 - 2.91]	———	0.67 [0.22 - 1.41]
Neo	0.29 [0.02 - 0.84]	0.16 [0.003 - 0.57]	———	Neo	0.27 [0.02 - 0.77]	0.16 [0.001 - 0.60]	———

Values in between brackets represent 95% HPD intervals

References

[1] Minh BQ, Klaere S, von Haeseler A. Phylogenetic diversity within seconds. Syst Biol. 2006 Oct;55(5):769–73.

[2] Minh BQ, Klaere S, von Haeseler A. Taxon Selection under Split Diversity. Syst Biol. 2009 Dec;58(6):586–94.

[3] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug;29(8):1969–73.

[4] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160–74.

[5] Shapiro B, Rambaut A, Drummond AJ. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol. 2006 Jan;23(1):7–9.

[6] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006 May;4(5):e88.

[7] Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol Biol Evol. 2008 Jul;25(7):1459–71.

[8] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol. 2009 Sep;5(9):e1000520.

[9] Kullback S, Leibler RA. On Information and Sufficiency. The Annals of Mathematical Statistics. 1951 March;22(1):79–86.

[10] Wang TH, Donaldson YK, Brettle RP, Bell JE, Simmonds P. Identification of shared populations of human immunodeficiency virus type 1 infecting microglia and tissue macrophages outside the central nervous system. J Virol. 2001 Dec;75(23):11686–99.

[11] Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Pathog. 2014 Feb;10(2):e1003932.

[12] Casella G, George EI. Explaining the Gibbs sampler. The American Statistician The American Statistician The American Statistician The American Statistician. 1992;46:167–174.

[13] Suchard MA, Kitchen CMR, Sinsheimer JS, Weiss RE. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst Biol. 2003 Oct;52(5):649–64.

[14] Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics. 1965;21:768–769.

[15] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2014. Available from: `http://www.R-project.org`.

[16] Bresenham JE. Algorithm for Computer Control of a Digital Plotter. IBM Syst J. 1965 Mar;4(1):25–30. Available from: `http://dx.doi.org/10.1147/sj.41.0025`.

[17] Pybus OG, Suchard MA, Lemey P, Bernardin FJ, Rambaut A, Crawford FW, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc Natl Acad Sci U S A. 2012 Sep;109(37):15066–71.