VIRUS EVOLUTION

Virus Evolution, 2022, **8(1)**, 1–5
DOI: https://doi.org/10.1093/ve/veac041
Advance access publication date: 18 May 2022
Review Article

Accommodating sampling location uncertainty in continuous phylogeography

SimonDellicour,[1,2,*,†] PhilippeLemey,[2,‡] Marc A.Suchard,[3,§] MariusGilbert,[1] and GuyBaele[2,**]

[1]Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP160/12, 50 av. FD Roosevelt, Bruxelles 1050, Belgium, [2]Department of Microbiology, Immunology and Transplantation, Laboratory of Clinical and Epidemiological Virology, Rega Institute, KU Leuven, Herestraat 49, Leuven 3000, Belgium and [3]Department of Biostatistics, Fielding School of Public Health, and Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA 90095-1766, USA
[†]https://orcid.org/0000-0001-9558-1052
[‡]https://orcid.org/0000-0003-2826-5353
[§]https://orcid.org/0000-0001-9818-479X
[**]https://orcid.org/0000-0002-1915-7732
*Corresponding author: E-mail: simon.dellicour@ulb.be

Abstract

Phylogeographic inference of the dispersal history of viral lineages offers key opportunities to tackle epidemiological questions about the spread of fast-evolving pathogens across human, animal and plant populations. In continuous space, i.e. when locations are specified by longitude and latitude, these reconstructions are however often limited by the availability or accessibility of precise sampling locations required for such spatially explicit analyses. We here review the different approaches that can be considered when genomic sequences are associated with a geographic area of sampling instead of precise coordinates. In particular, we describe and compare the approaches to define homogeneous and heterogeneous prior ranges of sampling coordinates.

Key words: virus; host species; continuous phylogeography; sampling precision; Bayesian inference; BEAST.

1. Introduction

Over the past decade, Bayesian phylogeographic inference methods have become popular approaches to reconstruct the dispersal history and dynamics of fast-evolving pathogens. Many examples exist for RNA viruses circulating in human (Faria et al. 2017; Zeller et al. 2021), animal (Torres et al. 2014; Duchatel, Bronsvoort, and Lycett 2019), and plant (Trovão et al. 2015; Kim et al. 2018) populations. Beyond descriptive epidemiological aspects, phylogeographic reconstructions have also been used to test hypotheses about the mode and tempo of viral spread. For instance, phylogeographic approaches have been used to test the importance of climatic, landscape, and host-related factors affecting bluetongue virus diffusion across Europe (Jacquot et al. 2017), rabies virus circulation in Tanzania (Brunker et al. 2018), and the dispersal dynamics of West Nile virus lineages in North America (Dellicour et al. 2020a). Recently, Guinat et al. also employed a phylogeographic approach to analyse several predictors of avian influenza H5N8 virus spread between poultry farms (Guinat et al. 2021).

The three most popular model-based phylogeographic approaches include (1) discrete phylogeographic inference using a continuous-time Markov chain model (Lemey et al. 2009) and inferring lineage transition events among discrete sampling locations, (2) inference using structured coalescent models (De Maio et al. 2015; Müller, Rasmussen, and Stadler 2018) and (3) continuous phylogeographic approaches aiming to infer geographic coordinates at ancestral nodes (Lemey et al. 2010; Pybus et al. 2012, see Fig. 1 in Baele et al. 2018 for a visual comparison). These different methods are implemented in the software packages BEAST 1 (Suchard et al. 2018) and BEAST 2 (Bouckaert et al. 2019). The choice between a discrete or continuous phylogeographic approach depends on the sampling pattern and on the ecology of the studied organism (or on the epidemiology of the studied pathogen) but also on the question under investigation (Faria et al. 2011; Rasmussen and Grünwald 2021). For instance, the discrete approach may be preferred in situations where multiple sequences are available for a limited number of locations and/or if the sampling distribution can be readily discretized into a limited set of locations (Faria et al. 2011). However, the continuous approach can be more relevant when samples are continuously distributed across space or when diffusion occurs across a landscape in a wave-like manner, making it frequently used to track the spread of wildlife diseases (Rasmussen and Grünwald 2021).

Compared to the discrete model and structured coalescent models, the continuous model does not require an arbitrary grouping of sampling locations, nor does it have to assume that ancestors are located at (one of) the sampling locations

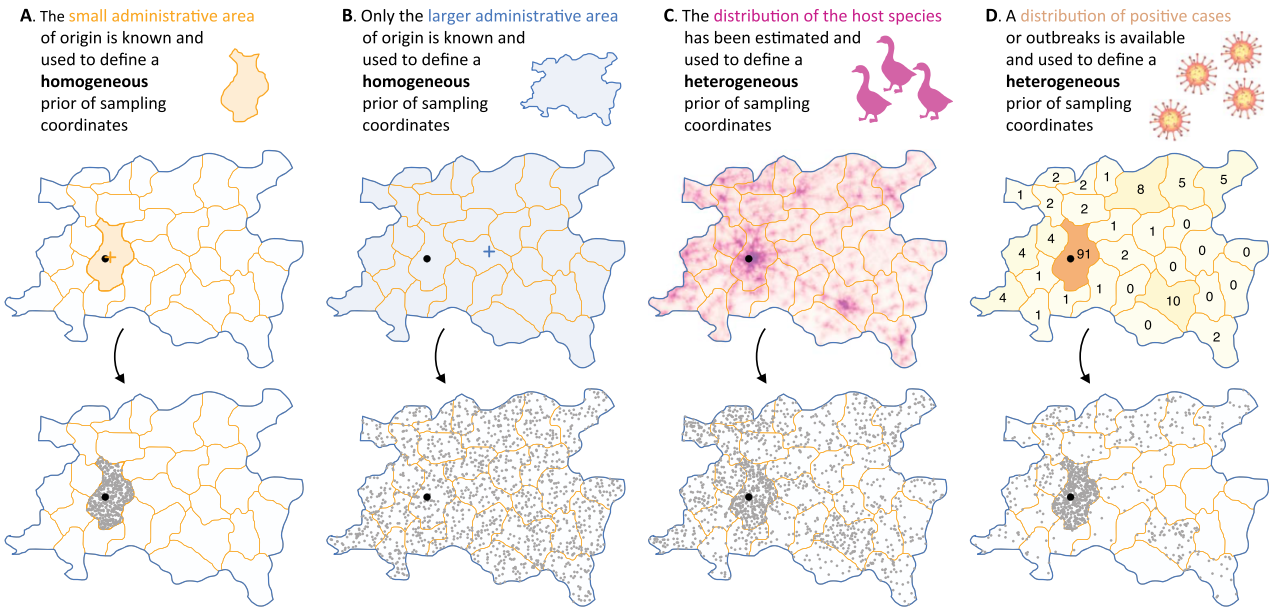

Figure 1. Illustration of the different procedures that can be used to define a homogeneous or heterogeneous prior range of sampling coordinates. To perform a continuous (i.e. spatially explicit) phylogeographic reconstruction of the dispersal history of a fast-evolving pathogen, geographic coordinates associated with the sampling location of each genomic sequence included in the analysis are required. However, precise sampling locations are frequently unknown, not available, or not accessible in the case of human cases protected by privacy data protection rules. When the small administrative area of origin is known (A), sampling coordinates can be integrated through a homogeneous prior range delineated by the polygon of this administrative area. On the contrary, when only the upper-level larger administrative area (e.g. province and state) is known (B), it becomes less relevant to consider the associated polygon to define the prior range of sampling coordinates. In the latter case, and in order to avoid having to discard the considered sample from the data set, external data can be used to define a heterogeneous prior range of sampling coordinates, which thus uses prior information to decrease the uncertainty associated with the geographic origin of the sample. Practically, two different types of external data can be considered: host species distribution (C) or, ideally, the spatial repartition of positive or outbreak cases recorded at the considered sampling time (D). In both cases, those external data are used to define the relative sampling probability assigned to a series of smaller polygon units. In all maps, the actual but unknown sampling point is indicated by a black dot. In Panels A and B, the centroid point of the small and larger administrative area of origin is displayed as an orange and blue cross, respectively.

(Dellicour et al. 2018). Furthermore, while the continuous phylogeographic approach is also impacted by heterogeneous sampling efforts (Kalkauskas et al. 2021), sampling bias is known to notably impact discrete phylogeographic reconstructions (De Maio et al. 2015; Baele et al. 2017) by directly affecting transition rates inferred between discrete locations. The discrete phylogeographic approach does however allow for sampling uncertainty among a given set of discrete locations (Scotch et al. 2019), can be parameterized in terms of covariates (Lemey et al. 2014) and can be extended to model complex temporal (Bielejec et al. 2014; Dudas et al. 2017) or phylogenetic (Faria et al. 2013) scenarios.

For the reasons outlined above, the continuous approach often offers a more realistic alternative to reconstruct the spread of viral lineages in space and time in addition to generating a more fine-grained reconstruction. This approach is however associated with at least two limitations. First, the continuous model is only adapted to dispersal processes that maintain some relationship with geographic distance. Second, it requires geographic coordinates associated with the sampling location of each genomic sequence included in the analysis. In practice, this latter requirement can represent an important limitation because precise sampling locations are frequently unknown, not available, or even not accessible in the case of human or veterinary cases protected by privacy data protection rules/laws. In public databases such as GenBank or GISAID, an important yet hardly estimable amount of genomic sequences are only associated with their country of sampling or a relatively broad administrative area of origin. When research teams aim to complement their new data sets with existing genomic data or to perform a meta-analysis, such a lack of sufficiently precise sampling metadata can prevent the inclusion of valuable genetic data in their analysis. To circumvent this issue and maximize the number of publicly available genomic sequences that can be included in large molecular epidemiological studies employing continuous phylogeographic inference, several methodological approaches have been proposed. We here detail, discuss and compare those different possibilities.

2. Standard approaches

When the reported administrative area of origin is relatively small (Fig. 1A), several relevant options can be considered. First, sampling coordinates could be approximated by the centroid point of the administrative polygon. The fictive example depicted in Fig. 1A illustrates that this may be a sensible approximation when the actual (unknown) sampling point is by chance not so distant from this centroid point. However, considering a centroid point is not necessarily the most adequate option for several reasons: (1) for some administrative areas, the centroid point can sometimes fall outside the border of the administrative polygon; (2) assigning precise coordinates of a fixed point ignores the inherent uncertainty associated with the sampling location of the considered genome sequence; (3) the relaxed random walk (RRW) model of diffusion used to perform continuous phylogeographic inference does not allow different sequences to be associated with identical geographic coordinates. If this is the case, a restricted amount of noise is frequently added to slightly differentiate identical sampling coordinates. In practice, such noise is added using a 'jitter' option that uniformly picks such noise from a user-defined square

area around the sampling point, a square that could problematically include areas falling outside the administrative polygon of the origin or even non-accessible areas (e.g. water areas in case of zoonoses impacting terrestrial species). For these different reasons, the so-called 'jitter' option should ideally be avoided when sampled sequences are only associated with an administrative area of origin (Dellicour et al. 2018).

3. The homogeneous prior approach

An alternative to centroid locations consists of drawing random sampling points within the administrative polygon (Dellicour et al. 2018) or, ideally, using the polygon to define a prior range of possible sampling coordinates and estimate the sampling location through Bayesian phylogeographic inference (Fig. 1A). Defining such a homogeneous spatial range of values has initially been proposed by Bouckaert and colleagues to trace the origins and expansion of the Indo-European languages (Bouckaert et al. 2012). For this purpose, they used the RRW diffusion model to model the language evolution from a data set made of basic vocabulary terms and geographic range assignments for more than 100 different languages. To account for the fact that languages are spoken in geographic areas, they extended the RRW model by allowing the specification of a geographic range (here associated with each language) rather than a point location to explicitly consider the uncertainty around the location assignment. With this novel approach, they found decisive support for an agricultural expansion from Anatolia beginning 8,000 to 9,500 years ago but also illustrated that phylogeographic reconstructions based on a diffusion model can find applications in other fields such as linguistics and anthropology.

In the context of a biogeographic analysis estimating ancestral areas of the plant genus *Centipeda* in Australia, Nylinder et al. subsequently proposed to apply a similar methodological approach accommodating shaped areas for tip locations (Nylinder et al. 2014). Specifically, each *Centipeda* species was assigned to a homogeneous prior range of spatial coordinates defined by its extant distribution. Their results shed light on how the evolutionary history of this plant genus was associated with the temporal increase of aridity since the Pliocene and indicate that *Centipeda* occurrences in western Australia resulted from a recent dispersal rather than an ancient vicariance. This study opened the perspective of biogeographic analyses of taxonomic groups for which the current species ranges cannot easily be delineated as discrete areas.

Since these two initial applications, which are actually outside the field of molecular epidemiology, homogeneous prior ranges of sampling coordinates have also been used for phylogeographic analyses of viruses such as the porcine deltacoronavirus (PDCoV) in China (He et al. 2020). In this study, the authors performed discrete as well as continuous phylogeographic analyses to reconstruct the dispersal history of PDCoV lineages across the Chinese territory. The continuous phylogeographic analysis was based on an alignment of newly sequenced and publicly available genomic sequences that were not associated with sampling coordinates or a sufficiently precise sampling location from which geographic coordinates could have been retrieved. The authors therefore employed the homogeneous prior approach to define ranges of sampling coordinates delineated by the administrative polygon of origin of each genomic sequence. Their resulting phylogeographic reconstruction highlighted frequent long-distance dispersal events that could have involved human-mediated transmission events.

More recently, the RRW diffusion model coupled with the specification of homogeneous prior ranges of trait values has been introduced for the ancestral inference of the climatic niche (as defined by the occupied two-dimensional temperature and precipitation niche space) of a group of bird species (Quintero, Suchard, and Jetz 2022). As introduced by the authors, one current challenge lies in developing analytical approaches that allow linking species niche characterization with describing their evolution. In their study, they analysed the evolution of the two-dimensional temperature and precipitation niche space occupied by different bird species. Their findings include the confirmation that extant birds coevolved from warm climatic niches into colder and drier environments. This recent study further opens the door for enhanced integrations between ecological niche modelling and evolutionary analyses, leading to methodological approaches that could further help understanding the impact of past climate and land-use changes on the ecological niche evolution of target living organisms (endangered species, invasive species, pathogens, etc.).

4. The heterogeneous prior approach

While the homogeneous prior range of geographic coordinates constitutes an interesting approach in the case of relatively small administrative polygons, it progressively loses its relevance when dealing with larger administrative sampling areas. As illustrated in Fig. 1B, sampling points drawn from such a large prior range can potentially fall quite far from the actual (unknown) sampling location. In the context of (viral) phylogeographic analyses, integrating sampling coordinates from too large polygons could result in a non-negligible amount of uncertainty that could in turn impact the precision associated with the phylogeographic reconstruction. To mitigate this issue, Dellicour et al. recently proposed to extend the homogeneous prior feature to a heterogeneous one (Dellicour et al. 2020b). In practice, they allowed specifying several non-overlapping sub-polygons, where each sub-polygon can be associated with a different sampling probability, the sum of which is constrained to 1 (Dellicour et al. 2020b). Implemented in the software package BEAST 1.10 (Suchard et al. 2018) alongside the homogeneous prior approach (Bouckaert et al. 2012; Nylinder et al. 2014), each series of sub-polygons assigned to a sampled sequence can be defined in a distinct external Keyhole Markup Language file. The advantage of this feature lies in the possibility to constrain the overall prior range and hence to reduce the sampling uncertainty.

In order to define the sampling probability associated with each sub-polygon, one can resort to external data such as the host distribution (Fig. 1C) or, ideally, a distribution of positive cases or outbreaks recorded at the time of sampling (Fig. 1D). In the former case, the sampling probability assigned to each sub-polygon can be defined according to the ratio between the estimated host counts in that sub-polygon and the estimated host counts for the entire polygon of origin (Fig. 1C). When only a species density layer is available, density values should ideally be summed rather than averaged across grid cells, at least if the objective is to define sampling probabilities according to the relative host presence and not density. However, when available, a distribution of positive cases or outbreaks recorded at the sampling time should be favoured to define the sampling probabilities assigned to each sub-polygon. Indeed, while the number of hosts estimated in each location reflects the relative potential for infections and thus to sample the pathogen, confirmed cases or outbreaks further certify that such infections were at least recorded in the considered areas. Similar to the situation where they are defined according to the

host species distribution, sampling probabilities can then be estimated proportionally to the number of positive cases or outbreaks recorded in each sub-polygon (Fig. 1D).

5. Limitations

Both the homogeneous and heterogeneous prior approaches are only applicable when the sampled sequences are associated with a well-delimited geographic area of origin, such as, for instance, an administrative polygon within which we can confidently assume that the sequence was sampled. In other words, if the registered administrative area can not be trusted as the actual polygon in which the sequence was sampled, none of those approaches are relevant. This can, for instance, be the case when the administrative area misleadingly corresponds to the location where the sample was conserved and/or analysed. Furthermore, selecting one of the two approaches is described above as a choice depending on the relative size of the known geographic area of origin. Yet, the notion of 'large administrative area' is of course subjective and it depends on the nature and epidemiology of the pathogen under investigation. For instance, one might consider the dispersal capacity of the pathogen in defining if a known geographic area of origin is more or less large, i.e. if a homogeneous or rather a heterogeneous prior approach should be selected to define/constrain the sampling uncertainty. Finally, in the context of the heterogeneous prior approach, using a collection of sub-polygons corresponding to smaller administrative areas each associated with a distinct sampling probability is not necessarily relevant in regard to the external variable used to define those probabilities. For example, when a host population count/density raster (i.e. geo-referenced grid of population density values) is used to define a heterogeneous prior range, directly considering each raster cell as a distinct square polygon might be more relevant than pooling host count/density values within a series of sub-polygons corresponding to administrative areas. In practice, the considered raster should then initially be converted into a set of square polygons each associated with a host count value and that will subsequently be subsampled to define heterogeneous prior ranges.

6. Conclusion

Defining heterogeneous prior ranges of sampling coordinates according to external data, such as the presence of host species or the distribution of confirmed infectious cases, has the potential to decrease the sampling uncertainty associated with an initially large administrative area of origin. In some cases, such an approach could increase the number of available genomic sequences that can be included in a continuous phylogeographic analysis without integrating too much uncertainty in the inference of ancestral locations. Even so, accessibility to sufficiently precise sampling locations remains a notable practical challenge for performing large-scale phylogeographic investigations involving publicly available genomic sequences. Indeed, an important proportion of publicly available genomic data does not include metadata on sampling location that could be exploited to define a precise sampling point, a homogeneous or heterogeneous prior range of sampling coordinates. Aiming for a more systematic integration (or even a required integration prior to submission) of precise sampling metadata could potentially increase the scope of epidemiological studies that could benefit from large-scale spatially explicit phylogeographic analyses.

7. Resources

The homogeneous and heterogeneous sampling prior approaches are both implemented in the software package BEAST 1.10 (Suchard et al. 2018). A detailed protocol on how to prepare and conduct a continuous phylogeographic analysis is available at https://doi.org/10.1093/molbev/msab031 (Dellicour et al. 2021) and has also been described on the BEAST community website using a yellow fever virus study case as an example: https://beast.community/workshop_continuous_diffusion_yfv (Faria et al. 2018). An example of how to prepare a continuous phylogeographic analysis using the homogeneous or heterogeneous sampling prior approach can be found at https://github.com/sdellicour/h5n1_mekong (Dellicour et al. 2020b).

Acknowledgements

We are grateful to two anonymous reviewers and the Associate Editor for their constructive comments on a previous version of the manuscript.

Funding

S.D. is supported by the *Fonds National de la Recherche Scientifique* (FNRS, Belgium). S.D. and P.L. acknowledge funding from the European Union Horizon 2020 project MOOD (grant agreement no. 874850). P.L. acknowledges support by the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek - Vlaanderen*, G051322N, G0D5117N and G0B9317N). S.D. and G.B. acknowledge support from the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek-Vlaanderen*, G098321N). P.L. and M.A.S. acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422 - ReservoirDOCS) from the Wellcome Trust through project 206298/Z/17/Z (Artic Network) and from the National Institutes of Health (grant R01 AI153044). M.A.S. acknowledges support from the National Institutes of Health (grant U19 AI135995). G.B. also acknowledges support from the Internal Funds KU Leuven under grant agreement C14/18/094 and the Research Foundation - Flanders (*Fonds voor Wetenschappelijk Onderzoek - Vlaanderen*, G0E1420N).

Conflict of interest: None declared.

References

Baele, G. et al. (2018) 'Recent Advances in Computational Phylodynamics', *Current Opinion in Virology*, 31: 24–32.

—— et al. (2017) 'Emerging Concepts of Data Integration in Pathogen Phylodynamics', *Systematic Biology*, 66: e47–65.

Bielejec, F. et al. (2014) 'Inferring Heterogeneous Evolutionary Processes through Time: From Sequence Substitution to Phylogeography', *Systematic Biology*, 63: 493–504.

Bouckaert, R. et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337: 957–60.

—— et al. (2019) 'BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis', *PLoS Computational Biology*, 15: e1006650.

Brunker, K. et al. (2018) 'Landscape Attributes Governing Local Transmission of an Endemic Zoonosis: Rabies Virus in Domestic Dogs', *Molecular Ecology*, 27: 773–88.

De Maio, N. et al. (2015) 'New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation', *PLoS Genetics*, 11: e1005421.

Dellicour, S. et al. (2018) 'Phylodynamic Assessment of Intervention Strategies for the West African Ebola Virus Outbreak', *Nature Communications*, 9: 2222.

—— et al. (2021) 'Relax, Keep Walking – a Practical Guide to Continuous Phylogeographic Inference with BEAST', *Molecular Biology and Evolution*, 38: 3486–93.

—— et al. (2020a) 'Epidemiological Hypothesis Testing Using a Phylogeographic and Phylodynamic Framework', *Nature Communications*, 11: 5620.

—— et al. (2020b) 'Incorporating Heterogeneous Sampling Probabilities in Continuous Phylogeographic Inference — Application to H5N1 Spread in the Mekong Region', *Bioinformatics*, 36: 2098–104.

Duchatel, F., Bronsvoort, B. M. D. C., and Lycett, S. (2019) 'Phylogeographic Analysis and Identification of Factors Ompacting the Diffusion of Foot-and-mouth Disease Virus in Africa', *Frontiers in Ecology and Evolution*, 7.

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

Faria, N. R. et al. (2018) 'Genomic and Epidemiological Monitoring of Yellow Fever Virus Transmission Potential', *Science*, 361: 894–9.

—— et al. (2017) 'Establishment and Cryptic Transmission of Zika Virus in Brazil and the Americas', *Nature*, 546: 406–10.

——— et al. (2011) 'Toward a Quantitative Understanding of Viral Phylogeography', *Current Opinion in Virology*, 1: 423–9.

——— et al. (2013) 'Simultaneously Reconstructing Viral Crossspecies Transmission History and Identifying the Underlying Constraints', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368: 20120196.

Guinat, C. et al. (2021) 'Disentangling the Role of Poultry Farms and Wild Birds in the Spread of Highly Pathogenic Avian Influenza Virus H5N8 in Europe', *bioRxiv*. 2021.10.22.465255.

He, W.-T. et al. (2020) 'Genomic Epidemiology, Evolution, and Transmission Dynamics of Porcine Deltacoronavirus', *Molecular Biology and Evolution*, 37: 2641–54.

Jacquot, M. et al. (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.

Kalkauskas, A. et al. (2021) 'Sampling Bias and Model Choice in Continuous Phylogeography: Getting Lost on a Random Walk', *PLoS Computational Biology*, 17: e1008561.

Kim, J. et al. (2018) 'Phylogeographic Analysis of the Full Genome of Sweepovirus to Trace Virus Dispersal and Introduction to Korea', *PLoS One*, 13: e0202174.

Lemey, P. et al. (2014) 'Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2', *PLoS Pathogens*, 10: e1003932.

—— 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.

Müller, N. F., Rasmussen, D., and Stadler, T. (2018) 'MASCOT: Parameter and State Inference under the Marginal Structured Coalescent Approximation', *Bioinformatics*, 34: 3843–8.

Nylinder, S. et al. (2014) 'On the Biogeography of Centipeda: A Species-tree Diffusion Approach', *Systematic Biology*, 63: 178–91.

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.

Quintero, I., Suchard, M. A., and Jetz, W. (2022) 'Macroevolutionary Dynamics of Climatic Niche Space', *Proceedings of the Royal Society B: Biological Sciences*, 289: 20220091.

Rasmussen, D. A., and Grünwald, N. J. (2021) 'Phylogeographic Approaches to Characterize the Emergence of Plant Pathogens', *Phytopathology*, 111: 68–77.

Scotch, M. et al. (2019) 'Incorporating Sampling Uncertainty in the Geospatial Assignment of Taxa for Virus Phylogeography', *Virus Evolution*, 5: vey043.

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

Torres, C. et al. (2014) 'Phylodynamics of Vampire Bat-transmitted Rabies in Argentina', *Molecular Ecology*, 23: 2340–52.

Trovão, N. S. et al. (2015) 'Host Ecology Determines the Dispersal Patterns of a Plant Virus', *Virus Evolution*, 1: vev016.

Zeller, M. et al. (2021) 'Emergence of an Early SARS-CoV-2 Epidemic in the United States', *Cell*, 184: 4939–4952.e15.