Emerging Concepts of Data Integration in Pathogen Phylodynamics

GUY BAELE[1], MARC A. SUCHARD[2,3,4], ANDREW RAMBAUT[5,6], AND PHILIPPE LEMEY[1,*]

[1]*Department of Microbiology and Immunology, Rega Institute, KU Leuven - University of Leuven, Leuven, Belgium;* [2]*Department of Biomathematics and*
[3]*Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA;* [4]*Department of Biostatistics,*
School of Public Health, University of California, Los Angeles, CA 90095, USA; [5]*Institute of Evolutionary Biology and* [6]*Centre for Immunity, Infection and*
Evolution, University of Edinburgh, Kings Buildings, Edinburgh EH9 3FL, UK
*Correspondence to be sent to: Philippe Lemey, Department of Microbiology and Immunology, KU Leuven - University of Leuven, Minderbroedersstaat 10,
3000 Leuven, Belgium; E-mail: philippe.lemey@kuleuven.be.

Abstract.—Phylodynamics has become an increasingly popular statistical framework to extract evolutionary and epidemiological information from pathogen genomes. By harnessing such information, epidemiologists aim to shed light on the spatio-temporal patterns of spread and to test hypotheses about the underlying interaction of evolutionary and ecological dynamics in pathogen populations. Although the field has witnessed a rich development of statistical inference tools with increasing levels of sophistication, these tools initially focused on sequences as their sole primary data source. Integrating various sources of information, however, promises to deliver more precise insights in infectious diseases and to increase opportunities for statistical hypothesis testing. Here, we review how the emerging concept of data integration is stimulating new advances in Bayesian evolutionary inference methodology which formalize a marriage of statistical thinking and evolutionary biology. These approaches include connecting sequence to trait evolution, such as for host, phenotypic and geographic sampling information, but also the incorporation of covariates of evolutionary and epidemic processes in the reconstruction procedures. We highlight how a full Bayesian approach to covariate modeling and testing can generate further insights into sequence evolution, trait evolution, and population dynamics in pathogen populations. Specific examples demonstrate how such approaches can be used to test the impact of host on rabies and HIV evolutionary rates, to identify the drivers of influenza dispersal as well as the determinants of rabies cross-species transmissions, and to quantify the evolutionary dynamics of influenza antigenicity. Finally, we briefly discuss how data integration is now also permeating through the inference of transmission dynamics, leading to novel insights into tree-generative processes and detailed reconstructions of transmission trees. [Bayesian inference; birth–death models; coalescent models; continuous trait evolution; covariates; data integration; discrete trait evolution; pathogen phylodynamics.]

Grenfell et al. (2004) originally introduced phylodynamics to describe "the melding of immunodynamics, epidemiology and evolutionary biology." Grown into a mature field that aims to enhance our understanding of infectious disease transmission and evolution, phylodynamics relies on phylogenetic inference as the core analytical tool to recover evolutionary and epidemic processes from the mutations that accumulate in the genomes of rapidly evolving pathogens during spread of an epidemic. These mutations may confer phenotypic differences that allow viruses to infect different cell types, to evade host immune responses or to transmit by different routes, hosts or vectors (Holmes et al. 1995), but the mutations may also represent the molecular footprint of epidemiological processes that can otherwise not directly be observed. Extracting such information from genetic data represents the primary goal of phylodynamics and requires the integration of additional data and models in a phylogenetic framework. Therefore, phylodynamics is not only considering the interplay between evolution and epidemiology from a conceptual stance, but the integration is made concrete through advances in statistical modeling and computational inference, a key focus of this review.

Rapidly evolving pathogens are unique in that their ecological and evolutionary dynamics occur on the same timescale and can therefore potentially interact. Time of sampling therefore represents important information to incorporate in phylodynamic analyses because it allows calibration of phylogenies, and hence epidemic histories, of rapidly evolving pathogenies in calendar time units (Pybus and Rambaut 2009). Molecular clock models that formalize the relationship between sequence divergence and evolutionary time have been extended specifically for this purpose, and models accommodating sampling time now represent the cornerstone of time-measured phylodynamics (Rambaut 2000; Shapiro et al. 2011). Populations from which "heterochronous" sequence data can be obtained are colloquially referred to as measurably evolving populations (MEPs) (Drummond et al. 2003), a concept that does not only apply to rapidly evolving pathogens but also extends to populations from which ancient DNA can be sampled (e.g., Hofreiter et al. (2001); Pääbo et al. (2004); Shapiro et al. (2004); Molak et al. (2015)). Not surprisingly, location of sampling has also received a great deal of attention because infectious disease transmission is an inherently spatial process. However, location in an epidemic network is not necessarily determined by geographical position, but may be more appropriately represented by position in a social or sexual network, the proximity to vector breeding sites, the movement of hosts or infectious agents through commerce, air travel, wind, or other factors (Smith 2005; Brockmann and Helbing 2013). By offering explicit patterns of connectivity in infectious disease populations, genetic data can be instrumental in

determining the importance of these factors in pathogen spread.

Together with spatial processes, the intensity of transmission and growth or decline in epidemic size can also leave an imprint in viral genomes. This has led to the widespread application of models that relate patterns of evolutionary ancestry to parameters quantifying population size changes or genealogical branching rates. Population genetic approaches based on the coalescent have enjoyed sustained popularity as "tree-generative" models, with the coalescent describing the relationship between the demographic history of a large population and the shared ancestry of individuals randomly sampled from it, as represented by a genealogical tree. This relationship is formalized by a probability distribution of times between the coalescence events in the sample genealogy, which depends on a demographic function that describes population-size change through time. Initially focusing on specific parametric functions of effective population size change through time (Pybus et al. 2000; Strimmer and Pybus 2001), coalescent modeling has evolved to allow for populations with stochastically varying sizes through the use of flexible nonparametric modeling of demographic history (Drummond et al. 2006b; Minin et al. 2008; Gill et al. 2013).

Birth–death models trace back to the work of Kendall (1948) and describe a stochastic process that typically starts with a single species, and allows species to give birth to a new species after an exponential waiting time or to die after an exponential waiting time. Calculation of the probability density of a genealogy generated by the birth–death process has been developed for complete sampling (Gernhard 2008) and incomplete sampling (Stadler 2010) of the population. The latter requires a combination of the birth–death process with a model of the sampling process (Volz and Frost 2014), which can be assumed to vary through time (Stadler et al. 2013).

Pathogen sequences sampled over time and space and their associated traits are now typically analyzed using Bayesian statistical approaches. Bayesian modeling and inference is particularly attractive for phylodynamics because it represents a natural framework for data integration and it also avoids the need to condition on data summary statistics and associated error propagation. By adequately taking into account the uncertainty of unobservables — for example the genealogy — when attempting to draw inference from processes giving rise to or unfolding on genealogies, Bayesian inference constitutes a general and coherent statistical framework that solely conditions on the observed sequences and associated information. Bayesian genealogical approaches can also naturally accommodate tree-generative process as tree priors. However, a Bayesian full probabilistic model requires the specification of appropriate prior distributions over all the parameters in the model, which can make practitioners feel uncomfortable at times and calls for an examination of the sensitivity of posterior estimates with respect to choices in prior distribution. MrBayes

(Ronquist et al. 2012) and BEAST (Drummond et al. 2012) represent two popular Bayesian inference packages in the field that offer a wide range of evolutionary and/or population genetic models, but we restrict ourselves to the BEAST framework in this review because of its traditional focus on measurably evolving pathogens and time-measured trees.

The increasing availability and quality of viral genome sequences, the growth in computer processing power, and the development of sophisticated statistical methods have all contributed to the current popularity of Bayesian phylodynamic inference in infectious disease research (Pybus and Rambaut 2009). However, adequately informing particular evolutionary and population genetic models from genetic data in isolation can be challenging. This may be particularly problematic for trait evolutionary models that are generally fitted to only a single observation of the trait associated with molecular sequences. Parameter uncertainty can hamper the establishment of definite associations with external covariates, and more generally, efficient hypothesis testing. A promising avenue of research that currently emerges is the integration of covariates or other data, like time series of case reports, in the evolutionary and population genetic reconstructions. This represents an additional level of data integration that offers many advantages.

In this review, we first highlight modern approaches to integrate trait and sequence evolution in pathogen phylodynamics and discuss an example of both discrete and continuous trait reconstruction. We expand on this by highlighting several applications that are not restricted to spatial problems. We subsequently discuss how potential covariates of sequence and trait evolutionary processes can be integrated and how additional information about epidemic dynamics can be incorporated in analyses that serve both to reconstruct and test hypotheses about phylodynamic history. We conclude this review by discussing recent developments and future perspectives in the field of phylodynamics.

Sequences and Traits

In terms of the sequence evolutionary process, we make a distinction between the mode of evolution or the relative intensities by which sequence characters are exchanged in evolutionary history and the tempo of evolution which determines how this exchange process scales in units of time. The former can be modeled by different parameterizations of continuous-time Markov chain (CTMC) models and for which we refer to the general phylogenetic literature (Felsenstein 2004; Yang 2006; Lemey et al. 2009b). For the latter, time of sampling provides the most important source for calibration in phylodynamic inference.

As mentioned above, molecular clock models allow quantification of the rate of substitution, and for MEPs, their calibration is generally based on the divergence that accumulates over the sampling time range. The extent
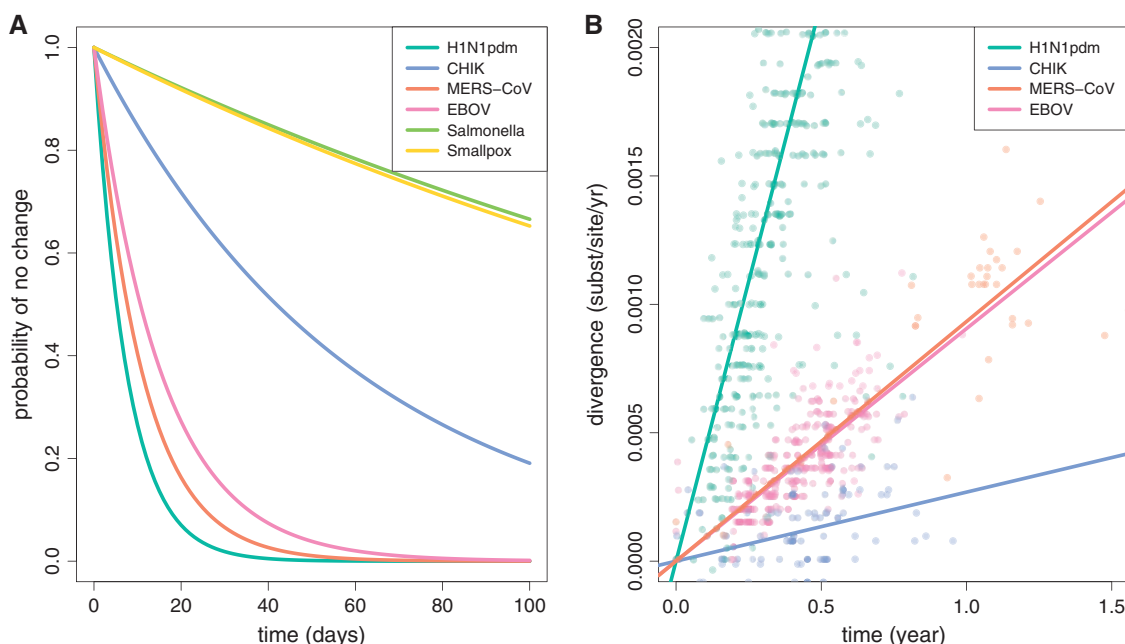
FIGURE 1.   Temporal signal in pathogen genomes sampled through time. A) Plot of the probability of observing no changes between two genomes of length (N) separated by the given number of days (x) for a given rate of evolution per site per year (μ). We used $N = 13,100$ and $\mu = 0.0037$ for Influenza A/H1N1 (Smith et al. 2009), $N = 30,100$ and $\mu = 0.0011$ for MERS-CoV (Cotten et al. 2014), $N = 19,800$ and $\mu = 0.0012$ for EBOV (Park et al. 2015), $N = 13,100$ and $\mu = 0.0005$ for CHIK, $N = 4,641,576$ and $\mu = 0.00000032$ for Salmonella Typhimurium DT104 (Mather et al. 2013), and $N = 190,000$ and $\mu = 0.0000082$ for Smallpox. An online application to plot such probabilities is available at http://epidemic.bio.ed.ac.uk/node/79. B. Root-to-tip divergences as a function of sampling time based on publicly available Influenza A/H1N1, MERS-CoV and EBOV genomes as well as for an unpublished CHIK data set. The regression plots were rescaled for each virus such that the oldest genome in each dataset was set at time = 0 and the regression line has zero divergence at time = 0.

to which we expect to observe genetic changes between viral sequences sampled at different times will depend on the overall rate of substitution (and its constancy), the length of the sequences that can be obtained and the differences in time of sampling. We illustrate this for complete genome sampling from four different viral pathogens that have caused relatively recent outbreaks: pandemic influenza A (A(H1N1)pdm09), Middle East respiratory syndrome coronavirus (MERS), Chikungunya virus and Ebola virus (Fig. 1). For comparison, we also include a DNA virus (smallpox) and a bacterial pathogen (Salmonella Typhimurium DT104). These pathogens differ both in genome length and short-term rates of substitutions, which leads to different expectations for observing changes as time elapses between sequence samples (Fig. 1A). Influenza A(H1N1)pdm09 genomes are expected to accumulate more changes over time, closely followed by MERS and Ebola genomes, whereas this occurs at a considerably slower rate for Chikungunya but still faster than smallpox and Salmonella. The divergence accumulating in actual sequence data sampled during outbreaks of these viruses is largely consistent with the respective probability plots (Fig. 1B). Further in line with these expectations and empirical patterns, Hedge et al. (2013) have been able to accurately estimate evolutionary and epidemiological parameters from A(H1N1)pdm09 genomes as early as two months after the first reported case, which demonstrates a clear potential

for real-time phylodynamic characterization during emerging epidemics. In general, assessing whether viral sequence samples contain sufficient temporal signal is a cautionary — if not a necessary — step prior to fitting dated-tip molecular clocks in order to extract epidemiological processes from viral sequences.

As in the general field of molecular evolution, making a molecular clock assumption serves two main purposes in phylodynamics: dating the phylogenetic history and providing a mechanistic description of the evolutionary process. The former is particularly useful to date cross-species transmissions, viral outbreaks, and the historical spread of epidemics. At a smaller scale, evolutionary estimates may also help to pinpoint the individual's infection time or assess transmission hypotheses (Vrancken et al. 2014b). Furthermore, incorporating sampling time can assist in adequately rooting pathogen phylogenies, which in turn may help in resolving the evolutionary origins of emerging viruses (e.g., for Ebola: Baize et al. (2014); Dudas and Rambaut (2014)). Coupled to tree-generative processes, time-measured genealogies also enable estimation of epidemic growth rates per unit time or basic reproduction rates, even early in a pandemic (Fraser et al. 2009; Volz et al. 2013; Hedge et al. 2013).

The nature by which the temporal information contained in the sampling times can be connected to the genetic similarities embedded in the sequences depends on the specific formulation of the molecular clock

hypothesis. The original constant or "strict" molecular clock model, postulating a single rate of substitution across all branches in the phylogeny, has proven to be too restrictive in many applications and can mislead divergence date estimation as well as phylogenetic inference (Yoder and Yang 2000; Ho and Jermiin 2007). This has motivated several developments to relax the strict molecular clock assumption, which can be largely subdivided into models that allow a limited number of discrete rate changes in the phylogeny ("local" molecular clocks, Yoder and Yang (2000); Drummond and Suchard (2010)) and models that allow the rate to change in a continuous fashion by either assuming or not assuming a relationship between the rates on ancestral and descendent branches (autocorrelated (Thorne et al. 1998) and uncorrelated (Drummond et al. 2006a) relaxed clocks respectively). Whereas the uncorrelated relaxed clock approach has gained considerable attraction as a generic clock modeling approach, it is worth noting that local molecular clock modeling also has interesting applications in pathogen phylodynamics. For example, assuming local molecular clocks for host-specific lineages can lead to more accurate reconstructions of phylogenetic history and this resulted in far more consistent evolutionary reconstructions across different segments of influenza A (Worobey et al. 2014). Host-specific rates of evolution also represent a particular scenario of local molecular clock modeling, but when such a specific hypothesis is not available a priori, a restricted number of discrete rate changes can still be identified through a random local molecular clock approach (Drummond and Suchard 2010).

When the aim is to estimate divergence times, molecular clock specification may be considered as a nuisance and Bayesian model averaging approaches are particularly attractive in this context (Li and Drummond 2012). Accurately modeling substitution variation becomes a primary interest when the tempo of evolution is scrutinized as a mechanistic process. Understanding the sources of rate variation, and in particular determining the role of host ecology, will be the major focus of our discussion on covariates for evolutionary processes. In the next section, we first outline how additional data can be integrated with genetic data through trait evolutionary modeling, both for discrete and continuously valued traits.

Discrete Trait Evolution

Examining traits associated with sequence data in an evolutionary context requires a model of how the traits evolve throughout phylogenetic history. Discrete trait modeling can take guidance from standard phylogenetics and borrow the process of exchange between sequence character states as a generic model for how traits substitute their state over tree branches. Arguably the most frequently considered traits in phylodynamics, and molecular evolution in general, are spatial locations. The interest in spatial dispersal, migration, or vicariance goes back to early naturalists studying the geographic distribution of species, even in the absence of genetic data (Haeckel 1866), and has developed into its own research field referred to as phylogeography. This field has witnessed several developments for reconstructing the geographic locations of ancestral lineages (e.g., Bloomquist et al. 2010), but here we focus on stochastic models of phylogenetic diffusion models and the structured coalescent.

The reconstruction of discrete character states on a phylogeny has traditionally relied on the principle of parsimony, which aims at minimizing the number of historical character changes required to produce the states of the trait we observe at the tree tips (e.g., (Maddison et al. 1984)). As in standard phylogenetics, maximum likelihood (ML) has been proposed as a probabilistic alternative to ancestral reconstruction. Using phylogenetic CTMC modeling, Pagel (1999b) introduced an ML approach to infer ancestral character states for binary discrete characters, which can be readily generalized to multistate characters (Pagel et al. 2004). As phylogenies are seldom known with certainty and generally represent the result of reconstruction procedures, Pagel et al. (2004) describe a general procedure for reconstructing ancestral character states across a statistically justified sample of trees estimated by a prior Bayesian phylogenetic analysis. This procedure combines information about the uncertainty of the phylogeny with uncertainty in the estimate of the ancestral state and avoids constraining the sample of trees to only those that contain the ancestral node or nodes of interest, which can lead to overconfidence in a particular reconstruction. A more efficient, simulation-free method for inferring ancestral traits is presented by Minin and Suchard (2008).

In the context of biogeography, Sanmartín et al. (2008) have also adopted Bayesian inference of discrete phylogenetic diffusion processes, this time proposing to jointly estimate the phylogeny and the trait evolutionary process (in MrBayes, Ronquist et al. (2012)). As discrete diffusion models, the authors consider the most general CTMC process with a different rate of exchange for each pair of states — akin to a General Time-Reversible (GTR) model in nucleotide space (Tavaré 1986) — and constrained versions thereof that represent specific scenarios of the island biogeography process they aim to characterize. A similar Bayesian full probabilistic connection between sequences and traits has also been implemented in BEAST, which initially focused on spatiotemporal reconstructions of viral spread (Lemey et al. 2009a). These sister approaches offer extensive modeling flexibility, but at the expense of a quadratic growth in number of instantaneous rate parameters in the CTMC matrix as a function of the state dimensionality of the trait. This is particularly problematic because the rate parameters are only informed by a single trait character observed at the tree tips.

The specific island biogeography application by Sanmartín et al. (2008) focused on a manageable problem by considering only a restricted number of spatial groups and by sharing a single rate matrix across different groups of organisms with independent phylogenies. However, Bayesian inference also offers specific ways to protect against over-parameterization. Prior specification is an obvious one that was exploited in viral epidemiology (Lemey et al. 2009a), for example, by simply proposing higher rates of diffusion between nearby locations a priori. Motivated by the argument that only a limited set of state transitions throughout evolutionary history can leave their trace in the distribution of a single character, Lemey et al. (2009a) adopted Bayesian stochastic search for variable selection (BSSVS) to reduce the number of rate parameters to a restricted set that provides the most adequate parsimonious description of the diffusion process. In the context of discrete diffusion models, BSSVS associates every pairwise rate with a binary indicator but a priori prefers to only invoke a minimal number of rates with an nonzero indicator to explain the tip trait observations. Variable selection and informative prior specification both increase statistical efficiency, which becomes even more important when drawing inference from sparse data under more complex models, for example, assuming two different rates of diffusion for each directionality between a pair of states (Edwards et al. 2011).

Simultaneous inference of sequences and traits implies that both data sources can impact the phylogeny. In most cases, however, this is of little practical importance because the information contained in the multiple sites comprising the sequence alignment can swamp the information present in a single trait character. However, in the presence of very shallow sequence diversity, the tendency of taxa to cluster together by trait state may become more prominent. It is probably not unreasonable to assume that taxa with the same trait character will be more closely related, but the relative strength of such a statement may be subject of debate. It is worth noting that the simultaneous inference capability has been explicitly leveraged to obtain "combined-data" trees in specific contexts, for example for morphological evolution (Nylander et al. 2004) and protein structure evolution (Scheeff and Bourne 2005).

Despite the overwhelming interest in geographic spread, various traits are now the subject of ancestral reconstruction analyses. To illustrate this, we present a sample of applications of the BEAST discrete trait modeling approach in Table 1, mostly focusing on nonspatial studies. These applications range from the reconstruction of morphological characters in animals and plants to antigenicity in influenza viruses and host species in pathogens and bacteria. In phylodynamics, host traits may be of particular interest as demonstrated by a seminal study on bat rabies viruses in the Americas (Streicker et al. 2010). Bat rabies in the Americas represents an interesting multihost system, in which rabies viruses jump between different bat

TABLE 1. Discrete trait applications

Trait	Organism	References
Spatial	Avian Influenza A H5N1	Lemey et al. (2009a)
	Mycobacterium tuberculosis	Comas et al. (2013)
	Influenza A H3N2	Lemey et al. (2014)
Spatial and temporal (season)	Influenza A H3N2	Bahl et al. (2011)
Host	Bat rabies	Streicker et al. (2010, 2012); Faria et al. (2013)
	Staphylococcus aureus CC398	Ward et al. (2014)
	Campylobacter	Dearlove et al. (2015)
	Salmonella Typhimurium DT104	Mather et al. (2013)
Antibiotic resistance genes	Staphylococcus aureus CC398	Ward et al. (2014)
Virulence determinants	Staphylococcus aureus CC398	Ward et al. (2014)
HA & NA subtypes (reassortment)	Avian Influenza Virus	Lu et al. (2014)
Antigenic clusters	Influenza A H3N2	Zinder et al. (2013)
Animal tissue	Small ruminant lentivirus (SRLV)	Ramírez et al. (2012)
Larval characters	Thecostraca	Pérez-Losada et al. (2012)
Gag and env subtypes (recombination)	HIV-1 group M	Ward et al. (2013)
Leaf traits	Triodiinae	Toon et al. (2015)
Morphological characters	Pradosia	Terra-Araujo et al. (2015)
	Pycnandra	Swenson et al. (2015)
Habitat preference	Pradosia	Terra-Araujo et al. (2015)
Phyllotaxy	Polygonatum	Meng et al. (2014)
Habitat types	Mus Nannomys	Bryja et al. (2014)
Cellular compartments	HIV-1	Cybis et al. (2013)

Notes: A non-exhaustive overview of applications of the discrete trait modeling approach by Lemey et al. (2009a), demonstrating the general applicability of the methodology beyond the inference of viral phylogeographic processes.

hosts, occasionally resulting in sustained transmission in the new host species. Over time, this process has produced a phylogenetic structure comprising lineages that are compartmentalized by a dominant bat host (Fig. 2). To study the ancestral host shifts giving rise to this pattern, Streicker et al. (2010) applied Bayesian phylogenetic inference to demonstrate that these events have the tendency to occur between more closely related bat species. We will return to this example when discussing the integration of covariates in sequence and trait evolutionary processes to test the causes and consequences of this host switching process. The study by Streicker et al. (2010) also highlights an alternative
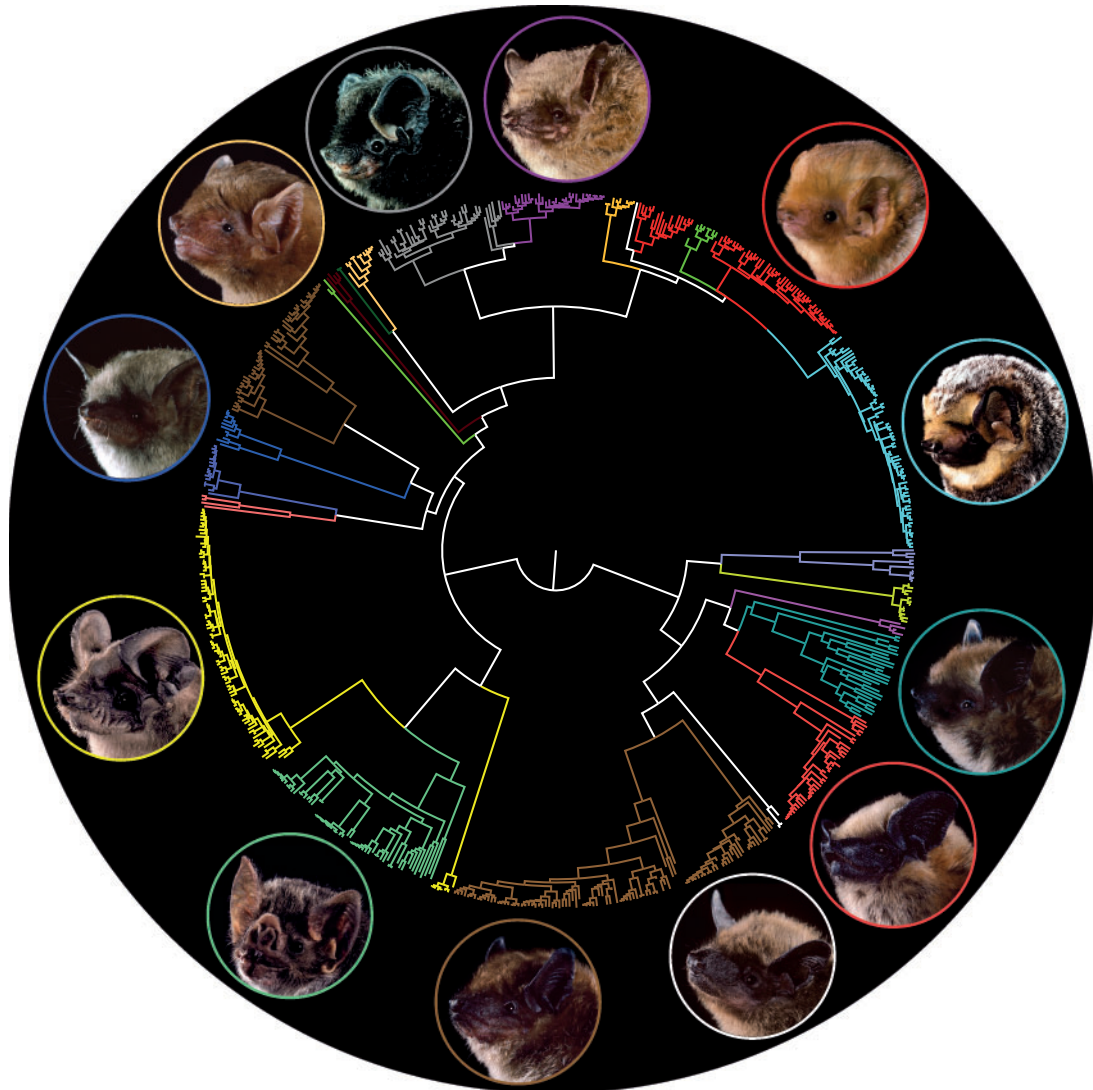
FIGURE 2.   The multihost transmission dynamics of rabies in American bat species. Streicker et al. (2010) identified 18 phylogenetic lineages of rabies virus that were statistically compartmentalized to particular bat taxa. These lineages are represented by differently colored clades in the phylogeny, along with a selection of bat species involved. CSTs are defined as jumps to bat species different from the dominant species within each lineage or clade. These dynamics have been quantified though structured coalescent approaches (Streicker et al. 2010). Host switches on the other hand are defined as the jumps along internal branches to new hosts followed by successful transmission in the new host species. Inferring the history of host jumping along the branches (mostly represented by white branches) has been the subject of discrete trait reconstruction (Streicker et al. 2010). We thank Daniel Streicker for providing the tree and MerlinTuttle.org for granting permission to use the bat portraits.

approach to discrete trait diffusion because the cross-species transmission (CST) dynamics within each host-associated lineage were quantified using a structured coalescent approach. By factoring in the diversity within each host, this approach allowed them to estimate *per capita* CST rates and test their association with several potential explanatory variables.

The structured coalescent also emerged in the context of spatial migration and goes back to the seminal work by Hudson (1990) and Notohara (1990). As an extension of Kingman's coalescent (Kingman 1982), the structured coalescent model builds on the standard Wright-Fisher model by specifying a number of discrete subpopulations, allowing location to explicitly affect the coalescent rate (Vaughan et al. 2014). Spatial subdivision into a number of distinct "demes" represents an obvious population structure, but as illustrated by the bat rabies example, any logical categorization of individuals can be considered by this model, including for example the temporal subdivision used in compartmental epidemiological models (discussed later in this review).

Probabilistic ancestral reconstruction and the structured coalescent rely on a very different set of assumptions, and have traditionally been interested in different estimates (ancestral states vs. population migration rates). Computational restrictions may have prevented the widespread adoption of structured coalescent approaches in phylodynamics. Interestingly,

this is now being addressed by recent developments in Bayesian inference under the structured coalescent. Vaughan et al. (2014) introduced a new set of MCMC transition kernels in BEAST that achieve significantly faster mixing compared with previous transition kernels (Ewing et al. 2004). Further, De Maio et al. (2015) tackle the impractical computational demands when confronted with large numbers of subpopulations and migration events by introducing a new model-based approach that achieves a close approximation to the structured coalescent, while still integrating over all possible migration histories. Their Bayesian structured coalescent approximation (BASTA) combines the accuracy of methods based on the structured coalescent with computational efficiency, offering less biased estimated compared to discrete trait reconstruction approaches. De Maio et al. (2015) attribute this to specific assumptions associated with modeling migration of lineages as a mutational process.

Continuous Trait Evolution

Modeling evolution of continuous characters stems from the field of comparative biology that was challenged with the problem of comparing continuous traits, for example to assess their correlation, across a number of taxa. Standard correlations assume independent data, which is obviously violated for taxa traits due to their shared ancestry. Felsenstein (1985) proposed the "independent contrasts" approach to deal with such nonindependence. This approach assumes that the changes in two characters follow a Brownian motion process, implying that they are drawn from a bivariate normal distribution with some degree of correlation between the characters along each branch of a phylogeny. In addition, the variance of the distribution of change along each branch is assumed to be proportional to the time elapsed on that branch. This procedure uses a phylogeny to identify a set of mutually independent comparisons between pairs of species, pairs of nodes, or a node and a species (Pagel 1997). Grafen (1989) presents a generalization of the independent contrasts approach that allows for trait adjustment by species-specific covariates. This approach connects general linear modeling to phylogenetic data and has become known as phylogenetic regression in the field of comparative biology.

In the conventional continuous random-walk model, traits evolve in each instant of time with a mean change of zero and unknown and constant variance. Originally introduced as an approximation of sequence evolution (Edwards and Cavalli-Sforza 1964), phylogenetic Brownian random walks have mostly been applied to phenotypic traits in comparative biology (see e.g., Martins (1994)). Analogous to discrete trait modeling, both maximum likelihood and Bayesian approaches have been proposed to draw inference under such models (Schluter et al. 1997; Pagel 1999a; Lemmon and Lemmon 2008). In phylodynamics, a bivariate Brownian random walk was again initially applied in a phylogeographic setting (Lemey et al. 2010). The Bayesian implementation proposed by this work not only avoids conditioning on the usual data summaries, for example a fixed tree and branch lengths, but also relaxes the constant diffusion rate assumption implied by a constant-variance random walk. Specifically, variation in diffusion rates among branches was modeled by borrowing from uncorrelated relaxed clock models (Drummond et al. 2006a), that is by rescaling the variance of the random walk (or precision in a Bayesian terminology) along each branch using a scalar drawn independently and identically from an underlying distribution. This relaxed random walk (RRW) appeared to be critical to accommodate a very high degree of heterogeneity that can underlie the spatial spread of pathogens, as was the case for the West Nile virus invasion in the United States (Pybus et al. 2012). The same analysis also challenged standard inference under these models, which initially relied on sampling internal node realizations of the bivariate locations but proved to be inefficient in this case. Analogous to efficient likelihood computation for discrete traits (Felsenstein 1981), Pybus et al. (2012) pursued an approach that integrates out internal node states in the Bayesian framework (see also Freckleton (2012)), which makes applications to large datasets more accessible.

Although diffusion rate variation can be taken into account, a phylogeographic RRW process still relies on dispersal as a function of geographic distance (or distance in Euclidean space). This may be unrealistic for pathogens that exploit human mobility or trade (e.g., human seasonal influenza (Lemey et al. 2014) and swine influenza respectively (Nelson et al. 2015)) and modeling efforts have demonstrated that pathogen spread follows "effective distances" measured along complex transportation networks (Brockmann and Helbing 2013). Due to these limitations, future applications of continuous trait modeling phylodynamics may also focus more on other traits, such as pathogen phenotypes and infection traits. A limited sample of applications of the continuous diffusion framework in BEAST (Lemey et al. 2010) (Table 2), mostly focusing on nonspatial traits, indeed highlights the general use of this approach even beyond the field of biology.

For pathogens, traits related to phenotype or infection severity, for example, will generally be measured through experimental assays. Although this does not hamper the general application of multivariate diffusion models, it can impose additional modeling challenges. One of these challenges is adequately incorporating measurement error (analogous to intraspecific variation for organismal traits), that is conditioning on the data (the repeated measures) rather than data summaries such as their mean and variance. The Bayesian multivariate diffusion approach naturally achieves this by numerically integrating the unobserved average tip trait values based on repeated measures (Vrancken et al. 2014a).

TABLE 2. Continuous trait applications

Trait	Organism or data	References
Spatial	Chorus frogs	Lemmon and Lemmon (2008)
	Raccoon rabies	Lemey et al. (2010)
	Indo-European languages	Bouckaert et al. (2012)
	Asian Languages	Dunn et al. (2013)
	Ainu language	Lee and Hasegawa (2013)
Normal mode	Enzymes	Lai et al. (2012)
Antigenic measurements	human Influenza A and B	Bedford et al. (2014)
(Set point) viral load	HIV-1	Vrancken et al. (2014a)
	Sigma viruses	Vrancken et al. (2014a)
Antibody neutralization	HIV-1	Vrancken et al. (2014a)
Femur length	Theropod dinosaurs	Lee et al. (2014)
Shell lengths	Chemosymbiotic deep-sea mussels	Lorion et al. (2013)
Generation times	Anthropoids	Schrago (2014a,b)
Gene flow	Rhesus macaques	Strickland et al. (2014)

Notes: A nonexhaustive overview of the applications of the continuous trait modeling approach by Lemey et al. (2010), demonstrating the broad range of applications even going beyond biology. For example, their approach can be used to simultaneously use sequence data with antigenic measurements to estimate antigenic cartography for influenza or to infer the spatiotemporal expansion of language families.

Another challenge may be that the trait is only indirectly observed through an assay. This is the case for influenza antigenic evolution, which is often assessed through pairwise measurements of cross-reactivity between influenza strains using the hemagglutination inhibition (HI) assay. The influenza virus population continually evolves in antigenic phenotype to escape host immunity in a process known as antigenic drift, which explains the continual need to update vaccine formulations. To capture this process from an incomplete table with noisy HI measurements (HI titres), Smith et al. (2004) proposed the use of multidimensional scaling (MDS) techniques to position viruses in a two-dimensional cartographic map such that the distance in the lower-dimensional space best fits the HI assay titres. More recently, Bedford et al. (2014) adopted a Bayesian formulation of MDS (Oh and Raftery 2001) that models the observed differences (HI titres) to be centered around their cartographic expectation with a Gaussian error (Fig. 3). As a prior on the unknown location parameters, their approach resorts to the phylogenetic Brownian diffusion process, which leads to an explicit connection between antigenic evolution and genetic relatedness. Bedford et al. (2014) apply this approach to HI data from all human influenza lineages, A/H3N2, A/H1N1, B/Victoria and B/Yamagata, and show that A/H3N2 evolves faster and in a more punctuated fashion than other influenza lineages. More recently, these differences in antigenic evolution, coupled to the nature of human behavior in seasonal influenza epidemiology, were shown to drive differences in migration rate and hence also epidemic success (Bedford et al. 2015).

Although the concept of phylogenetic Brownian diffusion was adopted by the comparative approach to specifically accommodate phylogenetic dependence among traits, the question may arise as to how much dependence really needs to be taken into account. In other words, to what extent does the phylogeny explain similarity among taxa traits? In pathogen phylodynamics, this question may apply to virulence or infection traits, which can also be heavily impacted by the host environment. Specifically, for chronic infections such as HIV-1 and HCV, the comparative framework is being deployed to determine to what extent the viral genotype can control for the rate of progression or infection outcome, sometimes leading to mixed conclusions (Alizon et al. 2010; Vrancken et al. 2014a; Hartfield et al. 2014; Hodcroft et al. 2014). Among the different approaches available to test or quantify this "phylogenetic signal," Pagel's λ (Pagel 1999a) is commonly used. Conceptually, this parameter scales internal node heights of a tree, on which Brownian evolution of traits are being modeled, in such a way that any trait correlation scenario can be accommodated from the estimated phylogeny down to a star-like tree (completely independent taxa traits). The BEAST framework we highlight here incorporates Pagel's λ estimator and therefore supports simultaneous estimation of evolutionary history and trait phylogenetic signal (Vrancken et al. 2014a).

In the framework of emerging infectious diseases, a similar question relates to host shifts and what determines the sensitivity to viral infection in the new host. Longdon et al. (2011) examined this in great detail for three host-specific sigma viruses in *Drosophila* and measured viral titres upon experimentally infecting 51 *Drosophila* species with each of these viruses. Using a phylogenetic mixed model (PMM), the authors found that the host phylogeny could explain most of the variation in viral replication and persistence between different host species. A PMM finds its analog in mixed modeling in quantitative genetics, where phenotypes of individuals related by a pedigree are partitioned into additive genetic (heritable) and residual (nonheritable) components (Henderson 1984; Lynch and Walsh 1998; Housworth et al. 2004). The implementations essentially differ in their specification of the correlation structure of the heritable components: a relationship matrix for pedigrees versus a matrix of shared common ancestry in a phylogeny. It is interesting to note that estimates of phylogenetic signal through Pagel's λ find a similar degree of phylogenetic association of sensitivity to sigma virus infection compared to a standard phenotypic trait as wing size in *Drosophila* (Vrancken et al. 2014a). In addition to the sensitivity to infection following host shifts, it may also be important to predict how much harm the pathogen will cause in the new hosts. Virulence may be considered as a direct consequence of pathogen replication, but also the host may be an important determinant. To examine how virulence may vary across new host species, Longdon et al. (2015) carried out a large cross-infection experiment of *Drosophila* C virus in 48 species of Drosophilidae. In this case as well, PMM analyses showed that changes in virulence, which can be
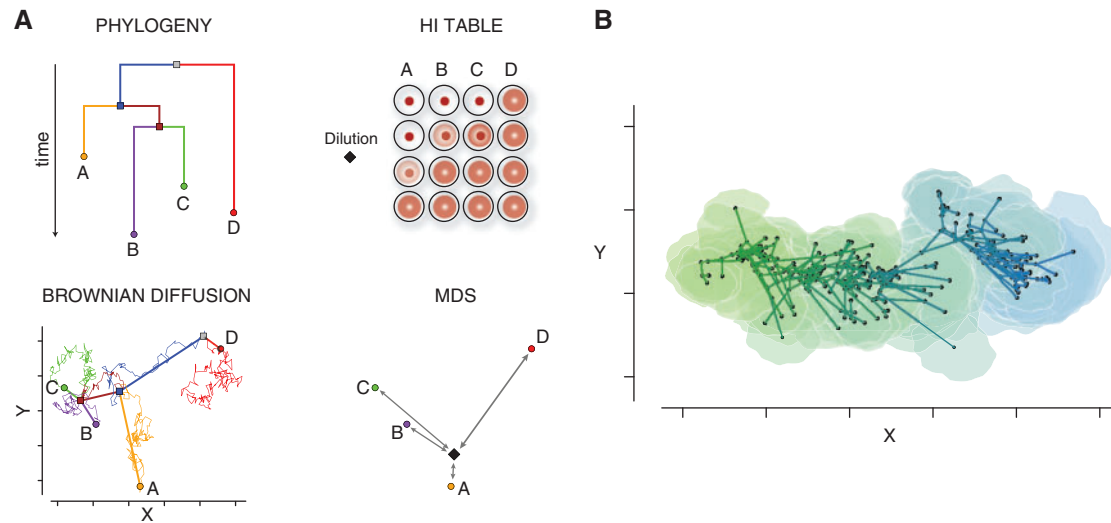
FIGURE 3. Antigenic cartography meets Brownian phylogenetic diffusion. A) Conceptual representation of the integration of genetic and antigenic evolution through a Bayesian MDS approach. In the HI assay (on the right), the antigenic phenotype is investigated by measuring the cross-reactivity of a virus (A, B, C, or D) strain to serum (♦) raised against another strain. Based on these HI measurements, MDS approaches allow to position viruses in lower-dimensional space such that the distances in this space best fit the HI assay titres. A probabilistic interpretation of MDS assumes that the observed differences are centered around their cartographic expectation, in which case the virus locations are estimable parameters. We refer to Bedford et al. (2014) for more information on how these locations are estimated in an integrated Bayesian phylogenetic framework. B) Visualization of antigenic drift dynamics reconstructed using Bayesian MDS in a two-dimensional map. These patterns are inferred from the 2002 to 2011 subset of the influenza A/H3N2 dataset analyzed by Bedford et al. (2014). X and Y represent the first and second antigenic dimensions. The contours represent the 80% HPD region for the node locations (both internal and external nodes). The colors range from green to blue for the lines, points and contours reflects the age between 2002 and 2011. This figure was made using SpreadD3 (Bielejec et al. 2016).

extremely large, were highly predictable from the host phylogeny (Longdon et al. 2015).

COVARIATES OF SEQUENCE AND TRAIT EVOLUTION

Although many phylodynamic hypotheses can be addressed through the analysis of genetic data, trait data, or their combination, additional data in the form of covariates may be required to further dissect the dynamic forces that determine the diversity of epidemiological and phylogenetic patterns. Integrating such data in phylodynamic approaches can serve two purposes: better informing reconstructions and identifying which covariates explain the evolutionary or epidemiological process. In this section, we discuss a number of examples to highlight recent methodological advances to achieve these goals for sequence evolution and trait evolution (tree diffusion processes).

THE TEMPO OF SEQUENCE EVOLUTION

The rate of evolution in different viral populations may be affected by their host environment, either through varying selective dynamics or through an impact on the replication rate. Although independent rate estimates could in theory be used to investigate this, a sparse heterochronous sequence sample from each population generally leads to uncertain evolutionary estimates which complicates formal statistical testing. In the context of HIV-1 evolution in different patient

groups, Edo-Matas et al. (2011) proposed the use of a Bayesian hierarchical phylogenetic model (HPM) to pool information across patients and improve estimate precision for patient-specific viral populations, while still allowing for evolutionary rate differences between the individual populations. Hierarchical modeling was first introduced in phylogenetics to share information across alignment partitions (Suchard et al. 2003). By applying independent parameters that share a hierarchical prior distribution with unknown estimable hyperparameters, HPMs hold a middle ground between independent and shared parameter estimation (Fig. 4).

Following the terminology of ANOVA and regression models, the specification of hierarchical priors can model random effects on the evolutionary response variable. In order to test the impact of patient groups defined by disease progression and host genetic status, Edo-Matas et al. (2011) further extend the HPM approach by incorporating fixed effects for N covariates, arriving at the following general form:

$$\log\theta_i = \beta_0 + \delta_1\beta_1 x_{i,1} + \ldots + \delta_N\beta_N x_{i,N} + \epsilon_i, \quad (1)$$

where θ_i is the evolutionary response variables in patient i, β_0 is an unknown grand mean, β is the estimated effect size of covariate x, δ is a binary indicator that tracks the posterior probability of the inclusion of covariate x in the model and ϵ_i are independent and normally distributed random variables with mean 0 and an estimable variance. The specification of indicator variables (δ) implements a variable selection procedure analogous to the approach aimed at reducing the
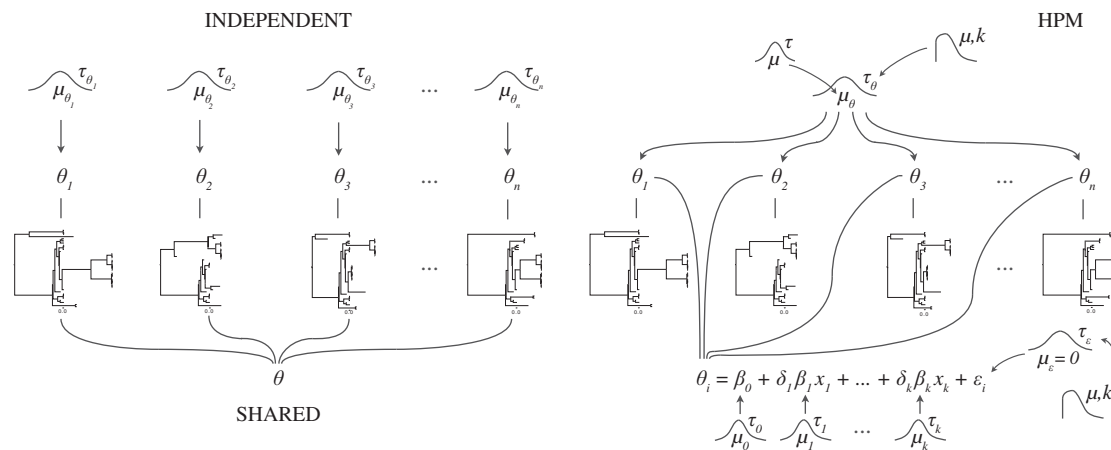
FIGURE 4. Contrasting models with completely linked and unlinked parameters to hierarchical modeling without and with fixed effects. Traditionally, two competing approaches were used when performing Bayesian inference to estimate parameters from a potentially large number of partitions, strata, or individuals: total evidence, where all the data across strata are pooled or shared to estimate a single parameter of interest, and unconditionally independent partitioning, where each stratum requires the estimation of a completely independent set of parameters. The latter is associated with independent prior specification on every parameter. Hierarchical phylogenetic models (HPMs) offer a middle ground between these two extremes by sharing a hierarchical prior distribution over all parameters with estimable mean and variance, which are drawn from hyperpriors. Edo-Matas et al. (2011) propose an HPM that employs a Bayesian mixed effects model that pools information across patients, affording more precise individual-patient parameter estimates when the data are sparse for a patient, but also allowing estimation of the effect of patient groups or continuous covariates.

number of rate parameters in discrete diffusion matrices; although it was considered to be an approach for increasing statistical efficiency in the latter context, it is used here as a model averaging approach that effectively integrates over all possible combinations of covariate inclusion. For this and all other applications of variable selection we highlight, we note that covariate or fixed effect support values can be readily computed in the form of Bayes factors based on the posterior and prior odds for their inclusion. By applying this approach to within-host HIV-1 data sampled from different patient groups, Edo-Matas et al. (2011) not only demonstrate significant shrinkage of the estimator variance, but they also provide support for faster viral evolutionary rates in patients that progressed to AIDS more rapidly.

A similar need for shrinkage and hypothesis testing emerged in a study on the evolutionary consequences of host switching in bat rabies viruses in the Americas (Fig. 2) (Streicker et al. 2012). Although considerable variation in rabies evolutionary rates was noticeable among lineages associated with different hosts on this longer evolutionary time scale (Streicker et al. 2012), accurate quantification remained difficult due to limited sequence samples and their variation across different host species. The authors therefore constructed an HPM over the evolutionary rate parameters at the third codon position — as a proxy for synonymous evolution — in 21 independent bat rabies virus lineages. The fixed effects in their full model included physiological (basal and torpid metabolic rate), environmental (climatic region: temperate vs. tropics/subtropics) and ecological traits (coloniality and seasonal activity). Also in this case, HPM estimates proved to be less sensitive to stochastic noise associated with sampling error, and the authors were able to show an accelerated rate of molecular

evolution in subtropical and tropical bats compared with temperate species. The association between geography and the tempo of evolution was explained by climate-associated differences in seasonality in bat activity and virus transmission.

The examples above consider categorical or continuous covariates for independent intrahost or interhost viral populations, but covariates may also represent categorical branch assignments in a phylogeny (conditionally independent evolutionary lineages). Vrancken et al. (2014b) took this into account in an examination of HIV-1 evolutionary rate differences within and between hosts in a known transmission chain. In this case, fixed effects were modeled as branch assignments that distinguish the transmitted lineage from other branches within specific hosts constituting the transmission chain. Random effects — in terms of possibly different rates for each branch — were modelled according to an uncorrelated relaxed clock process (following Drummond et al. (2006a)). This mixed effects modeling approach demonstrated a significantly slower rate for the transmitted lineage, which confirms earlier observations of evolutionary rate differences within and between hosts (Alizon and Fraser 2013) and provides support for the "store-and-retrieve" hypothesis in HIV transmission (Lythgoe and Fraser 2012; Fraser et al. 2014).

TREE DIFFUSION PROCESSES

By extending the covariate modeling approach for a particular evolutionary parameter to a matrix of transition rate parameters defining a CTMC process, Lemey et al. (2014) developed an approach to
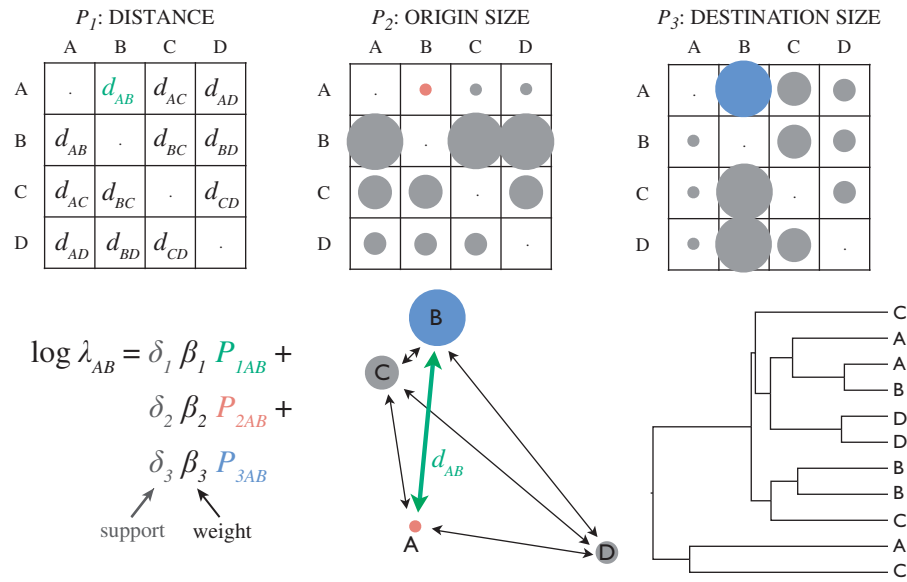
FIGURE 5.   GLM extension of discrete phylogenetic diffusion and examples of covariates or predictors (P) in a spatial context. The GLM parameterizes each rate of among-location movement in the phylogeographic model as a log linear function of various potential predictors. For each predictor P_{jXY} ($j \in \{1,2,3\}$; $X, Y \in \{A,B,C,D\}$), the GLM parameterization includes a coefficient β_j, which quantifies the contribution or effect size of the predictor (in log space), and a binary indicator variable δ_j, that allows the predictor to be included or excluded from the model. Since predictors are essentially matrices of pairwise measurements, location-specific measurements such as population size or density are separated into origin or destination size.

simultaneously reconstruct spatiotemporal history and identify which combination of covariates associates with the pattern of spatial spread. This extension of the discrete phylogeographic diffusion approach (Lemey et al. 2009a) shares the generalized linear model (GLM) formulation for fixed effects introduced above, in this case by parameterizing each rate of among-location movement in the phylogeographic model — typically denoted as the *ij*-th elements (Λ_{ij}) of the transition rate matrix Λ — as a log linear function of various potential covariates as following:

$$\log \Lambda_{ij} = \beta_1 \delta_1 x_{i,j,1} + \beta_2 \delta_2 x_{i,j,2} + \ldots + \beta_N \delta_N x_{i,j,N}, \quad (2)$$

where β and δ represent the same parameters as before. In this case, a single covariate, also referred to as predictors in Lemey et al. (2014), is a flattened vector of quantities corresponding to entries in the *i* to *j* rate matrix $\mathbf{x}_n = (x_{1,2,n}, \ldots x_{K-1,K,n})'$. In a phylogeographic setting, these predictors may, for example, represent a matrix of pairwise geographic distances or population sizes either at the origin or destination location (Fig. 5). As pointed out above, priors and posteriors for the inclusion probabilities (δ) can be used to express the support for each predictor in terms of BFs. Unlike the HPM or mixed effects modeling, the GLM parameterization does not consider random effects for the rates because of the difficulty to inform this large number of effects based on a single trait observation at the tree tips (but see (Trovão et al. 2015) for identifying exceptional random effects). The application of this approach to the seasonal dynamics of influenza H3N2

provided consistent support for air travel governing the global migration of the virus. We note that the GLM diffusion approach achieves both goals outlined in the introduction of covariates of evolutionary and epidemiological processes: it identifies the relevant covariates in the epidemiological process, but at the same time it helps to inform the ancestral reconstructions. For the influenza example, this may lead to more accurate reconstructions of the seasonal source-sink patterns, and in predictive modeling, the combination of genetic and air transportation data also outperformed both data sources in isolation (Lemey et al. 2014).

To demonstrate its generality, Faria et al. (2013) applied the GLM diffusion approach to host switching (jumps followed by successful transmission in the new host) and cross-species transmission (CST, spill-over in a viral clade associated with a dominant host) in bat rabies viruses (Fig. 2). These dynamics were separately investigated by Streicker et al. (2010) through standard ancestral reconstruction and structured coalescent analyses respectively, the latter resulting in rate estimates that were used to test several potential covariates as potential determinants of the CST dynamics. To simultaneously infer and test both host switching and CST in a single framework, Faria et al. (2013) made use of the flexibility of phylogenetic CTMC modeling to accommodate distinct host transition processes, parameterized by independent GLM diffusion models, on external and internal branches. To a large extent, this approach discriminates between recent CSTs, of which many are likely to result in dead-end infections, and the ancestral host shifts. The same predictors as used by

Streicker et al. (2010) were considered for both branch-specific GLM diffusion parameterizations: host genetic distance, geographical range overlap and similarities in roost structures, wing aspect ratio, wing loading, and body size. The latter three morphological measurements represent approximations of foraging niche overlap in bats (Streicker et al. 2010). The full Bayesian analysis revealed that host similarity was strongly predictive of host jumping intensity, both for host switching and CST, whereas a modest support for geographical range overlap was only recorded for CST.

Although the primary purpose is likely to remain the identification of relevant covariates to integrate in tree diffusion processes, we note that the GLM diffusion model may offer the only practical solution to discrete ancestral reconstruction for traits with high-state spaces because it reduces the number of parameters to a linear function of the number of predictors rather than a quadratic function of the number of states. High-state spaces still challenge likelihood calculations, but massively parallel computations offer considerable speed-up in such cases, and practical solutions are now in place both in terms of hardware (e.g., GPU cards) and software to support phylogenetic likelihood computations on such hardware (e.g., BEAGLE) (Suchard and Rambaut 2009; Ayres et al. 2012; Baele and Lemey 2013).

Data Integration to Elucidate Transmission Dynamics

In this section, we discuss how data integration is being considered to help uncover tree-generative processes in phylodynamics. We make a distinction between the reconstruction of large-scale epidemic dynamics based on a limited sample from the pathogen population and the reconstruction of transmission trees from densely sampled pathogen genetic sequences to recover the chain of transmission in extensive detail. Much of this work attempts to relate the transmission dynamics inferred from genetic data to mathematical epidemiology, or even explicitly builds a bridge between pathogen genetics and how transmission dynamics are formalized in compartmental models. As opposed to coalescent models describing the merging of lineages backward in time starting from a small sample of the population until a common ancestor has been reached, compartmental models like the susceptible-infected-removed (SIR) model describe the dynamics of an entire population going forward in time. Modeling infectious disease dynamics with compartmental models allows the description of nonlinear time series of prevalence of infection and the number of susceptible hosts, and leads to important predictions about pathogen spread, for example the prevalence and duration of the epidemic. They also offer a framework to assess the potential impact of intervention strategies on the outcome of an epidemic.

Reconstructing Large-Scale Epidemic Dynamics

Using time-varying coalescent models, phylogenies can be used to estimate how effective population sizes (N_e) change through time. This N_e represents the size of an idealized population that loses or gains genetic diversity at the same rate as the census population from which the sample has been drawn. Although frequently applied as the only source of information about past population dynamics, the dynamics of N_e have sometimes been contrasted against other data. For example, Biek et al. (2007) demonstrated that the expansion of rabies in the North American raccoon population shows similar dynamics to an epidemiological index that reflects the size of the area newly affected by rabies during each month of the outbreak.

Bennett et al. (2010) also report a correspondence between fluctuations in N_e and case counts for dengue serotype 4 between 1981 and 1998 in Puerto Rico. However, when formally assessing this relationship, they identify an offset of about seven months in the cyclical dynamics. Using a similar post-hoc procedure, Faria et al. (2012) also find a lag of about five years between prevalence counts and the growth in N_e over time for HIV-1 CRF02_AG in Cameroon. Although these studies illustrate opportunities for integrating time series covariates, they also point at complications in interpreting N_e estimates as "effective number of infections." Frost and Volz (2010) specifically address this issue and note that the rate of coalescence is primarily driven by new transmission — that is the incidence — and only indirectly by the number of infected individuals through sampling effects (Frost and Volz 2010).

By avoiding the concept of N_e, birth–death modeling represents an interesting alternative that is gaining popularity as a tree-generative process. Originally introduced in phylodynamics by Stadler et al. (2012), the stochastic linear birth–death process is parametrized by a rate at which infected individuals transmit and the rate at which they turn noninfectious (either by death or recovery). For a completely susceptible population, the ratio of these rates should in principle quantify the well-known R_0 parameter from mathematical epidemiology. Unlike the coalescent, birth–death models do not need to assume a small sample size from a large population because the sampling proportion is treated as a separate parameter in these models. Following the evolution in coalescent modeling, the standard birth–death model has also been extended to a nonparametric version that flexibly models varying infection rates (Stadler et al. 2013). Despite the flexible parameterizations of coalescent and birth–death modeling approaches, they have yet to be extended to allow the incorporation of covariates.

Arguably the most concrete step toward integrating covariates with tree-generative processes has been made by Rasmussen et al. (2011). The authors fitted a stochastic, nonlinear model of disease transmission to a combination of epidemiological data and phylogenetic

coalescence events. This approach follows the SIR coalescent modeling by Volz et al. (2009) and resorts to particle MCMC to fit state-space models to genealogies in order to avoid the need for an analytical likelihood function. Simulations suggest that genealogical information alongside of time series data may improve the estimates if the time series data suffers from observation error or from variation in reporting practices. The ability to combine coalescent information with covariates, or contrast it against them, also offers the opportunity to investigate which ecological factors need to be considered by the coalescent model to appropriately capture pathogen population dynamics. In a study of dengue serotype I in southern Vietnam, Rasmussen et al. (2014) demonstrate that nonparameteric coalescent modeling does not reproduce the highly seasonal incidence patterns observed in hospitalization data. In this case, incorporating spatial structure was critical to recapitulate seasonal fluctuations consistent with the hospitalization data. This illustrates that the panmictic population assumption in coalescent modeling can be problematic, and that inferring temporal epidemic dynamics cannot always be divorced from population structure, which we discussed in the context of trait or structured coalescent inference. Intriguingly, the authors also show the importance of accounting for vector population, suggesting a general need to consider ecological complexities in pathogen epidemiology.

RECONSTRUCTING TRANSMISSION TREES

In the first section of this review, we focused on the integration of time and location with genetic data in a general phylogenetic framework, but these sources of information are sometimes also combined in different ways when the aim is to reconstruct transmission trees. Recreating individual routes of transmission in infectious disease outbreaks is a challenging problem, but one of longstanding interest in infectious disease epidemiology that essentially goes back to the famous work by John Snow who traced the source of a cholera outbreak in London in 1854 (Snow 1855). A clear break with the inference of large-scale transmission dynamics is the requirement of a dense — if not complete — sampling from the pathogen population, at least for the initial developments in this direction.

Person-to-person transmission is frequently represented as a spanning tree, and various methods have therefore been proposed to find the spanning tree between sampled sequences that is most compatible with the genetic data, for example by minimizing a set of edge weights (see e.g., Jombart et al. (2011)). Cottam et al. (2008) introduce epidemiological data by first retrieving the set of transmission trees that are consistent with the available genetic data, and then evaluating these trees using data on their relative timings to identify the most plausible transmission history. A more formal integration of epidemiological and genetic data can be found in the work of Ypma et al. (2012), who apply their Bayesian procedure to temporal, geographical, and genetic data on poultry farms infected in an epidemic of avian influenza A (H7N7) in The Netherlands in 2003. Although genetic and epidemiological data arise from the same process, this approach assumes they are independent. To address this limitation, Morelli et al. (2012) develop an MCMC approach that integrates epidemiological data and pathogen sequences from infected hosts to estimate transmission trees and infection dates in a more coherent way.

Generally, disease dynamics are only partially observed, but methods based on transmission trees are not designed to handle large numbers of missing infections, and therefore require a dense sample of infected hosts from the outbreak. To remedy this, Mollentze et al. (2014) propose a generalization of the algorithm of Morelli et al. (2012) to allow its application to any directly transmitted disease and enable reconstruction of partially observed transmission trees as well as to estimate the number of cases missing from the sample. The extension specifically allows accommodation of a wide variety of spatial transmission patterns and allows multiple unobserved cases to arise anywhere in both space and time within the set of inferred transmissions. An application to endemic rabies virus in a province of South Africa shows that the method offers a better insight into the spatial epidemiological patterns (Mollentze et al. 2014). In related work, Jombart et al. (2014) propose a Bayesian framework that does not require all cases to be observed or assume a single introduction event at the origin of an outbreak. Their framework allows for the estimation of dates of infections, mutation rates, separate introductions of the pathogen, the presence of unobserved cases and the transmission tree, as well as the effective reproduction number over time, thereby overcoming the limitations of their previous method (Jombart et al. 2011) and of the methods of Ypma et al. (2012) and Morelli et al. (2012). The authors apply their method to the 2003 Severe Acute Respiratory Syndrome (SARS) outbreak in Singapore, providing new insights into the early stage of the epidemic.

RECENT DEVELOPMENTS AND FUTURE PERSPECTIVES

In this review, we highlighted several aspects of data integration in phylodynamics with a particular focus on the connection between sequences and traits. As mentioned in the respective sections, the initial applications of relatively simple but computationally efficient Bayesian implementations of random walk models in the field of pathogen evolution and epidemiology focused on reconstructions of spatial spread. There are, however, important caveats that need to be considered for phylogeographic applications, in particular related to sampling. First, spatial coverage may be restricted to particular geographic areas which

limits comprehensive understanding of the spatial spread dynamics. For example, although large genetic datasets have become available to study the global seasonal dynamics of all human influenza lineages, the sampling remains limited for Africa, Central America, the Middle East, and Russia (Bedford et al. 2015). This complicates assessment of their specific role in the influenza source-sink dynamics. In the discrete ancestral reconstruction approach, heterogeneity in sampling among the locations that can be represented in the analysis may also bias the results. Overrepresentation of location states is likely to be associated with high rates out of, or into, these specific states, and these transition rates will influence ancestral state probabilities in the phylogenies. Although subsampling may be used to investigate the sensitivity to sampling effects, it remains difficult if not impossible to remove sampling bias in these approaches. The structured coalescent methods we discuss as alternative approaches are less sensitive to sampling bias because they take the diversity in each location or "deme" into consideration, and may therefore have a more promising future in phylogeographic studies. In the continuous diffusion approach, the parameterization does not have such a direct connection with the sampling, but other restrictions are important to consider in this framework. Despite the ability to relax the assumption of a constant diffusion rate, the process still assumes a relationship between dispersal and geographic distance, which may provide a poor fit to pathogens that disperse through modern human mobility. In addition, standard Brownian diffusion considers distances in Euclidean space and therefore ignores the spherical nature of the globe (but see Bouckaert (2015)). Alternatively, Barton et al. (2013) discuss the extension of the coalescent to spatially structured populations, where populations are not disjointly subdivided, but instead are distributed across a spatial continuum. The authors point out that in such a setting, the system of stochastic ordinary differential equations known as Kimura's stepping stone model (Kimura 1953) has no solution in two spatial dimensions, as the system of coalescing random walks which describes the genealogy converges to a system of Brownian motions that will never meet.

Although the examples we discussed generally involve a single trait, there are no particular limitations on the number of traits that can be incorporated in a single analysis. Trovão et al. (2015) leverage this capacity to examine the contributions of different hosts to the spatial spread of avian influenza H5N1. Specifically, they jointly infer discrete host switching among members of different avian families/superorders and dispersal in continuous space underlying the H5N1 expansion across Eurasia. Although this study models the host and spatial dynamics as independent processes, it leads to host-specific summaries of spatial spread and testable differences among them. In addition to combining multiple traits, the possibility to incorporate covariates in discrete diffusion processes also prompts exploration of more detailed ecological information in pathogen

evolution and epidemiology. In a spatial context, cheap and mobile global positioning systems are now widely adopted in the recording of infectious disease spread, but also the variables that are associated with geo-located disease data (e.g., environmental, infrastructural, and socio-economic) are increasingly being characterized in great detail and distributed as publicly available datasets (see e.g., http://www.worldpop.org.uk). In addition, particular forms of animal trade or human mobility are extensively documented, or they can be modeled based on proxies such as the movement of marked banknotes (Brockmann et al. 2006), and anonymized mobile phone call records (González et al. 2008).

The example of trait combinations (Trovão et al. 2015) also raises the question of how correlations can be measured formally between different data types. This is a particularly pertinent question for phenotypic traits in evolutionary biology, which explains correlations as the result of genetic constraints or selective effects. The multivariate diffusion approach naturally models and estimates correlation among continuous traits, but for different data types (e.g., binary and multinomial ordered, or unordered data) additional modeling is required. Building on the phylogenetic threshold model (Felsenstein 2005), Cybis et al. (2015) demonstrate how to efficiently infer the evolution of all types of traits, and the correlations among them, through latent liability modeling in the BEAST framework. One of their applications to pathogens targeted the phenotypic correlation among the amino acid sites of the antigenic epitopes of the influenza hemagglutin surface protein, which plays an important role in antigenic drift dynamics. They find strong correlations among 11 sites in epitope A and B, including all sites that have been experimentally shown to be responsible for evolving antigenic novelty (Koel et al. 2013). Such methodologies open up new research opportunities at the interface of genotype and phenotype in pathogen evolution, and in conjunction with antigenic (Bedford et al. 2014) and integrated genetic and human mobility modeling (Lemey et al. 2014) in the same Bayesian framework, this may lead to holistic approaches in influenza phylodynamics for example. The latent liability or threshold model finds its nonphylogenetic roots in a model that goes back to Sewall Wright (Wright 1934), and has been used regularly in the pedigree analysis of discrete traits (Gianola 1982). So, in addition to the PMM we discussed in the context of trait heritability, this highlights another explicit link between phylogenetics and quantitive genetics.

Further in the phenotypic context, other assumptions than the constant variance also require attention when modeling trait evolution via Brownian motion along a phylogeny. Assuming a zero-mean displacement, for example, postulates that a single phenotypic value randomly increases or decreases each generation, which can only be expected for phenotypes undergoing random genetic drift or fluctuating directional selection. Under other forms of natural selection however, traits

may evolve toward some optimal value. To model such consistent selection toward a single optimum trait value, the Ornstein-Uhlenbeck (OU) model has been proposed as a mean-reverting extension of Brownian motion (Hansen 1997). Although this has proven useful to identify stabilizing selection regimes, a central tendency model does not offer a general solution to appropriately fit the complexity in trait evolution. Further advances in Bayesian trait evolutionary modeling may provide the flexibility to model displacements in traits along tree branches as a generalized stochastic process that allows the data to inform the degree of complexity required in the process.

Data integration in phylodynamic modeling is also transforming the inference of transmission dynamics. Although this is still somewhat in its infancy in the analyses of large-scale transmission dynamics based on coalescent or birth–death models, the possibility of using genetic data and time series data in tandem has been clearly demonstrated (Rasmussen et al. 2011), as well as the need to incorporate environmental stochasticity to accurately capture disease dynamics in some cases (Rasmussen et al. 2014). These developments are generally divorced from the sequence evolutionary process and still await their implementation in the commonly used Bayesian statistical software packages like BEAST (Drummond et al. 2012). Although BEAST implements a wide range of tree-generative processes, including time-variable coalescent and birth–death models, these have not yet been connected to covariate or time series data. Related to this, however, ongoing work in coalescent modeling acknowledges that sampling times may probabilistically depend on effective population size (Karcher et al. 2015), and preferential sampling is now being taken into account by modeling the sampling times as an inhomogeneous Poisson process dependent on effective population size.

Developments in transmission tree reconstruction have already been pursuing data integration for a longer time, and they are evolving toward methods that consider the likelihood of observing sequence and epidemiological data for a given transmission tree, and increasingly try to accommodate missing data. Many of these developments are however scattered in the field, and at least for the sequence evolutionary modeling, they could also benefit from an implementation in an integrated statistical framework. In addition, more complex relationships between phylogenies and transmission trees need to be taken into account (akin to gene/species tree modeling, Vrancken et al. (2014b)), in particular when within-host evolution plays an important role. Different aspects of within and between host evolution are not always easily reconciled, and this has been identified as one of the challenges in phylodynamic inference (Frost et al. 2015).

Accounting for nonvertical evolution represents another major challenge as all the methods we have discussed rely on a strictly bifurcating evolutionary process. Nonvertical evolution can be a prominent evolutionary force in many pathogen populations, for example in the form of recombination and reassortment in viruses and horizontal gene transfer in bacteria. Recombination analyses in viruses, for example based on the identification of tree incongruence, generally aim at avoiding the impact of recombination in downstream phylogenetic inferences (Martin et al. 2011). However, an alternative approach may be to explicitly accommodate nonvertical evolution through an ancestral recombination graph, which simultaneously describes vertical and nonvertical evolutionary events (Hudson 1983). Although the graph is a well-known model in coalescent inference, it has only been recently introduced as an explicit structure for analyzing phylogenetic data in BEAST (Bloomquist and Suchard 2010), and further modeling and computational development is required to promote its widespread use.

REFERENCES

Alizon S., Fraser C. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology 10:49.

Alizon S., von Wyl V., Stadler T., Kouyos R.D., Yerly S., Hirschel B., Boni J., Shah C., Klimkait T., Furrer H., Rauch A., Vernazza P.L., Bernasconi E., Battegay M., Bürgisser P., Telenti A., Günthard H.F., Bonhoeffer S., the Swiss Cohort Study. 2010. Phylogenetic approach reveals that virus genotype largely determines HIV set-point viral load. PLoS Path. 6:e1001123.

Ayres D.L., Darling A., Zwickl D.J., Beerli P., Holder M.T., Lewis P.O., Huelsenbeck J.P., Ronquist F., Swofford D.L., Cummings M.P., Rambaut A., Suchard M.A. 2012. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst. Biol. 61:170–173.

Baele G., Lemey P. 2013. Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. Bioinformatics 29:1970–1979.

Bahl J., Nelson M.I., Chan K.H., Chen R., Vijaykrishna D., Halpin R.A., Stockwell T.B., Lin X., Wentworth D.E., Ghedin E., Guan Y., Peiris J.S.M, Riley S., Rambaut A., Holmes E.C., Smith G.J.D. 2011. Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. Proc. Natl. Acad. Sci. USA 108:19359–19364.

Baize S., Pannetier D., Oestereich L., Rieger T., Koivogui L., Magassouba N., Soropogui B., Sow M.S., Keïta S., De Clerck H., Tiffany A., Dominguez G., Loua M., Traoré A., Kolié M., Malano E.R., Heleze E., Bocquin A., Mély S., Raoul H., Caro V., Cadar D., Gabriel M., Pahlmann M., Tappe D., Schmidt-Chanasit J.,

Impouma B., Diallo A.K., Formenty P., Van Herp M., Günther S. 2014. Emergence of Zaire ebola virus disease in Guinea. N. Engl. J. Med. 371:1418–1425.

Barton N.H., Etheridge A.M., Véber A. 2013. Modelling evolution in a spatial continuum. J. Stat. Mech. P01002.

Bedford T., Riley S., Barr I.G., Broor S., Chadha M., Cox N.J., Daniels R.S., Gunasekaran C.P., Hurt A.C., Kelso A., Klimov A., Lewis N.S., Li X., McCauley J.W., Odagiri T., Potdar V., Rambaut A., Shu Y., Skepner E., Smith D.J., Suchard M.A., Tashiro M., Wang D., Xu X., Lemey P., Russell C.A. 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. Nature 523:217–220.

Bedford T., Suchard M.A., Lemey P., Dudas G., Gregory V., Hay A.J., McCauley J.W., Russell C.A., Smith D.J., Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. eLife 3:e01914.

Bennett S.N., Drummond A.J., Kapan D.D., Suchard M.A., Muñoz-Jordán J.L., Pybus O.G., Holmes E.C., Gubler D.J. 2010. Epidemic dynamics revealed in Dengue evolution. Mol. Biol. Evol. 27:811–818.

Biek R., Henderson J.C., Waller L.A., Rupprecht C.E., Real L.A. 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proc. Natl. Acad. Sci. USA 104:7993–7998.

Bielejec F., Baele G., Vrancken B., Suchard M.A., Rambaut A., Lemey P. 2016. Spread3: Interactive visualization of spatiotemporal history and trait evolutionary processes. Mol. Biol. Evol.

Bloomquist E.W., Lemey P., Suchard M.A. 2010. Three roads diverged? Routes to phylogeographic inference. Trends Ecol. Evol. 25:626–632.

Bloomquist E.W., Suchard M.A. 2010. Unifying vertical and nonvertical evolution: a stochastic arg-based framework. Syst. Biol. 59:27–41.

Bouckaert R., Lemey P., Dunn M., Greenhill S.J., Alekseyenko A.V., Drummond A.J., Gray R.D., Suchard M.A., Atkinson Q.D. 2012. Mapping the origins and expansion of the indo-european language family. Science 337:957–960.

Bouckaert R. 2015. Phylogeography by diffusion on a sphere. bioRxiv doi: http://dx.doi.org/10.1101/016311.

Brockmann D., Helbing D. 2013. The hidden geometry of complex, network-driven contagion phenomena. Science 342:1337–1342.

Brockmann D., Hufnagel L., Geisel T. 2006. The scaling laws of human travel. Nature 439:462–465.

Bryja J., Mikula O., Šumbera R., Meheretu Y., Aghová T., Lavrenchenko L.A., Mazoch V., Oguge N., Mbau J.S., Welegerima K., Amundala N., Colyn M., Leirs H., Verheyen E. 2014. Pan-African phylogeny of mus (subgenus nannomys) reveals one of the most successful mammal radiations in Africa. BMC Evol. Biol. 14:256.

Comas I., Coscolla M., Luo T., Borrell S., Holt K.E., Kato-Maeda M., Parkhill J., Malla B., Berg S., Thwaites G., Yeboah-Manu D., Bothamley G., Mei J., Wei L., Bentley S., Harris S.R., Niemann S., Diel R., Aseffa A., Gao Q., Young D., Gagneux S. 2013. Out-of-Africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans. Nat. Genet. 45:1176–1182.

Cottam E.M., Thébaud G., Wadsworth J., Gloster J., Mansley L., Paton D.J., King D.P., Haydon D.T. 2008. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc. R. Soc. B 275:887–895.

Cotten M., Watson S.J., Zumla A.I., Makhdoom H.G., Palser A.L., Ong S.H., Al Rabeeah A.A., Alkaheem R.F., Assiri A., Al-Tawfiq J.A., Albarrak A., Barry M., Shibl A., Alrabiah F.A., Hajjar S., Balkhy H.H., Flemban H., Rambaut A., Kellam P., Memish Z.A. 2014. Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. mBio 5:e01062–13.

Cybis G.B., Sinsheimer J.S., Bedford T., Mather A.E., Lemey P., Suchard M.A. 2015. Assessing phenotypic correlations through the multivariate phylogenetic latent liability model. Ann. Appl. Stat. 9:969–991.

Cybis G.B., Sinsheimer J.S., Lemey P., Suchard M.A. 2013. Graph hierarchies for phylogeography. Phil. Trans. R. Soc. B 368:20120206.

Dearlove B.L., Cody A.J., Pascoe B., Méric G., Wilson D.J., Sheppard S.K. 2015. Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. The ISME Journal (in press).

De Maio N., Wu C.-H., O'Reilly K.M., Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. PLoS Genet. 11:e1005421.

Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006a. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A.J., Pybus O.G., Rambaut A., Forsberg R., Rodrigo A.G. 2003. Measurably evolving populations. Trends Ecol. Evol. 18: 481–488.

Drummond A.J., Rambaut A., Shapiro B., Pybus O.G. 2006b. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185–1192.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Drummond A.J., Suchard M.A. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8:114.

Dudas G., Rambaut A. 2014. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. PLoS Currents: Outbreaks 6.

Dunn T., Kruspe N., Burenhult N. 2013. Time and place in the prehistory of the Aslian languages. Human Biol. 85:383–400.

Edo-Matas D., Lemey P., Tom J.A., Serna-Bolea C., van den Blink A.E., 't Wout A.B.V., Schuitemaker H., Suchard M.A. 2011. Impact of CCR5delta32 host genetic background and disease progression on HIV-1 intrahost evolutionary processes: efficient hypothesis testing through hierarchical phylogenetic models. Mol. Biol. Evol. 28:1605–1616.

Edwards A.W.F., Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. In: Phenetic and Phylogenetic Classification V.H. Heywood, J. McNeill, editors. London: Systematics Association pub. no. 6. p 67–76.

Edwards C.J., Suchard M.A., Lemey P., Welch J.J., Barnes I., Fulton T.L., Barnett R., O'Connell T.C., Coxon P., Monaghan N., Valdiosera C.E., Lorenzen E.D., Willerslev E., Baryshnikov G.F., Rambaut A., Thomas M.G., Bradley D.G., Shapiro B. 2011. Ancient hybridization and an Irish origin for the modern polar bear matriline. Curr. Biol. 21:1251–1258.

Ewing G., Nicholls G., Rodrigo A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. Genetics 168:2407–2420.

Faria N.R., Suchard M.A., Abecasis A., Sousa J.D., Ndembi N., Bonfim I., Camacho R.J., Vandamme A.-M., Lemey P. 2012. Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. Infect. Genet. Evol. 12:453–460.

Faria N.R., Suchard M.A., Rambaut A., Streicker D.G., Lemey P. 2013. Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. Phil. Trans. R. Soc. B 368:20120196.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates, Inc.

Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between and within species. Phil. Trans. R. Soc. B: Biol. Sci. 360:1427–1434.

Fraser C., Donnelly C.A., Cauchemez S., Hanage W.P., Van Kerkhove M.D., Hollingsworth T.D., Griffin J., Baggaley R.F., Jenkins H.E., Lyons E.J., Jombart T., Hinsley W.R., Grassly N.C., Balloux F., Ghani A.C., Ferguson N.M., Rambaut A., Pybus O.G., Lopez-Gatell H., Alpuche-Aranda C.M., Chapela I.B., Zavala E.P., Guevara D.M.E., Checchi F., Garcia E., Hugonnet S., Roth C., The WHO Rapid Pandemic Assessment Collaboration. 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324:1557–1561.

Fraser C., Lythgoe K., Leventhal G.E., Shirreff G., Hollingsworth T.D., Alizon S., Bonhoeffer S. 2014. Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. Science 343:1243727.

Freckleton R.P. 2012. Fast likelihood calculations for comparative analyses. Methods Ecol. Evol. 3:940–947.

Frost S.D.W., Pybus O.G., Gog J.R., Viboud C., Bonhoeffer S., Bedford T. 2015. Eight challenges in phylodynamic inference. Epidemics 10: 88–92.

Frost S.D.W., Volz E.M. 2010. Viral phylodynamics and the search for an "effective number of infections". Phil. Trans. R. Soc. B 365: 1879–1890.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Gianola D. 1982. Theory and analysis of threshold characters. J. Anim. Sci. 54:1079–1096.

Gill M.S., Lemey P., Faria N.R., Rambaut A., Shapiro B., Suchard M.A. 2013. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol. Biol. Evol. 30:713–724.

González M., Hidalgo C., Barabási A. 2008. Understanding individual human mobility patterns. Nature 453:779–782.

Grafen A. 1989. The phylogenetic regression. Phil. Trans. R. Soc. Lond. B 326:119–157.

Grenfell B.T., Pybus O.G., Gog J.R., Wood J.L.N., Daly J.M., Mumford J.A., Holmes E.C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327–332.

Haeckel E. 1866. Generelle morphologie der organismen : allgemeine Grundzüge der organischen formen-wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte descendenz-theorie. Berlin: Georg Reimer.

Hansen T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hartfield M., Bull R., White P.A., Lloyd A., Luciani F., Alizon S., on behalf of the HITS investigators. 2014. Evidence that hepatitis C virus genome partly controls infection outcome. Evol. Appl. 7:533–547.

Hedge J., Lycett S.J., Rambaut A. 2013. Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol. Lett. 9:20130331.

Henderson C.R. 1984. Applications of linear models in animal breeding. Ontario: University of Guelph, Guelph.

Hodcroft E., Hadfield J.D., Fearnhill E., Phillips A., Dunn D., O'Shea S., Pillay D., Leigh Brown A.J., on behalf of the UK HIV Drug Resistance Database and the UK CHIC Study. 2014. The contribution of viral genotype to plasma viral set-point in HIV infection. PLoS Path. 10:e1004112.

Hofreiter M., Serre D., Poinar H.N., Kuch M., Päabo S. 2001. Ancient DNA. Nat. Rev. Genet. 2:353–359.

Holmes E.C., Nee S., Rambaut A., Garnett G.P., Harvey P.H. 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. Phil. Trans. R. Soc. Lond. B 349:33–40.

Housworth E.A., Martins E.P., Lynch M. 2004. The phylogenetic mixed model. Am. Nat. 163:84–96.

Ho S.Y.W., Jermiin L.S. 2007. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623–637.

Hudson R.R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183–201.

Hudson R.R. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7:1–44.

Jombart T., Cori A., Didelot X., Cauchemez S., Fraser C., Ferguson N. 2014. Bayesian reconstruction of disease outbreaks by combining epidemiological and genomic data. PLoS Comp. Biol. 10:e1003457.

Jombart T., Eggo R.M., Dodd P.J., Balloux F. 2011. Reconstructing disease outbreaks from genetic data: a graph approach. Heredity 106:383–390.

Karcher M.D., Palacios J.A., Bedford T., Suchard M.A., Minin V.N. 2015. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. arXiv:1510.00775 [stat.ME].

Kendall D.G. 1948. On the generalized "birth-and-death" process. Ann. Math. 19:1–15.

Kimura M. 1953. Stepping stone model of population. Ann. Rep. Nat. Inst. Genet. 3:62–63.

Kingman J.F.C. 1982. On the genealogy of large populations. J. Appl. Probab. 19:27–43.

Koel B.F., Burke D.F., Bestebroer T.M., van der Vliet S., Zondag G.C.M., Vervaet G., Skepner E., Lewis N.S., Spronken M.I.J., Russell C.A., Eropkin M.Y., Hurt A.C., Barr I.G., de Jong J.C., Rimmelzwaan G.F., Osterhaus A.D.M.E., Fouchier R.A.M., Smith D.J. 2013. Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. Science 342:976–979.

Lai J., Jin J., Kubelka J., Liberles D.A. 2012. A phylogenetic analysis of normal modes evolution in enzymes and its relationship to enzyme function. J. Mol. Biol. 422:442–459.

Lee M.S.Y., Cau A., Naish D., Dyke G.J. 2014. Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. Science 345:562–566.

Lee S., Hasegawa T. 2013. Evolution of the Ainu language in space and time. PLoS ONE 8:e62243.

Lemey P., Rambaut A., Bedford T., Faria N., Bielejec F., Baele G., Russell C.A., Smith D.J., Pybus O.G., Brockmann D., Suchard M.A. 2014. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. PLoS Path. 10:e1003932.

Lemey P., Rambaut A., Drummond A.J., Suchard M.A. 2009a. Bayesian phylogeography finding its roots. PLoS Comp. Biol. 5:e1000520.

Lemey P., Rambaut A., Welch J.J., Suchard M.A. 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27:1877–1885.

Lemey P., Salemi M., Vandamme A.-M., eds. 2009b. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing. Cambridge, UK.

Lemmon A., Lemmon E. 2008. A likelihood framework for estimating phylogeographical history on a continuous landscape. Syst. Biol. 57:544–561.

Li W.L.S., Drummond A.J. 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 29:751–761.

Longdon B., Hadfield J.D., Day J.P., Smith S.C.L., McGonigle J.E., Cogni R., Cao C., Jiggins F.M. 2015. The causes and consequences of changes in virulence following pathogen host shifts. PLoS Path. 11:e1004728.

Longdon B., Hadfield J.D., Webster C.L., Obbard D.J., Jiggins F.M. 2011. Host phylogeny determines viral persistence and replication in novel hosts. PLoS Path. 7:e1002260.

Lorion J., Kiel S., Faure B., Kawato M., Ho S.Y.W., Marshall B., Tsuchida S., Miyazaki J.-I., Fujiwara Y. 2013. Adaptive radiation of chemosymbiotic deep-sea mussels. Proc. R. Soc. B 280:20131243.

Lu L., Lycett S.J., Leigh Brown A.J. 2014. Reassortment patterns of avian influenza virus internal segments among different subtypes. BMC Evol. Biol. 14:16.

Lynch M., Walsh B. 1998. Genetics and analysis of quantitative traits. Sunderland, (MA): Sinauer.

Lythgoe K.A., Fraser C. 2012. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. Proc. Biol. Sci. 279:3367–3375.

Maddison W.P., Donogue M.J., Maddison D.R. 1984. Outgroup analysis and parsimony. Syst. Zool. 33:83–103.

Martins E.P. 1994. Estimating the rate of phenotypic evolution from comparative data. Am. Nat. 144:193–209.

Martin D.P., Lemey P., Posada D. 2011. Analysing recombination in nucleotide sequences. Mol. Ecol. Resour. 11:943–955.

Mather A.E., Reid S.W.J., Maskell D.J., Parkhill J., Fookes M.C., Harris S.R., Brown D.J., Coia J.E., Mulvey M.R., Gilmour M.W., Petrovska L., de Pinna E., Kuroda M., Akiba M., Izumiya H., Connor T.R., Suchard M.A., Lemey P., Mellor D.J., Haydon N.R.T.D.T. 2013. Distinguishable epidemics of multidrug-resistant salmonella typhimurium DT104 in different hosts. Science 341:1514–1517.

Meng Y., Nie Z.-L., Deng T., Wen J., Yang Y.-P. 2014. Phylogenetics and evolution of phyllotaxy in the Solomon's seal genus polygonatum (asparagaceae:polygonateae). Bot. J. Linn. Soc. 176:435–451.

Minin V.M., Bloomquist E.W., Suchard M.A. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Minin V.M., Suchard M.A. 2008. Fast, accurate and simulation-free stochastic mapping. Phil. Trans. R. Soc. Lond. B Biol. Sci. 363:3985–3995.

Molak M., Suchard M.A., Ho S.Y., Beilman D.W., Shapiro B. 2015. Empirical calibrated radiocarbon sampler: a tool for incorporating radiocarbon-date and calibration error into Bayesian phylogenetic analyses of ancient DNA. Mol. Ecol. Resour. 15:81–86.

Mollentze N., Nel L.H., Townsend S., le Roux K., Hampson K., Haydon D.T., Soubeyrand S. 2014. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc. R. Soc. B 281:20133251.

Morelli M.J., Thébaud G., Chadœuf J., King D.P., Haydon D.T., Soubeyrand S. 2012. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comp. Biol. 8:e1002768.

Nelson M.I., Viboud C., Vincent A.L., Culhane M.R., Detmer S.E., Wentworth D.E., Rambaut A., Suchard M.A., Holmes E.C., Lemey P. 2015. Global migration of influenza A viruses in swine. Nat. Commun. 6:6696.

Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. J. Math. Biol. 29:59–75.

Nylander J.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J.L. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.

Oh M., Raftery A. 2001. Bayesian multidimensional scaling and choice of dimension. J. Am. Stat. Assoc. 96:1031–1044.

Pääbo S., Poinar H., Serre D., Jaenicke-Després V., Hebler J., Rohland N., Kuch M., Krause J., Vigilant L., Hofreiter M. 2004. Genetic analyses from ancient DNA. Annu. Rev. Genet. 38:645–679.

Pagel M., Meade A., Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst. Biol. 53:673–684.

Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scripta 26:331–348.

Pagel M. 1999a. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Pagel M. 1999b. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Syst. Biol. 48:612–622.

Park D.J., Dudas G., Wohl S., Goba A., Whitmer S.L., Andersen K.G., Sealfon R.S., Ladner J.T., Kugelman J.R., Matranga C.B., Winnicki S.M., Qu J., Gire S.K., Gladden-Young A., Jalloh S., Nosamiefan D., Yozwiak N.L., Moses L.M., Jiang P.P., Lin A.E., Schaffner S.F., Bird B., Towner J., Mamoh M., Gbakie M., Kanneh L., Massally D.K.J.L., Kamara F.K., Konuwa E., Sellu J., Jalloh A.A., Mustapha I., Foday M., Yillah M., Erickson B.R., Sealy T., Blau D., Paddock C., Brault A., Amman B., Basile J., Bearden S., Belser J., Bergeron E., Campbell S., Chakrabarti A., Dodd K., Flint M., Gibbons A., Goodman C., Klena J., McMullan L., Morgan L., Russell B., Salzer J., Sanchez A., Wang D., Jungreis I., Tomkins-Tinch C., Kislyuk A., Lin M.F., Chapman S., MacInnis B., Matthews J., Bochicchio J., Hensley L.E., Kuhn J.H., Nusbaum C., Schieffelin J.S., Birren B.W., Forget M., Nichol S.T., Palacios G.F., Ndiaye D., Happi C., Gevao S.M., Vandi M.A., Kargbo B., Holmes E.C., Be T. 2015. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. Cell 161:1516–1526.

Pérez-Losada M., Høeg J.T., Crandall K.A. 2012. Deep phylogeny and character evolution in thecostraca (crustacea: Maxillopoda). Integr. Comp. Biol. 52:430–442.

Pybus O.G., Rambaut A., Harvey P.H. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155:1429–1437.

Pybus O.G., Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nat. Rev. Genet. 10:540–550.

Pybus O.G., Suchard M.A., Lemey P., Bernardin F.J., Rambaut A., Crawford F.W., Gray R.R., Arinaminpathy N., Stramer S.L., Busch M.P., Delwart E.L. 2012. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proc. Natl. Acad. Sci. USA 109:15066–15071.

Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16:395–399.

Ramírez H., Reina R., Bertolotti L., Cenoz A., Hernández M.-M., San Morán B., Glaria I., de Andrés X., Crespo H., Jáuregui P., Benavides J., Polledo L., Pérez V., García-Marín J.F., Rosati S., Amorena B., de Andrés D. 2012. Study of compartmentalization in the visna clinical form of small ruminant lentivirus infection in sheep. BMC Vet. Res. 8:8.

Rasmussen D.A., Boni M.F., Koelle K. 2014. Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol. Biol. Evol. 31:258–271.

Rasmussen D.A., Ratmann O., Koelle K. 2011. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Comp. Biol. 7:e1002136.

Ronquist F., Teslenko M., van der Mark P., Ayres D.L., Darling A., Höhna S., Larget B., Liu L., Suchard M.A., Huelsenbeck J.P. 2012. Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Sanmartín I., van der Mark P., Ronquist F. 2008. Inferring dispersal: a Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. J. Biogeogr. 35:428–449.

Scheeff E.D., Bourne P.E. 2005. Structural evolution of the protein kinase–like superfamily. PLoS Comp. Biol. 11:e49.

Schluter D., Price T., Mooers A., Ludwig D. 1997. Likelihood of ancestor states in adaptive radiation. Int. J. Org. Evol. 51:1699–1711.

Schrago C.G. 2014a. The effective population sizes of the anthropoid ancestors of the human–chimpanzee lineage provide insights on the historical biogeography of the great apes. Mol. Biol. Evol. 31:37–47.

Schrago C.G. 2014b. Estimation of the ancestral effective population sizes of African great apes under different selection regimes. Genetica 142:273–280.

Shapiro B., Drummond A.J., Rambaut A., Wilson M.C., Matheus P.E., Sher A.V., Pybus O.G., Gilbert M.T.P., Barnes I., Binladen J., Willerslev E., Hansen A.J., Baryshnikov G.F., Burns J.A., Davydov S., Driver J.C., Froese D.G., Harington C.R., Keddie G., Kosintsev P., Kunz M.L., Martin L.D., Stephenson R.O., Storer J., Tedford R., Zimov S., Cooper A. 2004. Rise and fall of the Beringian steppe bison. Science 306:1561–1565.

Shapiro B., Ho S.Y.W., Drummond A.J., Suchard M.A., Pybus O.G., Rambaut A. 2011. A Bayesian phylogenetic method to estimate unknown sequence ages. Mol. Biol. Evol. 28:879–887.

Smith D.L. 2005. Spatial heterogeneity in infectious disease epidemics. In: Ecosystem function in heterogeneous landscapes. G.M. Lovett, C.G. Jones, M.G. Turner, K.C. Weathers, editors. Springer. pp. 137–164.

Smith D.J., Lapedes A.S., de Jong J.C., Bestebroer T.M., Rimmelzwaan G.F., Osterhaus A.D.M.E., Fouchier R.A.M. 2004. Mapping the antigenic and genetic evolution of influenza virus. Science 305:371–376.

Smith G.J.D., Vijaykrishna D., Bahl J., Lycett S.J., Worobey M., Pybus O.G., Ma S.K., Cheung C.L., Raghwani J., Bhatt S., Peiris J.S.M., Guan Y., Rambaut A. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459:1122–1125.

Snow J. 1855. On the mode of communication of cholera. London, England: John Churchill.

Stadler T., Kouyos R., von Wyl V., Yerly S., Böni J., Bürgisser P., Klimkait T., Joos B., Rieder P., Xie D., Günthard H.F., Drummond A.J., Bonhoeffer S., the Swiss HIV Cohort Study. 2012. Estimating the basic reproductive number from viral sequence data. Mol. Biol. Evol. 29:347–357.

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Natl. Acad. Sci. USA 110:228–233.

Stadler T. 2010. Sampling-through-time in birth-death trees. J. Theor. Biol. 267:396–404.

Streicker D.G., Lemey P., Velasco-Villa A., Rupprecht C.E. 2012. Rates of viral evolution are linked to host geography in bat rabies. PLoS. Path. 8:e1002720.

Streicker D.G., Turmelle A.S., Vonhof M.J., Kuzmin I.V., McCracken G.F., Rupprecht C.E. 2010. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. Science 329:676–679.

Strickland S.L., Rife B.D., Lamers S.L., Nolan D.J., Veras N.M.C., Prosperi M.C.F., Burdo T.H., Autissier P., Nowlin B., Goodenow M.M., Suchard M.A., Williams K.C., Salemi M. 2014. Spatiotemporal dynamics of simian immunodeficiency virus brain infection in CD8$^+$ lymphocyte-depleted rhesus macaques with neuroAIDS. J. Gen. Virol. 95:2784–2795.

Strimmer K., Pybus O.G. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. Mol. Biol. Evol. 18:2298–2305.

Suchard M.A., Kitchen C.M., Sinsheimer J.S., Weiss R.E. 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst. Biol. 52:649–664.

Suchard M.A., Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. Bioinformatics 25:1370–1376.

Swenson U., Munzinger J., Lowry II P.P., Cronholm B., Nylinder S. 2015. Island life - classification, speciation and cryptic species of pycnandra (sapotaceae) in New Caledonia. Bot. J. Linn. Soc. 179:57–77.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Some mathematical questions in biology: DNA sequence analysis. M.S. Waterman, editor. Providence (RI): American Mathematical Society. p. 57–86

Terra-Araujo M.H., de Faria A.D., Vicentini A., Nylinder S., Swenson U. 2015. Species tree phylogeny and biogeography of the neotropical genus pradosia (sapotaceae, chrysophylloideae). Mol. Phyl. Evol. 87:1–13.

Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Toon A., Crisp M.D., Gamage H., Mant J., Morris D.C., Schmidt S., Cook L.G. 2015. Key innovation or adaptive change? A test of leaf traits using triodiinae in Australia. Sci. Rep. 5:12398.

Trovão N.S., Suchard M.A., Baele G., Gilbert M., Lemey P. 2015. Bayesian inference reveals host-specific contributions to the epidemic expansion of influenza A H5N1. Mol. Biol. Evol. 32:3264–3275.

Vaughan T.G., Kühnert D., Popinga A., Welch D., Drummond A.J. 2014. Efficient Bayesian inference under the structured coalescent. Bioinformatics 30:2272–2279.

Volz E.M., Frost S.D.W. 2014. Sampling through time and phylodynamic inference with coalescent and birth–death models. J. R. Soc. Interface 11:20140945.

Volz E.M., Koelle K., Bedford T. 2013. Viral phylodynamics. PLoS Comp. Biol. 9:e1002947.

Volz E.M., Pond S.L.K., Ward M.J., Brown A.J.L., Frost S.D.W. 2009. Phylodynamics of infectious disease epidemics. Genetics 183:1421–1430.

Vrancken B., Lemey P., Rambaut A., Bedford T., Longdon B., Gunthard H.F., Suchard M.A. 2014a. Simultaneously estimating evolutionary history and repeated traits phylogenetic signal: applications to viral and host phenotypic evolution. Methods Ecol. Evol. 6:67–82.

Vrancken B., Rambaut A., Suchard M.A., Drummond A.J., Baele G., Derdelinckx I., Wijngaerden E.V., Vandamme A.-M., Laethem K.V., Lemey P. 2014b. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. PLoS Comp. Biol. 10:e1003505.

Ward M.J., Gibbons C.L., McAdam P.R., van Bunnik B.A.D., Girvan E.K., Edwards G.F., Fitzgerald J.R., Woolhouse M.E.J. 2014. Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of staphylococcus aureus clonal complex 398. Appl. Environ. Microbiol. 80:7275–7282.

Ward M.J., Lycett S.J., Kalish M.L., Rambaut A., Leigh Brown A.J. 2013. Estimating the rate of intersubtype recombination in early HIV-1 group M strains. J. Virol. 87:1967–1973.

Worobey M., Han G.-Z., Rambaut A. 2014. A synchronized global sweep of the internal genes of modern avian influenza virus. Nature 508:254–257.

Wright S. 1934. An analysis of variability in the number of digits in an inbred strain of guinea pigs. Genetics 19:506–536.

Yang Z. 2006. Computational molecular evolution. Oxford, UK: Oxford University Press.

Yoder A.D., Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

Ypma R.J.F., Bataille A.M.A., Stegeman A., Koch G., Wallinga J., van Ballegooijen W.M. 2012. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc. R. Soc. B. 279:444–450.

Zinder D., Bedford T., Gupta S., Pascual M. 2013. The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. PLoS Path. 9:e10031054.