Epidemics 10 (2015) 88–92

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

# Eight challenges in phylodynamic inference

Simon D.W. Frost[a,b,*], Oliver G. Pybus[c], Julia R. Gog[d], Cecile Viboud[e], Sebastian Bonhoeffer[f], Trevor Bedford[g]

[a] Department of Veterinary Medicine, University of Cambridge, Cambridge, UK
[b] Institute of Public Health, University of Cambridge, Cambridge, UK
[c] Department of Zoology, University of Oxford, Oxford, UK
[d] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK
[e] Fogarty International Center, National Institutes of Health, Bethesda, USA
[f] Institute of. Integrative Biology, ETH Zurich, Zurich, Switzerland
[g] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, USA

ARTICLE INFO

*Article history:*
Received 21 February 2014
Received in revised form 30 August 2014
Accepted 2 September 2014
Available online 16 September 2014

*Keywords:*
Phylodynamics
Coalescent models
Selection
Recombination

ABSTRACT

The field of phylodynamics, which attempts to enhance our understanding of infectious disease dynamics using pathogen phylogenies, has made great strides in the past decade. Basic epidemiological and evolutionary models are now well characterized with inferential frameworks in place. However, significant challenges remain in extending phylodynamic inference to more complex systems. These challenges include accounting for evolutionary complexities such as changing mutation rates, selection, reassortment, and recombination, as well as epidemiological complexities such as stochastic population dynamics, host population structure, and different patterns at the within-host and between-host scales. An additional challenge exists in making efficient inferences from an ever increasing corpus of sequence data.



## Introduction

'Phylodynamics' is a term used to describe the 'melding of immunodynamics, epidemiology, and evolutionary biology' in order to understand how infectious diseases are transmitted and evolve (Grenfell et al., 2004). Since the term was coined ten years ago, many studies have taken up this concept, driven by the increasing availability of pathogen sequence data. Publicly available software such as BEAST (Drummond and Rambaut, 2007) has enabled individuals to apply complex evolutionary models to these data. New conceptual models (Volz et al., 2009, 2012; Frost and Volz, 2010; Rasmussen et al., 2011; Stadler et al., 2012; Dearlove and Wilson, 2013) have added to our understanding of how the process of disease transmission may shape a phylogeny, and of how the population genetics concept of 'effective population size' relates to pathogens. Here we present open challenges in using sequence data to infer disease dynamics.

## 1. How can we account for sequence sampling patterns?

While there are a vast amount of publicly-available sequence data – currently there are over one million viral sequences in GenBank – sampling is often highly biased. Sampling may be biased towards trying to capture a diverse taxonomic sample or may be biased by sampling a restricted geographic area, impacting estimates of effective population size (Holmes et al., 1999). Sampling effects may also be important when studying the 'phylogeography' of a pathogen. Some widely used phylogeographic models treat the migration of a pathogen as though it were analogous to mutation (Kühnert et al., 2011), mainly for computational expediency. However, spatial oversampling of specific areas may lead to these areas becoming apparent 'sinks', where overrepresentation of a deme causes estimates of migration into that deme to increase.

Previous work on the impact of temporal sampling has demonstrated that sampling protocols designed to capture sequences at specific points in the epidemic cycle result in more accurate inference using coalescent models (Stack et al., 2010). Current birth-death models used for phylodynamic inference assume a constant probability of sampling throughout the evolutionary history, which may result in biased estimates of quantities such as the effective population size when the sampling process is misspecified. Formal investigations of the potentially confounding effects

* Corresponding author at: Department of Veterinary Medicine, University of Cambridge, Cambridge, UK. Tel.: +44 01223764201; fax: +44 01223765760.
*E-mail address:* sdf22@cam.ac.uk (S.D.W. Frost).

http://dx.doi.org/10.1016/j.epidem.2014.09.001

of both spatial and temporal non-random sampling, and how they may be ameliorated, are well overdue. In order to make the best use of currently available data, methods would strongly benefit from recalibrating samples based on surveillance information; the development of realistic models of the sampling process could in theory increase statistical power while reducing bias.

## 2. How can more realistic evolutionary models be used to improve phylodynamic inferences?

For rapidly evolving pathogens, a range of sampling times can be used to calibrate a 'molecular clock' to estimate divergence times on a phylogenetic tree. Despite 'relaxed' clocks that can capture some degree of variation in evolutionary rates (Drummond et al., 2006), current models may fall short of capturing true variation. For example, Wertheim et al. (2012) analysed major subtypes of pandemic HIV-1 group M, which are thought to exemplify closely related lineages with different substitution rates, and found that the times to the most recent common ancestor differed markedly when subtypes were analysed separately compared to jointly. This suggests that current models fail to capture higher-order temporal correlations in the evolutionary rate. However, recent work on influenza by Worobey et al. (2014) found that incorporating outside information, in this case host species, substantially improved temporal calibration.

In addition, most studies do not consider how the epidemiological dynamics may feed back to the pattern of evolution. For pathogens such as influenza A, mutations that induce antigenic change and allow escape from the predominating herd immunity are likely to spread. This results in phylogenies that depart strongly from the neutral coalescent expectation, showing increased asymmetry in branching topology and skewed patterns of coalescence (Bedford et al., 2011). Analytically tractable coalescent models that directly incorporate such selection pressures do not currently exist. Recent progress has been made applying non-Kingman coalescent processes, such as the Bolthausen–Sznitman coalescent, to capture some of the broad effects of selection on phylogenetic shape and scale (Neher and Hallatschek, 2013). However, even with such coalescent models, there will remain the assumption that the observed phylogeny is independent of the substitution process. This is problematic as a major goal is to link viral mutations to evolutionary outcomes, and identify strains that may have a competitive advantage.

## 3. What is the role of stochastic effects in phylodynamics?

The vast majority of phylodynamic studies assume a time-varying coalescent model (Pybus and Rambaut, 2009; Volz et al., 2013) that specifies that changes at the population level are deterministic, which have demonstrated a variety of dynamic patterns for different viral systems (Table 1 in Frost and Volz (2010)). Demographic stochasticity may play a central role in infections that exhibit recurrent epidemics, such as influenza A virus and norovirus, due to seasonal troughs in incidence. However, even infections that are now endemic in many populations, such as HIV-1 and hepatitis C virus, were once at low frequency, and also sporadically appear in new populations. Hence, stochastic effects may play a role close to the time of the most recent common ancestor for many pathogens. Stochastic effects due to demography may also be important when the number of infected individuals is relatively small and/or infection and recovery rates are high.

Several recent studies have employed a stochastic linear birth–death process, where birth corresponds to transmission, and death to either recovery or death of infected individuals (Stadler et al., 2012). However, these models assume constant rates, and hence may be inappropriate if infection is not spreading exponentially. Extensions to the basic birth–death model, such as the birth–death skyline (Stadler et al., 2013), which involves fitting a piecewise constant birth–death process, may help to capture varying infection rates. However, fitting a stochastic, nonlinear model of disease transmission may be preferable to such non-parametric approaches, as it may offer mechanistic insights. One approach to incorporate stochasticity in such models is to apply the coalescent to an ensemble of stochastic simulations, an approach taken by Rasmussen et al. (2011), who fitted a stochastic differential equation model jointly to epidemiological data and to coalescence events inferred from a phylogenetic tree. Another possible approach to allow use of the coalescent likelihood would be to perform a stochastic change in timescale (Kaj and Krone, 2003). Kühnert et al. (2014) fitted a stochastic epidemiological model by simulating epidemiological trajectories, which were used to parameterise a stochastic birth-death process with piecewise constant rates. One can take inspiration from developments in fitting stochastic epidemic models to incomplete data. Leventhal et al. (2014) fitted a stochastic model using numerical approximation to the solution of the underlying master equation, in order to integrate out the (unknown) number of transmission events in the population that occur between coalescent intervals in the sample.

Environmental stochasticity, which is important in the dynamics of many vector-borne diseases, has received relatively little attention in the phylodynamics literature to date. Recent developments in fitting models that can accommodate both stochasticity (via stochastic differential equations, a common framework for including environmental stochasticity) (Rasmussen et al., 2011) and structure (Rasmussen et al., 2014) are encouraging, and await wider availability of sequence data on vector-borne pathogens.

## 4. How does the structure of the host population relate to pathogen genetic variation?

There has been recent progress in understanding how classical compartmental epidemiological models, which can incorporate population structure by considering multiple classes of individual, relate to the resulting pathogen phylogeny (Stadler et al., 2012; Volz et al., 2012; Volz, 2012; Frost and Volz, 2013). These models may be more robust to biased sampling of specific groups than models that consider population structure as a trait that evolves independently of the underlying phylogeny (Lemey et al., 2009). Including population structure may be essential for accurate inference, as recently demonstrated by Rasmussen et al. (2014a), who showed that a panmictic epidemiological model failed to capture the classical oscillatory dynamics of dengue virus, while splitting the hosts into separate but linked urban and rural populations was sufficient to recapitulate observed oscillations in hospital admissions with dengue. Phylogeographic models have also been applied to consider host species jumps in multiple host systems such as rabies Streicker et al. (2010), although there are issues with sparse sampling of such systems, and the potential for stochastic effects (Buhnerkempe et al., 2014; Lloyd-Smith et al., 2014).

However, compartmental models may not fully capture heterogeneity in contacts and transmissions among individuals. A high variance among hosts in onward transmission – sometimes termed super-spreading – is characteristic of many infectious diseases (Lloyd-Smith et al., 2005), and this may impact the phylogeny (Leventhal et al., 2012). There are many challenges in developing network models that aim to capture deviations from the 'well-mixed pot' assumption of many compartmental epidemiological models (Pellis et al., 2014), and more challenging still to include such structure into phylodynamic models. This relates to the wider challenge of incorporating individual-level variation, rather than aggregating individuals into groups.