PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2010) **365**, 1879–1890
doi:10.1098/rstb.2010.0060

Viral phylodynamics and the search for an 'effective number of infections'

Simon D. W. Frost[1,*] and Erik M. Volz[2]

[1]*Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, Cambridgeshire CB3 0ES, UK*
[2]*Department of Epidemiology, University of Michigan – Ann Arbor, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA*

Information on the dynamics of the effective population size over time can be obtained from the analysis of phylogenies, through the application of time-varying coalescent models. This approach has been used to study the dynamics of many different viruses, and has demonstrated a wide variety of patterns, which have been interpreted in the context of changes over time in the 'effective number of infections', a quantity proportional to the number of infected individuals. However, for infectious diseases, the rate of coalescence is driven primarily by new transmissions i.e. the incidence, and only indirectly by the number of infected individuals through sampling effects. Using commonly used epidemiological models, we show that the coalescence rate may indeed reflect the number of infected individuals during the initial phase of exponential growth when time is scaled by infectivity, but in general, a single change in time scale cannot be used to estimate the number of infected individuals. This has important implications when integrating phylogenetic data in the context of other epidemiological data.

Keywords: phylodynamics; effective population size; viral evolution; coalescent; epidemiological models

1. INTRODUCTION

Viruses, especially RNA viruses such as human immunodeficiency virus type 1 (HIV-1), hepatitis C virus (HCV) and influenza A virus, may exhibit a great deal of genetic variation at the population level, allowing the reconstruction of viral phylogenies that reflect the past transmission of the virus. The shape of the phylogeny can tell us a great deal about population processes, such as changes in population size and geographic population structure. It can also indicate the effects of immunological processes, such as selection of escape variants (Pybus & Rambaut 2009). For example, 'star-like' phylogenies are typical of populations that are growing exponentially, while 'ladder-like' phylogenies are consistent with a model where one variant is replaced by another due to immune escape. This integration of ecological, epidemiological and evolutionary processes has been dubbed 'phylodynamics' (Grenfell *et al.* 2004).

Sophisticated statistical methods have been developed which allow time-stamped phylogenies to be obtained from viral sequence data (Rambaut 2000; Drummond *et al.* 2006), and these have been used in conjunction with coalescent models borrowed from population genetics (Kingman 2000; Pybus *et al.* 2000; Drummond *et al.* 2005) to determine different patterns of changes in population size over time (figure 1). These methods have been used to study the phylodynamics of many different viruses, mostly RNA viruses, but also to a lesser extent, DNA viruses (table 1). While not an exhaustive review of viral phylodynamic studies, table 1 reveals a wide range of phylodynamic patterns, ranging in complexity from a constant population size to multiple phases of growth, including oscillations. Most of these studies have used a model of the coalescent in a time-varying population, which considers the genealogical process of a small sample of taxa taken from a large population that changes in time deterministically. The population size is assumed to be homogeneous and under neutral evolution. Although in practice these assumptions are broken, it is often the case that an 'effective population size', N_e, can be derived, which gives the same coalescence rate as an idealized population of size N. To date, phylodynamic studies of viral evolution have assumed that N_e is equivalent to the (effective) number of infected individuals. Although some studies argue that the effective population size may be lower than expected due to variability between individuals in infectiousness, all assume that an 'effective number of infections' that is proportional to the number of infected individuals.

Using simple epidemiological models, we have recently demonstrated that the coalescence rate of an infectious disease is related to the rate of transmission (i.e. the incidence) and not directly to the absolute number of infected individuals (i.e. the prevalence; Volz *et al.* 2009). Prevalence does affect the shape of the phylogeny, but only indirectly through sampling

* Author for correspondence (sdf22@cam.ac.uk).

One contribution of 14 to a Theme Issue 'New experimental and theoretical approaches towards the understanding of the emergence of viral infections'.

This journal is © 2010 The Royal Society

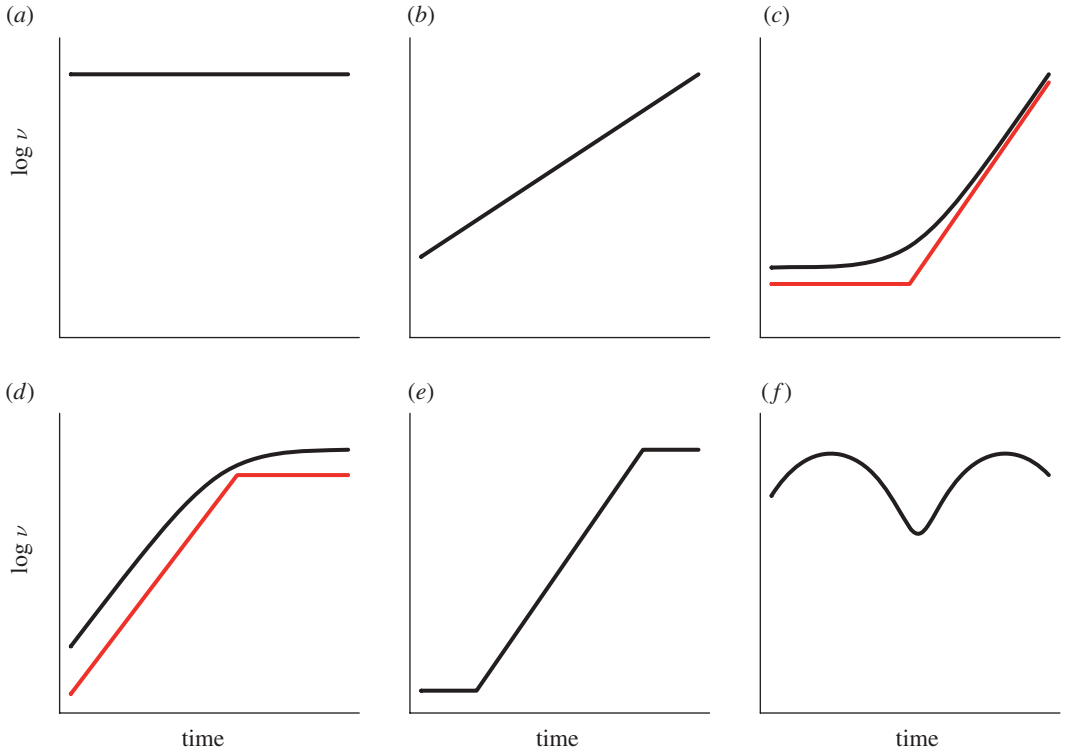

Figure 1. Schematic of different phylodynamic patterns for the relative size function, ν over time. (*a*) constant; (*b*) exponential; (*c*) piecewise expansion; (*d*) piecewise logistic; (*e*) constant–expansion–constant; (*f*) oscillatory.

effects; when a higher proportion of infected individuals are sampled, more coalescent events are evident near the tips of the phylogeny. In this study, we examine whether there are conditions under which the coalescence rate may indeed reflect the 'effective number of infections', by comparing coalescence in epidemiological models with classical population genetics models. We also address how the conclusions of previous studies may be affected by interpreting phylodynamic patterns as being driven by incidence rather than prevalence.

2. PHYLODYNAMIC PATTERNS UNDER DIFFERENT EPIDEMIOLOGICAL SCENARIOS

(a) *The time-varying coalescent model*

The model used most commonly for viral phylodynamics is the time-varying coalescent model (Griffiths & Tavaré 1994), which considers the genealogical process in a population that changes size in a deterministic fashion according to some relative size function, $\nu(\tau)$, where τ is the time measured in generations, starting with the present and going backwards. For example, for a constant population size, we have $\nu(\tau) = 1$. A variety of different parametric models have been proposed for $\nu(\tau)$, including a constant population size, exponential growth, logistic growth and expansion growth. These can be strung together in series to make more complicated patterns. In addition, a number of 'nonparametric' models have been proposed for $\nu(\tau)$ (Pybus *et al.* 2000; Strimmer & Pybus 2001; Drummond *et al.* 2005; Opgen-Rhein *et al.* 2005; Minin *et al.* 2008), which when fitted to data have sometimes demonstrated complex patterns, including oscillatory dynamics (table 1).

So, what does the relative size actually mean, and how does it relate to the coalescence rate? Let us consider a sample of n individuals taken at time $\tau = 0$, and assume that the sample can be traced back to a single common ancestor with probability 1 (i.e. $\sum_0^\infty \nu^{-1}(\tau)\mathrm{d}\tau = \infty$). The dynamics of the number of distinct ancestors of the sample at time τ is modelled as a stochastic process $\{\mathrm{A}_n(\tau), \tau \geq 0\}$, which starts at $A_n(0) = n$, and moves down in steps of 1 until reaching 1, at which point the sample has been traced back to a most recent common ancestor. In a small time step h, the transition probabilities are determined by the following:

$$\mathcal{P}(A_n(\tau+h) = j | A_n(\tau) = i) = \begin{cases} \binom{i}{2}\frac{1}{\nu(\tau)}h + o(h) & j = i-1, \\ 1 - \binom{i}{2}\frac{1}{\nu(\tau)}h + o(h) & j = i, \\ 0 & \text{otherwise.} \end{cases}$$

Equation (2.1) shows that the rate of coalescence increases with the number of distinct ancestors, and decreases with a greater relative size. Under a Wright–Fisher model of a haploid population, the relative size function is simply the population size, i.e. $\nu(\tau) = N(\tau)$. This is only an approximate result for the Wright–Fisher model, however, which holds when the sample size is small relative to the population size, as equation (2.1) assumes that only one

Table 1. Phylodynamic patterns of viruses.

pattern	virus
constant	Canine distemper virus (Pomeroy *et al.* 2008); Hepatitis B virus (van Houdt *et al.* 2010); Hepatitis C virus (Golemba *et al.* 2010); HIV-1 (Deng *et al.* 2008; Tee *et al.* 2008); Measles virus (Pomeroy *et al.* 2008); Mumps virus (Pomeroy *et al.* 2008); Rabbit haemorrhagic disease virus (Kinnear & Linde 2010); Rabies virus (Hughes *et al.* 2004; Davis *et al.* 2007); Ross River virus (Jones *et al.* 2010); Simian foamy virus (Liu *et al.* 2008); St Louis encephalitis virus (Twiddy *et al.* 2003).
expansion	Hepatitis B virus (Zehender *et al.* 2008); Hepatitis C virus (Jiménez-Hernández *et al.* 2007); HIV-1 (Worobey *et al.* 2008); Influenza A (Goñi *et al.* 2009).
exponential	Dengue virus (Twiddy *et al.* 2003); Hepatitis C virus (Jiménez-Hernández *et al.* 2007; Pybus *et al.* 2003, 2005); HIV-1 (Lemey *et al.* 2004; Salemi *et al.* 2005; Walker *et al.* 2005; Salemi *et al.* 2008); Influenza A (Chen & Holmes 2006; Fraser *et al.* 2009; Rambaut & Holmes 2009); Human rhinovirus (Briese *et al.* 2008); Measles virus (Pomeroy *et al.* 2008); Rabies virus (Hughes *et al.* 2004; Davis *et al.* 2007); West Nile virus (Snapinn *et al.* 2007).
logistic	Canine parvovirus (Pereira *et al.* 2007); Dengue virus (Carrington *et al.* 2005); Hepatitis B virus (Zehender *et al.* 2008); Hepatitis C virus (Verbeeck *et al.* 2006; Jiménez-Hernández *et al.* 2007). HIV-1 (Robbins *et al.* 2003; Hu *et al.* 2005; Walker *et al.* 2005; Bello *et al.* 2007; Tee *et al.* 2008); Human erythrovirus B19 (de Freitas *et al.* 2008); West Nile virus (Snapinn *et al.* 2007).
piecewise logistic	Hepatitis C virus (Tanaka *et al.* 2004; Pybus *et al.* 2003);
piecewise expansion	Hepatitis B virus (Michitaka *et al.* 2006); Hepatitis C virus (Pybus *et al.* 2005; Nakano *et al.* 2004; Kurbanov *et al.* 2007); Hepatitis delta virus (Kurbanov *et al.* 2007); Hepatitis E virus (Tanaka *et al.* 2006); HIV-1 (Kurbanov *et al.* 2003); HIV-2 (Lemey *et al.* 2003*a*); Infectious bursal disease virus (Hon *et al.* 2006); Japanese encephalitis virus (Twiddy *et al.* 2003); Rabies virus (Hughes *et al.* 2004)
two-phase exponential	Dengue virus (Twiddy *et al.* 2003).
nonparametric	
constant	Epizootic haemorrhagic disease virus (Biek 2007); St Louis encephalitis virus (Baillie *et al.* 2008).
constant/exponential phases	Avian metapneumovirus (Padhi & Poss 2009); Dengue virus (Schreiber *et al.* 2009); Feline immunodeficiency virus (Biek *et al.* 2006); Hepatitis C virus (Nakano *et al.* 2006; Njouom *et al.* 2007; Njouom *et al.* 2009; Pybus *et al.* 2009); HIV-1 (Bello *et al.* 2009; Pérez-Losada *et al.* 2010); JC virus (Kitchen *et al.* 2008); Rabies virus (Biek *et al.* 2007).
decline	Buggy Creek virus (Padhi *et al.* 2008); Hepatitis A virus (Moratorio *et al.* 2007); Hepatitis B virus (van Ballegooijen *et al.* 2009); Toscana virus (Zehender *et al.* 2009)
oscillatory	Dengue virus (Bennett *et al.* 2009); Influenza A (Rambaut *et al.* 2008); Influenza B (Chen & Holmes 2008) West Nile virus (Amore *et al.* in press).

coalescence can occur at a time. When a large proportion of the population is sampled, multiple coalescent events may occur in a single generation. In such a case, more general coalescent models that the commonly used Kingman coalescent may be more appropriate, which allow multiple 'collisions' of lineages (Pitman 1999; Sagitov 1999, 2003; Schweinsberg 2000; Mohle & Sagitov 2001). Although populations may deviate from the assumptions of a Wright–Fisher model—for example, they may show geographical structure—in many, but not all cases, the relative size function can be assumed to be proportional to the population size, in which case, it is referred to as the 'effective population size', N_e, and the relative size function is $\nu(\tau) = N_e(\tau)$.

If g_i is the length of time during which the ancestral process is in state $A_n = i$ and τ_i is the time that the interval starts, then under model (2.1), g_i is distributed as follows (Pybus *et al.* 2000).

$$\mathcal{P}(g_i|\tau_i) = \frac{\binom{i}{2}}{\nu(\tau_i + g_i)} \exp\left[-\int_{\tau=\tau_i}^{\tau_i+g_i} \frac{\binom{i}{2}}{\nu(\tau)} d\tau\right]. \qquad (2.1)$$

Since $\mathcal{P}(g_i|\tau_i)$ depends only on the relative size function, equation (2.1) allows coalescent intervals to be simulated for a given relative population density $\nu(\tau)$, and also allows the model to be fitted to coalescent intervals estimated from phylogenetic trees. Although branch lengths in a phylogeny are typically in units of expected substitutions per site, in many viral phylodynamic studies, a strict or 'relaxed' molecular clock is often used in conjunction with serial samples of

sequences (Rambaut 2000; Seo *et al.* 2002*a*,*b*; Lemey *et al.* 2003*b*; Sanderson 2003; Drummond *et al.* 2006; Yang *et al.* 2007), such that branch lengths are scaled in absolute time. Many studies do not assume a specific generation time, and in doing so, generate estimates of the product of the generation time and $\nu(\tau)$ as the 'effective population size'. To avoid making assumptions regarding how time is rescaled, some studies simply refer to estimates of $\nu(\tau)$ obtained from the data as 'genetic diversity' (Carrington *et al.* 2005; Rambaut *et al.* 2008; Schreiber *et al.* 2009; van Ballegooijen *et al.* 2009).

(b) *Deterministic models for the coalescent*

A common framework for modelling infectious diseases is compartmental models, in which the population is divided up into subpopulations called *compartments*, such as susceptible and infected individuals. The rate of change in the size of these compartments as we go forward in time, t, is modelled using differential equations. We can also consider a differential equation for the dynamics of the number of lineages over time based on equation (2.1).

$$\frac{\mathrm{d}A(\tau)}{\mathrm{d}\tau} = -\binom{A(\tau)}{2}\frac{1}{\nu(\tau)}. \tag{2.2}$$

There are two different ways of interpreting this equation. Firstly, we could consider A as an approximation to the number of lineages when the sample size is very large i.e. $A = \lim_{n\to\infty} A_n$, in which case we could approximate equation (2.2) by $-A(\tau)^2/\ 2\nu(\tau)$, as $n(n-1)$ tends to n^2 as n gets large. This approximation is surprisingly good, even when the number of distinct lineages is small (e.g. only an 11% difference when $n = 10$). Another way to look at A is as an approximation to the mean number of lineages over time i.e. $A \approx E(A_n)$; we adopt the latter interpretation. Recently, we showed that for many simple epidemiological models, the rate of coalescence in a phylogeny is a function of the number of infected individuals, Y and the rate at which susceptible individuals, X, become infected, f_{XY}. If we denote time going backwards from the present as s, the dynamics of the number of ancestral lineages over time can be modelled using the following differential equation (Volz *et al.* 2009).

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\binom{A(s)}{2}\frac{f_{XY}}{\binom{Y(s)}{2}} \approx -f_{XY}\frac{A(s)^2}{Y(s)^2}. \tag{2.3}$$

The rationale underlying equation (2.3) is that coalescence occurs at a rate equal to the transmission rate, f_{XY}; coalescence can occur between any pair of infected individuals, but will only result in a decrease in the number of lineages in the sample if both the source of infection and the recipient of infection are sampled, either directly (through these individuals being included in the sample) or indirectly (through sampling their descendant viral lineages). In our previous work, we modelled the number of lineages using expression (2.3); in order to assist comparisons between the coalescent and epidemiological models, we assume that the population size is large, such that $\binom{Y}{2} \approx Y^2/2$, but not the number of samples. Hence, we model the number of lineages over time as follows.

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\binom{A(s)}{2}\frac{2f_{XY}}{Y(s)^2}. \tag{2.4}$$

Note that the time scale in equation (2.4) is in real time, and the coalescence rate is determined by a combination of the number of new infections per unit time (the absolute incidence) and the level of sampling (which, for a fixed sample size, is dependent on the absolute prevalence of infection). The term $2f_{XY}/Y(s)^2$ on the right-hand side of equation (2.4) is simply the probability that a pair of ancestral lineages are descended from a common ancestor, and this probability is the same as that under a Moran model, because one of the lineages we are following must be the 'offspring' and the other must be the 'parent', and there are two ways for this to occur. This is in contrast to the haploid Wright–Fisher model, in which the probability of a pair of ancestral lineages being descended from a common ancestor is the inverse of the population size. Despite being based on differential equations, extensive simulation results show that this model is surprisingly good at recapitulating the dynamics of the number of lineages over time (at least on average) for a range of population sizes and sample sizes (Volz *et al.* 2009 and this study), although it should be noted that the variance in the number of lineages can be large. Although this may be an issue when trying to estimate parameters from data (for example, using equation (2.1)), equation (2.4) is extremely useful to help understand the connection between the epidemiological and evolutionary dynamics. To illustrate this, we considered the dynamics of the number of lineages over time for a variety of epidemiological scenarios using two simple, but commonly used, epidemiological models.

(c) *A model with a constant number of infected individuals*

A useful 'null model' to study the change in effective population size over time is a model with a constant population size. In an epidemiological model, this corresponds to an endemic equilibrium. As an example of a model with an endemic equilibrium, we consider a simple model commonly used to study the spread of HIV among men who have sex with men (for a comparison of the deterministic and stochastic version of this model, see Jacquez & Simon 1993). If X denotes the number of susceptible individuals and Y denotes the number of infected individuals, the rates of change of X and Y are as follows:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \Lambda - \beta c X(t)\frac{Y(t)}{N(t)} - \mu X(t) \tag{2.5}$$

and

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = \beta c X(t)\frac{Y(t)}{N(t)} - (\mu+\gamma)Y(t), \tag{2.6}$$

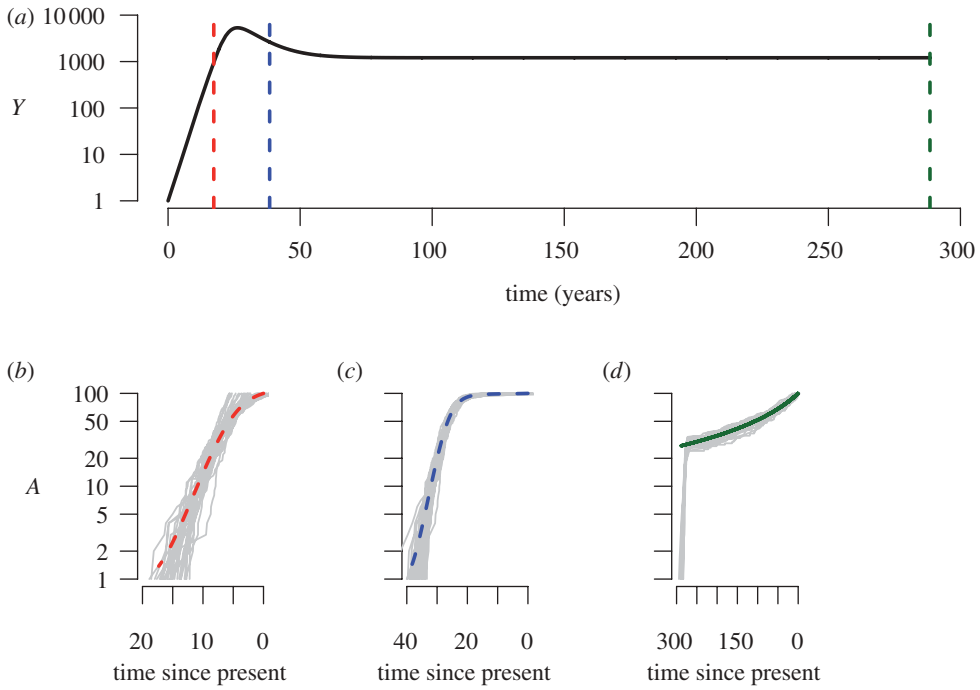

Figure 2. Phylodynamics of a simple susceptible-infected model in an open population (equations (2.7) and (2.8) in the main text). (*a*) Dynamics of the number of infected individuals, I over time in years. The vertical lines denote sampling times, and the number of lineages over time (*b*) during exponential growth (red), (*c*) following the peak of infected individuals (blue) and (*d*) at equilibrium (green). The grey lines represent stochastic simulations; in order to generate a fair comparison between the deterministic model and the stochastic simulations, time was shifted for each simulation such that the peak prevalence occurred at the same time as in the deterministic model. Parameter values are as follows (with time in years); $\beta c = 52$, $\gamma = 1/10$, $\mu = 1/70$, $\Lambda = 10000/70$. The initial conditions were: $X(0) = 9999$, $Y(0) = 1$. Sampling times were set at 900/52, 2000/52 and 15000/52 years, and a sample size of 100 was assumed, i.e. $A = 100$. Numerical simulations were performed in R (R Development Core Team 2009) using the simecol library (Petzoldt & Rinke 2007). Stochastic simulations were performed with SimPy (http://simpy.sourceforge.net). All code is available from S.D.W.F. on request.

where

$$N(t) = X(t) + Y(t).$$

Here, β denotes the probability of infection per contact, c, the contact rate, μ, the natural mortality rate, γ, the excess mortality caused by infection, and Λ, the rate of immigration/birth of new susceptibles. The dynamical behaviour of the model depends on the value basic reproductive number $R_0 = \beta c/\mu + \gamma$. If $R_0 > 1$, the number of infected individuals initially increases exponentially, plateaus, and finally reaches an equilibrium (figure 2).

By substituting $f_{XY} = \beta cXY/N$ into equation (2.4), we obtain the following expression.

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\binom{A(s)}{2}\frac{2\beta cX(s)}{Y(s)N(s)}. \tag{2.7}$$

If we denote the equilibrium population sizes of the number of susceptibles, infecteds and the total population size as X^*, Y^* and $N^* = X^* + Y^*$, respectively, the rate of change of lineages going backwards in time, $\mathrm{d}A/\mathrm{d}s$, is as follows.

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\kappa A(s)(A(s) - 1). \tag{2.8}$$

The solution of which is

$$A(s) = \left(1 - \frac{A(0) - 1}{A(0)}\mathrm{e}^{-\kappa s}\right)^{-1}, \tag{2.9}$$

where

$$\kappa = \frac{\beta cX^*}{(X^* + Y^*)Y^*} = \frac{(\beta c - \gamma)(\mu + \gamma)^2}{\Lambda(\beta c - (\mu + \gamma))}. \tag{2.10}$$

Equation (2.8) shows that the coalescence rate is not proportional to the number of infected individuals, but is also a function of the number of susceptible individuals. Consequently, even for this relatively simple model, the expression (2.10) for the rate parameter κ is a nonlinear combination of several parameters, and shows that in the absence of other information about the epidemiological processes, the dynamics of lineages through time may provide very little information about individual parameters. Note that by starting the system at equilibrium, the number of infected individuals going backwards in time is constant i.e. all information on when the susceptible population was invaded with an infected individual is lost.

We compared the number of lineages over time using equation (2.9) with stochastic simulations (figure 2). The analytical solution gives a good approximation to the mean number of lineages over time for the period during which the system is close to equilibrium.

(d) *An exponentially growing population*

During the early phase of epidemic growth, when $X(t)/N(t) \approx 1$, the number of infected individuals increases

exponentially over time.

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = (\beta c - (\mu + \gamma))Y(t)$$

and

$$Y(t) = Y(0)\mathrm{e}^{(\beta c - (\mu+\gamma))t}.$$

Going backwards in time, the expressions for $Y(s)$ and $A(s)$ are as follows:

$$Y(s) = Y(0)\mathrm{e}^{-rs},$$

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\frac{\beta c \mathrm{e}^{rs}}{Y(0)}A(s)^2$$

and

$$A(s) = \left(1 - \frac{A(0)-1}{A0}\mathrm{e}^{-\beta c/rY(0)(\mathrm{e}^{rs}-1)}\right)^{-1},$$

where

$$r = \beta c - (\mu + \gamma).$$

It is also informative to examine the expression for $\mathrm{d}A(s)/\mathrm{d}s$ as a function of $Y(s)$ in the case of exponential growth.

$$\frac{\mathrm{d}A(s)}{\mathrm{d}s} = -\binom{A(s)}{2}\frac{2\beta c}{Y(s)}. \tag{2.11}$$

During exponential growth, there is a linear relationship between the prevalence and the incidence, and hence the coalescence rate is directly proportional to the number of infected individuals.

(e) *Logistic growth*

The model given by equations (2.5) and (2.6) also exhibits similar dynamics to logistic growth. Although closed expressions for $X(t)$ and $Y(t)$ cannot be obtained for this model, we can obtain the number of lineages through time by numerically solving for A backwards in time, either by simulating the complete system of differential equations backwards (as in Volz *et al.* 2009), or by simulating the epidemic forwards in time, and storing f_{XY} and Y, which can then be used as inputs into a single differential equation for A. Figure 2 demonstrates that the number of lineages over time, for a sample taken just after peak prevalence, is well described on average by the differential equation model (2.5) and (2.6). During exponential growth, incidence is high and lineages increase rapidly, while after the peak, incidence is low, and the rate of increase of lineages drops.

(f) *Relationship between coalescence rate and estimates of effective population size*

Many previous studies have estimated the 'effective population size', N_{e} of an epidemic without considering an explicit model of disease transmission. To investigate the relationship between estimates of effective population size obtained using standard coalescent models, transmission rates, and number of infected individuals, we fitted generalized skyline plots to stochastic simulations of the model based on equations (2.5) and (2.6). When branch lengths in the phylogeny are measured in continuous time, as is common for viral phylodynamic studies, then assuming model (2.1), the use of this approach will generate estimates of the product of the generation time and N_{e}. From a comparison of equations (2.2) and (2.11), it might initially appear that the application of standard coalescent models would give estimates of $2\beta cY$. However, as shown in figure 3, during exponential growth, the skyline is a good estimate of βcY. This arises as epidemiological models that operate in continuous time bear a closer resemblance to the Moran population model, where generations overlap in continuous time and only one coalescent event can occur at a time. The 'coalescent effective population size', defined as the average time to a coalescent event measured in units of the average time back to a birth event is $N_{\mathrm{e}} = N$ for a Wright–Fisher model, and $N_{\mathrm{e}} = N/2$ for a Moran model (Wakeley & Sargsyan 2009). Consequently, we have to halve the estimates of effective population size obtained assuming a Wright–Fisher model. In addition, the appropriate scaling in time is determined by the infectivity, which determines the average time back to a transmission event (analogous to a birth event), and not by the duration of infectiousness. Figure 3 demonstrates that standard models perform well in terms of both the absolute number of infected individuals, and the rate of change over time, suggesting that previous studies may have obtained good estimates of epidemic doubling time, despite making the erroneous assumption that coalescence is directly related to prevalence. However, as the relationship between the transmission rate, f_{XY} and the number of infected individuals Y is different during exponential growth and at equilibrium, we cannot find a single transformation of time such that the coalescence rate corresponds to the number of infected individuals over the entire epidemic. In this model, as the time between infections changes, the use of a single transformation of time to fit the early stages of the epidemic results in an overestimation of the true number of infected individuals in the later stages.

(g) *Oscillatory dynamics*

By application of the non-parametric 'skyline' type approaches used in the previous section, a number of studies have demonstrated oscillations in the relative size, ν, over time. Oscillations in the number of infected individuals in an epidemiological model can arise, for example, from seasonally changing contact rates. A simple example of this, appropriate to study the dynamics of an acute infection under seasonality, which considers susceptible, X, infected, Y, and immune individuals, Z, is as follows.

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \mu N - \beta_0(1 + \beta_1 \sin(\omega t))X(t)\frac{Y(t)}{N(t)} - \mu X(t), \tag{2.12}$$

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = \beta_0(1 + \beta_1 \sin(\omega t))X(t)\frac{Y(t)}{N(t)} - (\mu + \gamma)Y(t) \tag{2.13}$$

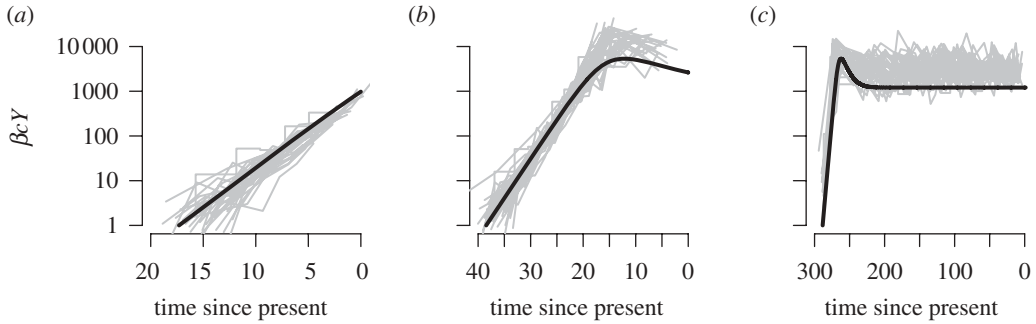

Figure 3. The product of the transmission probability, contact rate and number of infected individuals, βcY, at different stages of the epidemic depicted in figure 2 (smooth black line) obtained from numerically solving equations (2.7) and (2.8), along with numerical estimates of 'effective population size' estimated using generalized skyline plots fitted to stochastic simulations (grey lines) on the same scale. During the exponential growth period, the skyline generates good estimates of βcY. Parameter values and initial conditions are as described in figure 2. Skyline plots were generated using the APE library (Paradis *et al.* 2004) in R (R Development Core Team 2009). (*a*) Exponential growth; (*b*) after peak; (*c*) at equilibrium.

and

$$\frac{\mathrm{d}Z(t)}{\mathrm{d}t} = \gamma Y(t) - \mu Z(t), \tag{2.14}$$

where

$$N(t) = X(t) + Y(t) + Z(t).$$

We chose parameter values that gave annual fluctuations in the number of infected individuals, and numerically simulated the epidemic over a ten year period. We then simulated the dynamics of the number of lineages, sampling at the last peak of infection. Figure 4 shows the prevalence of infection, $Y(t)$ over time. This looks very different from the transmission rate, $f_{XY} = \beta_0\ (1 + \beta_1\ \sin(\omega t))X(t)Y(t)/N(t)$, which determines the rate at which lineages coalesce. If we were to mistakenly interpret the coalescence rate as proportional to the number of infected individuals, we would conclude that the prevalence was at a peak when it was in a trough, and vice versa, as for these parameter values, $Y(t)$ and f_{XY} are out of phase. For more complex oscillatory dynamics, such as biennial cycles, the relative magnitudes of $Y(t)$ and f_{XY} may also differ. These model results also reinforce previous assertions (Rambaut *et al.* 2008; Stack *et al.* in press), that a sample taken at a single point in time may provide relatively little information about the past population dynamics, as the population bottlenecks result in all sequences sampled at a single timepoint having a relatively recent common ancestor.

3. DISCUSSION

Using simple differential equation based models to gain insights into the phylodynamics of viral infections, we have demonstrated that the pattern of coalescence for an infectious disease is dominated by the transmission rate, while the number of infected individuals is of secondary importance. Although Holmes *et al.* (1995) recognized that coalescence in an infectious disease was related to transmission, this was not taken into account in later phylodynamic studies, which referred to the 'effective number of infections', i.e. the prevalence. Some studies also noted that the generation time is effectively the time between infections (Pomeroy *et al.* 2008; van Ballegooijen *et al.* 2009), and not the duration of infectiousness, but did not recognize that this changes throughout an epidemic. Hence, a single transformation of time, which is commonly used to estimate N_e from temporally sampled sequence data, cannot be used to recover the 'effective number of infected individuals'. In some cases, such as during exponential growth, there is a linear relationship between the transmission rate and the number of infected individuals, and with an appropriate choice of time scale (dividing time by βc in the models here) it is possible to estimate the number of infected individuals, but this is not true in general. Some studies (e.g. Rambaut *et al.* 2008) have been vague in the interpretation of the coalescence rate, relating it to 'genetic diversity'. We believe that this is a little too cautious—the rate of coalescence can be related to epidemiological parameters, but we have to explicitly consider the underlying transmission dynamics for this to be done correctly. For example, in the case of endogenous retroviruses (Romano *et al.* 2008), the transmission tracks the reproduction of the host, and standard coalescent models used for human populations can be used. In the case of viruses where there is significant vertical and horizontal transmission, more sophisticated models that incorporate coalescence in both the host and the virus will be required to interpret the phylodynamics patterns in the context of transmission parameters. A particularly pertinent quote comes from a review by Donnelly & Tavaré (1995) in their discussion of the time-varying coalescent (equation (2.1)):

> [T]he results described above do not apply in general. It is true for very general neutral models that unless there are discontinuities, i.e. sudden changes, in the processes governing the population size, the ancestral process can be represented as a time change of the process described in (equation (2.1)). However, the form of the time change, which is in general different from (equation (2.1)), depends on properties of the random process governing the rate at which individuals are born in the population, about which little is known in many practical contexts. It thus appears

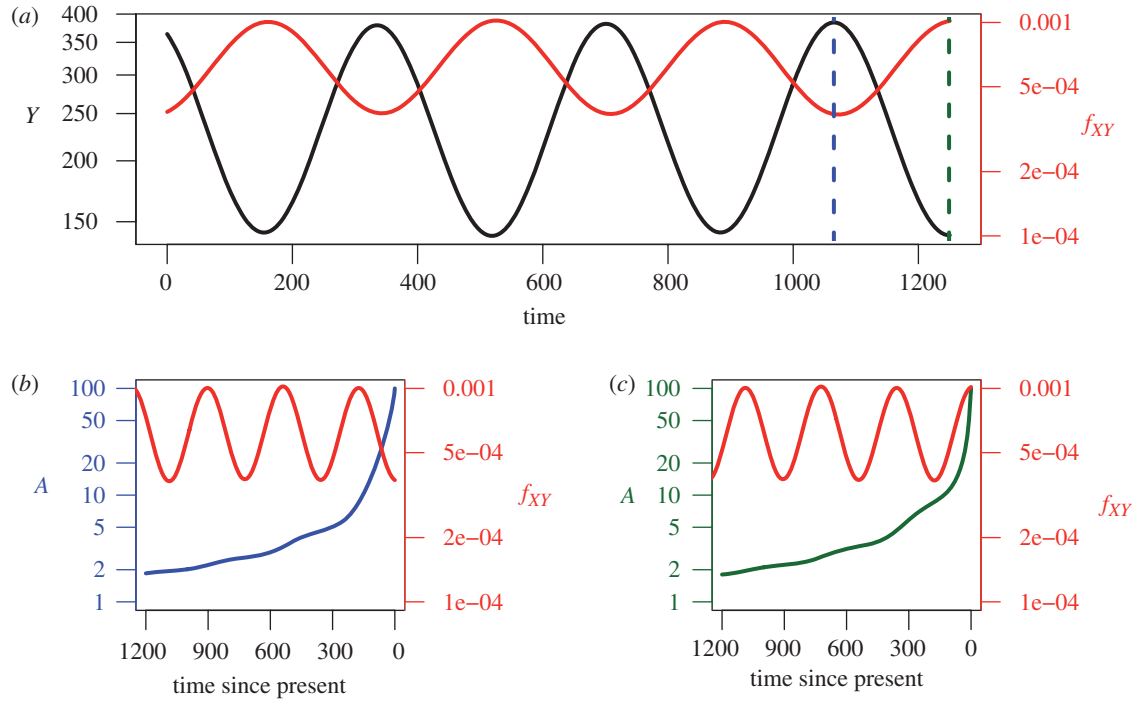

Figure 4. (*a*) Dynamics of the number of infected individuals, Y, and the transmission rate f_{XY} for a susceptible-infected-recovered model with seasonal forcing, given by equations (2.20)–(2.22) in the main text. Parameter values are as follows (with time in days): $\beta_0 = 10/7$, $\beta_1 = 0.05$, $\omega = 2\pi/365$, $\gamma = 1/7$, $\mu = 1/25550$. The population size, N, was assumed to be 10^6. Initial conditions: $S = 100029.946$, $I = 142.978$, $R = 899827.076$. (*b*) The time of sampling for the high prevalence scenario was $t = 3465$, when $I = 384.477$ (38.4 per 100 000) (*c*) and the time of sampling for the low prevalence scenario was $t = 3649$, when $I = 140.7068$ (14.0 per 100000).

> that some caution is appropriate in applying the above results on the coalescent in populations of variable size.
>
> (Donnelly & Tavaré 1995, p. 408)

That coalescence is related to transmission has important implications when interpreting phylodynamic patterns in the context of other data, such as information on the timing of external events or on disease prevalence. For example, in a recent study of dengue (DENV-4) in Puerto Rico (Bennett *et al.* 2009), although both N_e and case counts fluctuated over time, changes in N_e preceded changes in case counts by about seven months. This puzzling result is easily explained when one recognizes that the coalescence rate is a measure of incidence; as shown in our simple model of an oscillating epidemic, we expect incidence and prevalence to be out of phase, and in general, peaks of incidence precede peaks of prevalence. There was also no simple relationship between the amplitude of the fluctuations in N_e compared with the amplitude in case counts; in order to derive a meaningful comparison between these data, we would have to compare fluctuations in estimated incidence with N_e. Multiple studies have interpreted the timing of changes in phylodynamic patterns in the context of changes in other factors. For example, a decline in a skyline plot obtained from hepatitis A sequences sampled in France coincided with the introduction of vaccination (Moratorio *et al.* 2007), while a massive expansion in the 'effective number of infections' of hepatitis C virus in Egypt fell within a time period when the general population was treated with parenteral antischistosomal treatment (Pybus *et al.* 2003). Such external forces have a more immediate impact on transmission than prevalence.

The phylodynamic patterns can also be affected by sampling; sampling a higher fraction of the infected individuals at a time results in more recent coalescent times, and shorter terminal (external) branches of the tree, and a different tree shape (Mooers 1995; Rannala *et al.* 1998; Pybus *et al.* 2000, 2002; Purvis & Agapow 2002; Huelsenbeck & Lander 2003; Volz *et al.* 2009). As many viral phylodynamic studies employ serial samples of viral sequences, it is important to correct for possible differences in sampling depth, which will be a function ofthe temporal pattern of the sampling and the number of infected individuals. In a heterogeneous epidemic, the extent to which specific subpopulations are over- or under-sampled also has to be taken into account. The model framework we present here can be extremely informative to help understand the potential effects of sampling on phylodynamic patterns, and offers a more computationally faster approach to studying sampling effects than approaches based on full epidemic simulations coupled with computationally intensive Bayesian approaches for estimating N_e (Stack *et al.* in press).

Deterministic models of the phylodynamics of infectious disease can be very informative due to their relative simplicity. However, in some cases, such as the very early stages of an epidemic, or an endemic infection in a small population, a stochastic model may be more appropriate. In the simple case of a susceptible-infected (SI) model in a closed population (i.e. equations (2.5) and (2.6) with $\Lambda = \mu = 0$),

the timing of the coalescent events coincides with each transmission, and hence in this case, we can use the widely studied stochastic version of the SI model to model changes in ancestral lineages through time. However, in general, we cannot simply borrow from the epidemiological or population genetic literature. Most work on coalescent theory in finite populations has focused on birth–death processes (Hey 1992; Nee *et al.* 1994; Rannala 1997), either homogenous or non-homogenous, which are too simple for our purposes, while stochastic epidemiological models generally consider the dynamics of the process forward in time, rather than backwards, and do not consider the number of lineages. Unlike the deterministic models, in general we cannot simply run the nonlinear epidemiological models backwards in time from the present; for example, the stochastic version of the model (2.5) and (2.6) reaches a quasistationary state, at which point, the system has no 'memory' of when the first infection occurred.

The simple nature of the epidemiological models considered here allowed us to draw direct comparisons between population genetics models such as the Wright–Fisher and the Moran model, and epidemiological models. The correspondence between population genetic and epidemiological models becomes more complex in the case of heterogeneous populations; the models described here can be extended to consider heterogeneous populations, such as different contact rates, different infectivities, spatial structure and so on. For example, previously we considered a model of HIV infection which assumed two stages of infection, a brief, highly infectious acute period, followed by a long, less infectious chronic period (Volz *et al.* 2009), such that there is no longer a single rate of coalescence that applies to all individuals. In addition, for the simple models discussed here, the shape of the tree is captured by the dynamics of the number of lineages over time. However, phylogenetic trees contain more information than simply the number of lineages over time, for example tree balance, the distribution of the length of the terminal branches, and in the case of a heterogeneous population, the relative distribution of subpopulations across the tree. The development of new phylodynamic models will help to elucidate the role of epidemiological processes in generating these patterns.

S.D.W.F. is supported in part by a Royal Society Wolfson Research Merit Award. E.M.V. gratefully acknowledges support from the National Institutes of Health (grants T32 AI07384 and U01 GM087719). We would like to thank Santiago Elena and Rémy Froissart for organizing this symposium.

REFERENCES

Amore, G., Bertolotti, L., Hamer, G., Kitron, U., Walker, E., Ruiz, M., Brawn, J. & Goldberg, T. In press. Multi-year evolutionary dynamics of West Nile virus 1 in suburban Chicago, USA, 2005–2007. *Phil. Trans. R. Soc. Lond. B* **365**, 1871–1878. (doi:10.1098/rstb.2010.0054)

Baillie, G. J., Kolokotronis, S.-O., Waltari, E., Maffei, J. G., Kramer, L. D. & Perkins, S. L. 2008 Phylogenetic and evolutionary analyses of St Louis encephalitis virus genomes. *Mol. Phylogenet. Evol.* **47**, 717–728. (doi:10.1016/j.ympev.2008.02.015)

Bello, G., Eyer-Silva, W. A., Couto-Fernandez, J. C., Guimarães, M. L., Chequer-Fernandez, S. L., Teixeira, S. L. M. & Morgado, M. G. 2007 Demographic history of HIV-1 subtypes B and F in Brazil. *Infect. Genet. Evol.*, **7**, 263–270. (doi:10.1016/j.meegid.2006.11.002)

Bello, G., Guimarães, M. L., Passaes, C. P. B., Almeida, S. E. M., Veloso, V. G. & Morgado, M. G. 2009 Short communication: evidences of recent decline in the expansion rate of the HIV type 1 subtype C and CRF31_BC epidemics in southern Brazil. *AIDS Res. Hum. Retroviruses* **25**, 1065–1069. (doi:10.1089/aid.2009.0106)

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. 2009 Epidemic dynamics revealed in dengue evolution. *Mol. Biol. Evol.* **27**, 811–818. (doi:10.1093/molbev/msp285)

Biek, R. 2007 Evolutionary dynamics and spatial genetic structure of epizootic hemorrhagic disease virus in the eastern United States. *Infect. Genet. Evol.* **7**, 651–655. (doi:10.1016/j.meegid.2007.04.005)

Biek, R., Walsh, P. D., Leroy, E. M. & Real, L. A. 2006 Recent common ancestry of ebola Zaire virus found in a bat reservoir. *PLoS Pathog.* **2**, e90. (doi:10.1371/journal.ppat.0020090)

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. (doi:10.1073/pnas.0700741104)

Briese, T. *et al.* 2008 Global distribution of novel rhinovirus genotype. *Emerg. Infect. Dis.* **14**, 944–947. (doi:10.3201/eid1406.080271)

Carrington, C. V. F., Foster, J. E., Pybus, O. G., Bennett, S. N. & Holmes, E. C. 2005 Invasion and maintenance of dengue virus type 2 and type 4 in the Americas. *J. Virol.* **79**, 14 680–14 687. (doi:10.1128/JVI.79.23.14680-14687.2005)

Chen, R. & Holmes, E. C. 2006 Avian influenza virus exhibits rapid evolutionary dynamics. *Mol. Biol. Evol.* **23**, 2336–2341. (doi:10.1093/molbev/msl102)

Chen, R. & Holmes, E. C. 2008 The evolutionary dynamics of human influenza B virus. *J. Mol. Evol.* **66**, 655–663. (doi:10.1007/s00239-008-9119-z)

Davis, P. L., Rambaut, A., Bourhy, H. & Holmes, E. C. 2007 The evolutionary dynamics of canid and mongoose rabies virus in southern Africa. *Arch. Virol.* **152**, 1251–1258. (doi:10.1007/s00705-007-0962-9)

Deng, X., Liu, H., Shao, Y., Rayner, S. & Yang, R. 2008 The epidemic origin and molecular properties of B': a founder strain of the HIV-1 transmission in Asia. *AIDS* **22**, 1851–1858. (doi:10.1097/QAD.0b013e32830f4c62)

de Freitas, R. B., Melo, F. L., Oliveira, D. S., Romano, C. M., Freitas, M. R. C., Macdo, O., Linhares, A. C., de A Zanotto, P. M. & Durigon, E. L. 2008 Molecular characterization of human erythrovirus B19 strains obtained from patients with several clinical presentations in the Amazon region of Brazil. *J. Clin. Virol.* **43**, 60–65. (doi:10.1016/j.jcv.2008.03.033)

Donnelly, P. & Tavaré, S. 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**, 401–421. (doi:10.1146/annurev.ge.29.120195.002153)

Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192. (doi:10.1093/molbev/msi103)

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)

Fraser, C. *et al.* 2009 Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* **324**, 1557–1561. (doi:10.1126/science.117602)

Golemba, M. D., Lello, F. A. D., Bessone, F., Fay, F., Benetti, S., Jones, L. R. & Campos, R. H. 2010 High prevalence of hepatitis C virus genotype 1b infection in a small town of Argentina. Phylogenetic and Bayesian coalescent analysis. *PLoS One* **5**, e8751. (doi:10.1371/journal.pone.0008751)

Goñi, N., Fajardo, A., Moratorio, G., Colina, R. & Cristina, J. 2009 Modeling gene sequences over time in 2009 H1N1 influenza A virus populations. *Virol. J.* **6**, 215. (doi:10.1186/1743-422X-6-215)

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. (doi:10.1126/science.1090727)

Griffiths, R. C. & Tavaré, S. 1994 Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* **344**, 403–410. (doi:10.1098/rstb.1994.0079)

Hey, J. 1992 Using phylogenetic trees to study speciation and extinction. *Evolution* **46**, 627–640.

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. (doi:10.1098/rstb.1995.0088)

Hon, C.-C. *et al.* 2006 Phylogenetic analysis reveals a correlation between the expansion of very virulent infectious bursal disease virus and reassortment of its genome segment B. *J. Virol.* **80**, 8503–8509. (doi:10.1128/JVI.00585-06)

Hu, S., Pillay, D., Clewley, J. P. & Pybus, O. G. 2005 Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl Acad. Sci. USA* **102**, 4425–4429. (doi:10.1073/pnas.0407534102)

Huelsenbeck, J. P. & Lander, K. M. 2003 Frequent inconsistency of parsimony under a simple model of cladogenesis. *Syst. Biol.* **52**, 641–648.

Hughes, G. J., Páez, A., Bóshell, J. & Rupprecht, C. E. 2004 A phylogenetic reconstruction of the epidemiological history of canine rabies virus variants in Colombia. *Infect. Genet. Evol.* **4**, 45–51. (doi:10.1016/j.meegid.2003.12.001)

Jacquez, J. A. & Simon, C. P. 1993 The stochastic SI model with recruitment and deaths. I. Comparison with the closed SIS model. *Math. Biosci.* **117**, 77–125.

Jiménez-Hernández, N., Torres-Puente, M., Bracho, M. A., García-Robles, I., Ortega, E., del Olmo, J., Carnicer, F., González-Candelas, F. & Moya, A. 2007 Epidemic dynamics of two coexisting hepatitis C virus subtypes. *J. Gen. Virol.* **88**(Pt 1), 123–133. (doi:10.1099/vir.0.82277-0)

Jones, A., Lowry, K., Aaskov, J., Holmes, E. C. & Kitchen, A. 2010 Molecular evolutionary dynamics of Ross River virus and implications for vaccine efficacy. *J. Gen. Virol.* **91**(Pt 1), 182–188. (doi:10.1099/vir.0.014209-0)

Kingman, J. F. 2000 Origins of the coalescent: 1974–1982. *Genetics* **156**, 1461–1463.

Kinnear, M. & Linde, C. C. 2010 Capsid gene divergence in rabbit hemorrhagic disease virus. *J. Gen. Virol.* **91**(Pt 1), 174–181. (doi:10.1099/vir.0.014076-0)

Kitchen, A., Miyamoto, M. M. & Mulligan, C. J. 2008 Utility of DNA viruses for studying human host history: case study of JC virus. *Mol. Phylogenet. Evol.* **46**, 673–682. (doi:10.1016/j.ympev.2007.09.005)

Kurbanov, F. *et al.* 2003 Human immunodeficiency virus in Uzbekistan: epidemiological and genetic analyses. *AIDS Res. Hum. Retroviruses* **19**, 731–738. (doi:10.1089/088922203769232520)

Kurbanov, F., Tanaka, Y., Elkady, A., Oyunsuren, T. & Mizokami, M. 2007 Tracing hepatitis C and delta viruses to estimate their contribution in HCC rates in Mongolia. *J. Viral Hepat.* **14**, 667–674. (doi:10.1111/j.1365-2893.2007.00864.x)

Lemey, P., Pybus, O. G., Wang, B., Saksena, N. K., Salemi, M. & Vandamme, A.-M. 2003*a* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl Acad. Sci. USA* **100**, 6588–6592. (doi:10.1073/pnas.0936469100)

Lemey, P., Salemi, M., Wang, B., Duffy, M., Hall, W. H., Saksena, N. K. & Vandamme, A. M. 2003*b* Site stripping based on likelihood ratio reduction is a useful tool to evaluate the impact of non-clock-like behavior on viral phylogenetic reconstructions. *FEMS Immunol. Med. Microbiol.* **39**, 125–132.

Lemey, P., Pybus, O. G., Rambaut, A., Drummond, A. J., Robertson, D. L., Roques, P., Worobey, M. & Vandamme, A.-M. 2004 The molecular population genetics of HIV-1 group O. *Genetics* **167**, 1059–1068. (doi:10.1534/genetics.104.026666)

Liu, W. *et al.* 2008 Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. *PLoS Pathog.* **4**, e1000 097. (doi:10.1371/journal.ppat.1000097)

Michitaka, K., Tanaka, Y., Horiike, N., Duong, T. N., Chen, Y., Matsuura, K., Hiasa, Y., Mizokami, M. & Onji, M. 2006 Tracing the history of hepatitis B virus genotype D in western japan. *J. Med. Virol.* **78**, 44–52. (doi:10.1002/jmv.20502)

Minin, V. N., 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. (doi:10.1093/molbev/msn090)

Mohle, M. & Sagitov, S. 2001 A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29**, 1547–1562. (doi:10.1214/aop/1015345761)

Mooers, A. O. 1995 Tree balance and tree completeness. *Evolution* **49**, 379–384.

Moratorio, G., Costa-Mattioli, M., Piovani, R., Romero, H., Musto, H. & Cristina, J. 2007 Bayesian coalescent inference of hepatitis A virus populations: evolutionary rates and patterns. *J. Gen. Virol.* **88**(Pt 11), 3039–3042. (doi:10.1099/vir.0.83038-0)

Nakano, T., Lu, L., Liu, P. & Pybus, O. G. 2004 Viral gene sequences reveal the variable history of hepatitis C virus infection among countries. *J. Infect. Dis.* **190**, 1098–1108. (doi:10.1086/422606)

Nakano, T., Lu, L., He, Y., Fu, Y., Robertson, B. H. & Pybus, O. G. 2006 Population genetic history of hepatitis C virus 1b infection in China. *J. Gen. Virol.* **87**(Pt 1), 73–82. (doi:10.1099/vir.0.81360-0)

Nee, S., May, R. M. & Harvey, P. H. 1994 The reconstructed evolutionary process. *Phil. Trans. R. Soc. Lond. B* **344**, 305–311. (doi:10.1098/rstb.1994.0068)

Njouom, R., Nerrienet, E., Dubois, M., Lachenal, G., Rousset, D., Vessière, A., Ayouba, A., Pasquier, C. & Pouillot, R. 2007 The hepatitis C virus epidemic in Cameroon: genetic evidence for rapid transmission between 1920 and 1960. *Infect. Genet. Evol.* **7**, 361–367. (doi:10.1016/j.meegid.2006.10.003)

Njouom, R. *et al.* 2009 Predominance of hepatitis C virus genotype 4 infection and rapid transmission between

1935 and 1965 in the Central African Republic. *J. Gen. Virol.* **90**(Pt 10), 2452–2456. (doi:10.1099/vir.0.011981-0)

Opgen-Rhein, R., Fahrmeir, L. & Strimmer, K. 2005 Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol. Biol.* **5**, 6. (doi:10.1186/1471-2148-5-6)

Padhi, A. & Poss, M. 2009 Population dynamics and rates of molecular evolution of a recently emerged paramyxovirus, avian metapneumovirus subtype C. *J. Virol.* **83**, 2015–2019. (doi:10.1128/JVI.02047-08)

Padhi, A. *et al.* 2008 Phylogeographical structure and evolutionary history of two Buggy Creek virus lineages in the western Great Plains of North America. *J. Gen. Virol.* **89**(Pt 9), 2122–2131. (doi:10.1099/vir.0.2008/001719-0)

Paradis, E., Claude, J. & Strimmer, K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290.

Pereira, C. A. D., Leal, E. S. & Durigon, E. L. 2007 Selective regimen shift and demographic growth increase associated with the emergence of high-fitness variants of canine parvovirus. *Infect. Genet. Evol.* **7**, 399–409. (doi:10.1016/j.meegid.2006.03.007)

Pérez-Losada, M., Jobes, D. V., Sinangil, F., Crandall, K. A., Posada, D. & Berman, P. W. 2010 Phylodynamics of HIV-1 from a phase-III AIDS vaccine trial in North America. *Mol. Biol. Evol.* **27**, 417–425. (doi:10.1093/molbev/msp254)

Petzoldt, T. & Rinke, K. 2007 simecol: an object-oriented framework for ecological modeling in R. *J. Statist. Softw.* **22**, 1–31.

Pitman, J. 1999 Coalescents with multiple collisions. *Ann. Probab.* **27**, 1870–1902. (doi:10.1214/aop/1022677552)

Pomeroy, L. W., Bjornstad, O. N. & Holmes, E. C. 2008 The evolutionary and epidemiological dynamics of the paramyxoviridae. *J. Mol. Evol.* **66**, 98–106. (doi:10.1007/s00239-007-9040-x)

Purvis, A. & Agapow, P.-M. 2002 Phylogeny imbalance: taxonomic level matters. *Syst. Biol.* **51**, 844–854.

Pybus, O. G. & Rambaut, A. 2009 Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550. (doi:10.1038/nrg2583)

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., Holmes, E. C. & Harvey, P. H. 2002 New inferences from tree shape: numbers of missing taxa and population growth rates. *Syst. Biol.* **51**, 881–888.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H. & Rambaut, A. 2003 The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. *Mol. Biol. Evol.* **20**, 381–387.

Pybus, O. G., Cochrane, A., Holmes, E. C. & Simmonds, P. 2005 The hepatitis C virus epidemic among injecting drug users. *Infect. Genet. Evol.* **5**, 131–139. (doi:10.1016/j.meegid.2004.08.001)

Pybus, O. G. *et al.* 2009 Genetic history of hepatitis C virus in East Asia. *J. Virol.* **83**, 1071–1082. (doi:10.1128/JVI.01501-08)

R Development Core Team. 2009 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Rambaut, A. & Holmes, E. 2009 The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Curr. Influenza* RRN1003.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K. & Holmes, E. C. 2008 The genomic and epidemiological dynamics of human influenza A virus. *Nature* **453**, 615–619. (doi:10.1038/nature06945)

Rannala, B. 1997 Gene genealogy in a population of variable size. *Heredity* **78** (Pt 4), 417–423.

Rannala, B., Huelsenbeck, J. P., Yang, Z. & Nielsen, R. 1998 Taxon sampling and the accuracy of large phylogenies. *Syst. Biol.* **47**, 702–710.

Robbins, K. E., Lemey, P., Pybus, O. G., Jaffe, H. W., Youngpairoj, A. S., Brown, T. M., Salemi, M., Vandamme, A.-M. & Kalish, M. L. 2003 U.S. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.* **77**, 6359–6366.

Romano, C. M., de A Zanotto, P. M. & Holmes, E. C. 2008 Bayesian coalescent analysis reveals a high rate of molecular evolution in GB virus C. *J. Mol. Evol.* **66**, 292–297. (doi:10.1007/s00239-008-9087-3)

Sagitov, S. 1999 The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36**, 1116–1125. (doi:10.1239/jap/1032374759)

Sagitov, S. 2003 Convergence to the coalescent with simultaneous multiple mergers. *J. Appl. Probab.* **40**, 839–854.

Salemi, M., de Oliveira, T., Soares, M. A., Pybus, O., Dumans, A. T., Vandamme, A.-M., Tanuri, A., Cassol, S. & Fitch, W. M. 2005 Different epidemic potentials of the HIV-1B and C subtypes. *J. Mol. Evol.* **60**, 598–605. (doi:10.1007/s00239-004-0206-5)

Salemi, M., de Oliveira, T., Ciccozzi, M., Rezza, G. & Goodenow, M. M. 2008 High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. *PLoS One* **3**, e1390. (doi:10.1371/journal.pone.0001390)

Sanderson, M. J. 2003 r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302.

Schreiber, M. J. *et al.* 2009 Genomic epidemiology of a dengue virus epidemic in urban Singapore. *J. Virol.* **83**, 4163–4173. (doi:10.1128/JVI.02445-08)

Schweinsberg, J. 2000 Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5**, 1–50.

Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. 2002*a* Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics* **160**, 1283–1293.

Seo, T.-K., Thorne, J. L., Hasegawa, M. & Kishino, H. 2002*b* A viral sampling design for testing the molecular clock and for estimating evolutionary rates and divergence times. *Bioinformatics* **18**, 115–123.

Snapinn, K. W., Holmes, E. C., Young, D. S., Bernard, K. A., Kramer, L. D. & Ebel, G. D. 2007 Declining growth rate of West Nile virus in North America. *J Virol*, **81**, 2531–2534. (doi:10.1128/JVI.02169-06)

Stack, J. C., Welch, J. D., Ferrari, M. J., Shapiro, B. U. & Grenfell, B. T. In press. Protocols for sampling viral sequences to study epidemic dynamics. *J. R. Soc. Interface*. (doi:10.1098/rsif.2009.0530)

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

Tanaka, Y. *et al.* 2004 Exponential spread of hepatitis C virus genotype 4a in Egypt. *J. Mol. Evol.* **58**, 191–195. (doi:10.1007/s00239-003-2541-3)

Tanaka, Y. *et al.* 2006 Molecular tracing of Japan-indigenous hepatitis E viruses. *J. Gen. Virol.* **87**(Pt 4), 949–954. (doi:10.1099/vir.0.81661-0)

Tee, K. K., Pybus, O. G., Li, X.-J., Han, X., Shang, H., Kamarulzaman, A. & Takebe, Y. 2008 Temporal and spatial dynamics of human immunodeficiency virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. *J. Virol.* **82**, 9206–9215. (doi:10.1128/JVI.00399-08)

Twiddy, S. S., Pybus, O. G. & Holmes, E. C. 2003 Comparative population dynamics of mosquito-borne flaviviruses. *Infect. Genet. Evol.* **3**, 87–95.

van Ballegooijen, W. M., van Houdt, R., Bruisten, S. M., Boot, H. J., Coutinho, R. A. & Wallinga, J. 2009 Molecular sequence data of hepatitis B virus and genetic diversity after vaccination. *Am. J. Epidemiol.* **170**, 1455–1463. (doi:10.1093/aje/kwp375)

van Houdt, R., Bruisten, S. M., Geskus, R. B., Bakker, M., Wolthers, K. C., Prins, M. & Coutinho, R. A.. 2010 Ongoing transmission of a single hepatitis B virus strain among men having sex with men in Amsterdam. *J. Viral Hepat.* **17**, 108–114. (doi:10.1111/j.1365-2893.2009.01158.x)

Verbeeck, J. *et al.* 2006 Investigating the origin and spread of hepatitis C virus genotype 5a. *J. Virol.* **80**, 4220–4226. (doi:10.1128/JVI.80.9.4220-4226.2006)

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. (doi:10.1534/genetics.109.106021)

Wakeley, J. & Sargsyan, O. 2009 Extensions of the coalescent effective population size. *Genetics* **181**, 341–345. (doi:10.1534/genetics.108.092460)

Walker, P. R., Pybus, O. G., Rambaut, A. & Holmes, E. C. 2005 Comparative population dynamics of HIV-1 subtypes B and C: subtype-specific differences in patterns of epidemic growth. *Infect. Genet. Evol.* **5**, 199–208. (doi:10.1016/j.meegid.2004.06.011)

Worobey, M. *et al.* 2008 Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661–664. (doi:10.1038/nature07390)

Yang, Z., O'Brien, J. D., Zheng, X., Zhu, H.-Q. & She, Z.-S. 2007 Tree and rate estimation by local evaluation of heterochronous nucleotide data. *Bioinformatics* **23**, 169–176. (doi:10.1093/bioinformatics/btl577)

Zehender, G., Maddalena, C. D., Giambelli, C., Milazzo, L., Schiavini, M., Bruno, R., Tanzi, E. & Galli, M. 2008 Different evolutionary rates and epidemic growth of hepatitis B virus genotypes A and D. *Virology* **380**, 84–90. (doi:10.1016/j.virol.2008.07.009)

Zehender, G., Maddalena, C. D., Canuti, M., Zappa, A., Amendola, A., Lai, A., Galli, M. & Tanzi, E. 2009 Rapid molecular evolution of human bocavirus revealed by Bayesian coalescent inference. *Infect. Genet. Evol.* (doi:10.1016/j.meegid.2009.11.011)