Phylogenetic and epidemic modeling of rapidly evolving infectious diseases

Originally published in:
Infection, Genetics and Evolution 11(8), https://doi.org/10.1016/j.meegid.2011.08.005

Infection, Genetics and Evolution 11 (2011) 1825–1841

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

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

Review

Phylogenetic and epidemic modeling of rapidly evolving infectious diseases

Denise Kühnert [1], Chieh-Hsi Wu [1], Alexei J. Drummond *

Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand
Department of Computer Science, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 1 April 2011
Received in revised form 9 August 2011
Accepted 9 August 2011
Available online 31 August 2011

Keywords:
Coalescent
Phylodynamics
Statistical phylogeography
Phylogenetic epidemiology
Rapidly evolving viruses
Stochastic SIR

ABSTRACT

Epidemic modeling of infectious diseases has a long history in both theoretical and empirical research. However the recent explosion of genetic data has revealed the rapid rate of evolution that many populations of infectious agents undergo and has underscored the need to consider both evolutionary and ecological processes on the same time scale. Mathematical epidemiology has applied dynamical models to study infectious epidemics, but these models have tended not to exploit – or take into account – evolutionary changes and their effect on the ecological processes and population dynamics of the infectious agent. On the other hand, statistical phylogenetics has increasingly been applied to the study of infectious agents. This approach is based on phylogenetics, molecular clocks, genealogy-based population genetics and phylogeography. Bayesian Markov chain Monte Carlo and related computational tools have been the primary source of advances in these statistical phylogenetic approaches. Recently the first tentative steps have been taken to reconcile these two theoretical approaches. We survey the Bayesian phylogenetic approach to epidemic modeling of infection diseases and describe the contrasts it provides to mathematical epidemiology as well as emphasize the significance of the future unification of these two fields.

Contents

1. Introduction

Molecular phylogenetics has had a profound impact on the study of infectious diseases, particularly rapidly evolving infectious

* Corresponding author.
E-mail address: alexei@cs.auckland.ac.nz (A.J. Drummond).
[1] These authors contributed equally to this work.

1567-1348 © 2011 Elsevier B.V. Open access under CC BY-NC-ND license.
doi:10.1016/j.meegid.2011.08.005

agents such as RNA viruses. It has given insight into the origins, evolutionary history, transmission routes and source populations of epidemic outbreaks and seasonal diseases. One of the key observations about rapidly evolving viruses is that the evolutionary and ecological processes occur on the same time scale (Pybus and Rambaut, 2009). This is important for two reasons. First, it means that neutral genetic variation can track ecological processes and population dynamics, providing a record of past evolutionary events (e.g., genealogical relationships) and past ecological/population events (geographical spread and changes in population size and structure) that were not directly observed. Second, the concomitance of evolutionary and ecological processes leads to their interaction that, when non-trivial, necessitates joint analysis.

Arguably the most studied infectious disease agent to date has been human immunodeficiency virus (HIV) and it has been the subject of thousands of phylogenetic studies. These have shed light on many aspects of HIV evolutionary biology, epidemiology, origins, phylogeography, transmission dynamics and drug resistance. In fact, the vast body of literature on HIV makes it clear that almost every aspect of the biology of a rapidly evolving pathogen can be better understood in the context of the evolution of the virus. Whether it is retracing the zoonotic origins of the HIV pandemic or describing the interplay between the virus population and its host's immune system, a phylogenetic analysis frequently sheds light.

Although probabilistic modeling approaches to phylogenetics predate Sanger sequencing (Edwards and Cavalli-Sforza, 1965), it was not until the last decade that probabilistic modeling became the dominant approach to phylogeny reconstruction. Part of that dominance has been due to the rise of Bayesian inference (Huelsenbeck et al., 2001), with its great flexibility in describing prior knowledge, its ability to be applied via the Metropolis-Hastings algorithm to complex highly parametric models, and the ease with which multiple sources of data can be integrated into a single analysis. The history of probabilistic models of molecular evolution and phylogenetics is a history of gradual refinement; a process of selection of those modeling variations that have the greatest utility in characterizing the ever-growing empirical data. The utility of a new model has been evaluated either by how well it fits the data (formal model comparison or goodness-of-fit tests) or by the new questions that it allows a researcher to ask of the data. In this review we will describe the modern phylogenetic approach to the field of infectious diseases, and particularly with reference to Bayesian inference of the phylogenetic epidemiology of rapidly evolving viral pathogens such as Hepatitis C virus (HCV), HIV and Influenza A virus. The review is separated into two main sections. In Section 2 we discuss phylogenetic methods for reconstructing the history of infectious epidemics, including identification of origins, dating of common ancestors, relaxed phylogenetics and coalescent-based population dynamics. In Section 3 we review epidemiological models and finish by outlining progress in the development of phylodynamical models that marry statistical phylogenetics with dynamical modeling.

2. Reconstructing the history of infectious epidemics

The introduction of an efficient means of calculating the probability of a sequence alignment given a phylogenetic tree (known as the phylogenetic likelihood; Felsenstein, 1981) heralded the beginning of practical phylogenetic tree reconstruction in a statistical framework. At around the same time the coalescent was introduced: a theory relating the shape of the genealogy of a random sample of individuals to the size of the population from which they came (Kingman, 1982; see Section 2.3 for details). Both of these advances have been subsequently developed to the point that, together they enable the estimation of viral evolutionary histories and past population dynamics.

Bayesian inference brings together the *likelihood*, $\Pr(D|\theta)$ (the probability of the data given the model parameters) and the *prior*, $P(\theta)$ (the probability of the model parameters prior to seeing the data), so that the *posterior* probability of the model parameters (θ) given the data is:

$$P(\theta|D) = \frac{\Pr(D|\theta)P(\theta)}{\int \Pr(D|\theta)P(\theta)d\theta} \quad (1)$$

In a standard phylogenetic setting, the probabilistic model parameters include the phylogenetic tree, coalescent times and substitution parameters, and a prior probability distribution over these parameters must be specified. By using Kingman's coalescent as a prior density on trees, Bayesian inference can be used to simultaneously estimate the phylogeny of the viral sequences and the demographic history of the virus population (Drummond et al., 2002, 2005, 2006, see Box 1). Extension of phylogenetic inference methods to accommodate time-stamped sequence data (Rambaut, 2000; Drummond et al., 2002) and relaxation of the assumption of a strict molecular clock (Thorne et al., 1998; Kishino et al., 2001; Sanderson, 2002; Drummond et al., 2006; Rannala and Yang, 2007) provided sophisticated methods for ancestral divergence time estimation. For virus species that occupy more than one host species (e.g Influenza A), models that aim to detect cross-species transmission may provide clues to the origin of a virus strain in a host population (Reis et al., 2009).

2.1. Reconstructing the origins of an infectious disease

When a new epidemic emerges, one of the first goals is to trace it back to its genetic and geographic origin. The reconstruction of phylogenetic trees to infer the evolutionary relationships has been a key tool to uncover the origin of regional epidemics such as those resulting from HIV (Gao et al., 1999; Santiago et al., 2002), HCV (Pybus et al., 2009; Markov et al., 2009) and SARS coronavirus (SARS-CoV) (Li et al., 2005). Some studies have also attempted to use phylogenetic trees to draw conclusions about transmission history and geographic spread of viral epidemics (Motomura et al., 2003; Santiago et al., 2005; Gilbert et al., 2007). However, great care should be taken when coming to conclusions about aspects of the epidemic process that are not explicitly modeled in the reconstruction of the phylogenetic tree and even if they are, the user needs to consider the appropriateness of the underlying model assumptions.

One common and straightforward method used to identify the origin of an epidemic involves determining the non-epidemic genotype or lineage most closely related to the epidemic, i.e., the molecular sequences clustered most closely with the epidemic strain on a phylogenetic tree. While the method is intuitive, its success heavily depends on the collected data.

The closest simian immunodeficiency virus (SIV) relative of HIV-1 is SIVcpz (Gao et al., 1999; Santiago et al., 2002), which is harbored in chimpanzee sub-species *Pan troglodytes troglodytes* and *P.t. schweinfurthii* in the form of the respective sub-species specific SIV lineages SIVcpz*Ptt* and SIVcpz*Pts*. Although SIVcpz became the prime candidate for the zoonotic source of HIV-1 as soon as it was identified, alternative sources could not be ruled out due to the paucity of identified chimpanzee infections (Vanden Haesevelde et al., 1996). The source of HIV-1 was confirmed much later after the collection of SIVcpz from fecal samples of wild *P. t. troglodytes* apes in the Cameroon forest (Keele et al., 2006). HIV-1 groups M and N are much more closely related to sequences from the fecal samples than previously identified SIVcpz strains. This finding uncovered the distinct origins of HIV-1 group M (pandemic) and group N (non-pandemic) traced to chimpanzee communities of southeastern and central Cameroon respectively. The

precise geographic identification of these wildlife chimpanzee reservoirs of HIV-1 by phylogenetic techniques provided the crucial evidence that SIVcpz gave rise to the HIV/AIDS pandemic.

Conversely, if strains sufficiently closely related to the epidemic strain cannot be identified then phylogenetic trees are not able to easily provide answers about origins. For example, there has been much heated debate on the origin of the 1918 H1N1 Influenza A pandemic – whether its source was avian, non-human mammalian or even human. The uncertainty mainly stems from the absence of sequences from the immediate ancestral source population of the 1918 virus (Gibbs and Gibbs, 2006).

A similar, though less severe problem has been encountered with the search for the origin of HIV-1 O group. Strains of HIV-1 O group have been revealed to be most closely related to SIVgor found in Western lowland gorillas (*Gorilla gorilla gorilla*) (Van Heuverswyn et al., 2006; Takehisa et al., 2009). However, HIV-O sequences are moderately divergent from the known SIVgor sequences and consequently, the route of transmission that has given rise to HIV-1 O group and SIVgor is still indeterminate.

The interspersion of an emergent viral strain with other strains in a phylogenetic tree is often interpreted as evidence supporting multiple independent viral introductions. For example, HIV lineages are paraphyletic with SIV lineages creating several separate clusters of HIV suggesting multiple zoonotic viral transmissions into the human population (Santiago et al., 2005; Keele et al., 2006). While it is intuitive that separate clusters of the emergent virus suggest multiple introductions, it is not clear from the number of clusters alone how many independent events are responsible for the observed pattern. Incomplete taxon sampling will lead to undercounting. For example, there may exist an unsampled sequence that will split an emergent viral cluster, or an additional unsampled emergent cluster. Both scenarios, if detected, would increase the lower bound of the inferred number of events. The number of events could also be incorrectly estimated due to phylogenetic estimation error. Finally, in situations where the event is potentially reversible, such as with drug-resistance mutations, e.g., adamantane resistance in H3N2 influenza virus (Nelson et al., 2009), it is quite possible that reversions are also present in the phylogenetic history, and these are not always detectable by a simple parsimony reconstruction, again leading to undercounting. For all these reasons, the applications of Bayesian modeling of phylogeography and character evolution on phylogenies is crucial to quantitatively assess the uncertainty generated from these different sources of error (see Section 2.4).

In contrast to HIV-1, it has been clearly established for almost two decades that the progenitor of HIV-2 is SIVsm from sooty mangabey (*Cerocebus torquatus atys*) (Hirsch et al., 1989; Gao et al., 1992). It was suggested by (Santiago et al., 2005) that the geographic origin of HIV-2 groups A and B are in the eastern sooty mangabey range according to the clear geographic clustering displayed in the phylogenetic tree and branching position of the HIV-2 strains. Although this heuristic approach to locating phylogeographic origins is commonly used, it has several disadvantages aside from the sampling error mentioned earlier. First, it relies on strong geographic signals to produce an unambiguous geographic clustering pattern in the trees. Second, the lack of a formal statistical framework results in an inability to quantify the associated uncertainty with the geographic estimates. A number of statistical phylogenetic methods aim to reconstruct the migration process by treating geographic locations as another state that evolves down the tree. The states are either discrete (Lemey et al., 2009b), denoted by names of cities or provinces, or continuous represented by the latitude and longitude of the location (Biek et al., 2006, 2007; Lemey et al., 2010).

Even with comprehensive sampling, using a single phylogenetic tree is insufficient to reflect the complex genetic origin of virus species that undergo recombination or reassortment. Reassortment arises when segments of the viral genome come from different viruses, while recombination also requires the genetic material from one source to (break and) join with that from another. These two processes enable the generation of novel combinations from two existing genotypes. Moreover, these often large genetic changes may provide the potential for adaptation to a new host species (Parrish et al., 2008). Reassortment has played an important role in the evolution of the Influenza A virus (Lindstrom et al., 2004; Holmes et al., 2005; Nelson et al., 2008). Evidence for recombination have also been found in Dengue (Holmes et al., 1999), HIV, HCV and SARS-CoV (Li et al., 2006).

There are many phylogenetic methods that aim to detect recombination by identifying discordance in the topologies of different parts of the alignment (Grassly and Holmes, 1997; Salminen et al., 1995; Lole et al., 1999; Smith, 1992; Robertson et al., 1995; Paraskevis et al., 2005), which is a potential consequence of recombination. Most of these methods use a sliding window approach to compute a summary statistic along the length of sequence. Phylogenetic approaches are based on estimating either (i) bootstrap values or (ii) clade posterior probabilities for each window and a sudden change in bootstrap value, clade posterior probability or site percentage identity is an indication of the presence of a breakpoint around the region. Other methods explicitly estimate the position of the breakpoint in an alignment, providing access to test the strength of support for recombination (Holmes et al., 1999). Finally, some approaches portray the evolutionary history by networks to incorporate horizontal transfer (Huson, 1998) or ancestral recombination graphs (Bloomquist and Suchard, 2010).

2.2. Dating of ancestors

As a rule, RNA viruses mutate rapidly, so that viruses isolated only a few months apart may exhibit measurable genetic differences (Drummond et al., 2003a and references therein). Indeed, the mutation rate of some RNA viruses is so high that it can result in evolutionary changes within a host during the course of infection. This is particularly true of long term chronic infections caused by viruses such as HIV and HCV. It is therefore not appropriate to consider the analysis of sequences that have been sampled years apart as if they are contemporaneous. Sequence data with this type of temporal structure are called heterochronous and from such

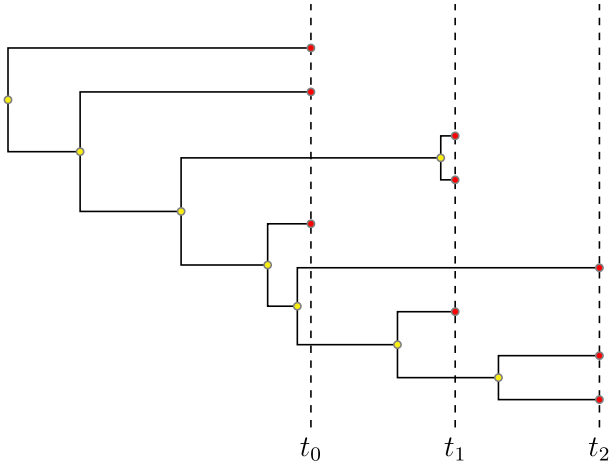

Fig. 1. A serially sampled time tree of a rapidly evolving virus, showing that the sampling time interval $[t_0,t_2]$ represents a substantial fraction of the time back to the common ancestor. Red circles represent sampled viruses (three viruses sampled at each of three times) and yellow circles represent hypothetical common ancestors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data the substitution rate can be estimated and divergence times calibrated to a calendar scale. Here, a tree with branch lengths in calendar units is termed a "time tree". Fig. 1 depicts an example of a serially sampled time tree of a rapidly evolving virus.

To account for temporal structure in sequence data, the earliest methods estimated the time scale by estimating a gene tree with unconstrained branch lengths and then performing a linear regression of root-to-tip genetic distance against sampling times (see for review Drummond et al., 2003b). This method was used to provide the first estimate of the time of the most recent common ancestor (t_{MRCA}) of HIV-1 M group, placing it in the 1930s (Korber et al., 2000). Despite its simplicity, this method also accurately estimated the age of the oldest HIV sequence sampled in 1959. A maximum likelihood based method (the single rate dated tips (SRDT) model; Rambaut, 2000), estimates ancestral divergence times and overall substitution rate on a fixed tree, assuming a strict molecular clock. The SRDT model was used to date the most recent common ancestor of HIV-2 subtype A in 1940 ±16 and that of subtype B in 1945 ±14 (Lemey et al., 2003). Using the serial coalescent as a tree prior in Bayesian coalescent methods (Drummond et al., 2002, 2005; Drummond and Rambaut, 2007) allows the time scale to be simultaneously estimated with other phylogenetic and demographic parameters. Recently, a relaxed clock Bayesian coalescent analysis that included two historical viral samples from 1959 (ZR59) and 1960 (DRC60) (Worobey et al., 2008), pushed back the estimated t_{MRCA} of HIV-1 M group to 1908 (1884–1924).

Besides estimating the time of an epidemic outbreak, it may also be important to know how long the ancestors of the epidemic strain had circulated in the source population prior to the epidemic. This can sometimes be indicated by the length of the branch ancestral to the epidemic clade. In the case of the 2009 Swine-origin Influenza A virus, the length of the branch leading to 2009 S-OIV strains is estimated to be 9–17 years depending on the viral segment analyzed, suggesting roughly a decade of unsampled diversity (Smith et al., 2009).

To estimate the age of the common ancestor of SIVsm strains, the t_{MRCA} of HIV-2/SIVsm has been dated, indicating that the common ancestry prior the zoonosis of HIV-2 group A and B spans only the last few centuries (Wertheim and Worobey, 2009). This does not necessarily indicate that SIVsm first arose only centuries ago, just that the common ancestor of all current SIVsm may be recent. However, even this conclusion has recently been questioned (Worobey et al., 2010) as a result of independent calibration evidence that suggests the t_{MRCA} could in fact be greater than 32,000 years ago, leading to debate about the fidelity of the statistical substitution models commonly employed for divergence time dating when the true divergence times are very ancient compared to the sampling interval. As demonstrated by Wertheim and Pond, 2011, substitution models that do not take into account the effects of selection can produce underestimated branch lengths leading to much younger age estimates in presence of purifying selection. This will be more problematic for data sets for which the total sampling interval is only a small fraction of the total age of the tree.

While incorporating sampling dates provides additional information to phylogenetic inference, it also implies that the reliability of those dates has a heavy impact on the validity of the inference. The H1N1 influenza virus that re-emerged in 1977 was found to have missed decades of evolution and was genetically remarkably similar to the H1N1 1950 virus (Nakajima et al., 1978). It is thus thought to be descended from a strain that was kept frozen in an unknown laboratory for perhaps decades before again becoming a "wild" strain again (Zimmer and Burke, 2009). If the missing evolution is not corrected for, analyses including the re-emergent strains produce biased date estimates and increased variances of the t_{MRCA} of the re-emergent lineages and across the phylogeny (Wertheim, 2010). In cases where the sampling dates of sequences are contentious or unknown, a method that can handle sequences with unknown dates is required. For example, the leaf-dating method estimates the unknown date or age of a sequence as a parameter, treating it the same way as the age of internal nodes (Drummond et al., 2003c; Nicholls and Gray, 2008; Shapiro et al., 2010).

Unrealistic sampling dates may also be the result of human error and are thus not recognized prior to an analysis. Therefore, diagnostics for unrealistic dates are important to pick up errors in the recorded dates. One possible method is to plot the root-to-tip genetic distance against sampling year if the virus does not display significant departure from constant rate (Wertheim, 2010). Another is to check calibrations by dropping each calibration point in turn and re-estimating the date to confirm that the estimated dates are consistent (Shapiro et al., 2010; Ryder and Nicholls, 2011).

2.2.1. Relaxed molecular clocks

Early methods that accommodated heterochronous data assumed a strict clock model. However, a comprehensive study of heterochronous RNA viral sequences using the SRDT model (Rambaut, 2000) demonstrated that the majority of the 50 RNA viral species studied rejected the constant rate molecular clock hypothesis (Jenkins et al., 2002). The unrooted phylogeny is the other extreme of the scale of rate variability across branches of a phylogenetic tree. Neither of them is a realistic representation of the underlying evolutionary process and the reality lies somewhere between the two. This has spawned the development of numerous methods that relax the molecular clock assumption and differ in their assumption of the pattern of rate variation across the branches.

The local clock model approach assigns different rates to clades/regions of the tree. However, without external information, it is difficult to know *a priori* what is the best partitioning of the tree into local clock models. Bayesian model averaging overcomes the challenge of rate assignment by averaging over all possible local clock models (Drummond and Suchard, 2010), estimating the substitution rates, and the number and position of changes in substitution rate, simultaneously.

Another category of relaxed clock models is based on 'rate smoothing', including non-parametric rate smoothing (Sanderson, 1997), penalized likelihood (Sanderson, 2002) and Bayesian autocorrelated relaxed clock methods (Thorne et al., 1998; Kishino et al., 2001; Aris-Brosou and Yang, 2002; Rannala and Yang, 2007). These methods restrict the rates on parent and descendant branches to be similar by penalizing large departures from parent branch rates. Hence, rate variation is expected to occur through small and frequent changes. Different Bayesian autocorrelated clock models differ in the distribution used to model a branch rate given its parent rate (Thorne et al., 1998; Kishino et al., 2001).

However, analysis of sequence data from Influenza A and Dengue-4 do not provide any evidence of autocorrelation of branch rates (Drummond et al., 2006) suggesting that autocorrelated models may not be appropriate when analyzing a genealogy of sequences from a single virus species. Whereas lineage-effects may be expected to cause autocorrelation of rates (through incremental changes to life-history, metabolic rate *et cetera*), the gene-specific action of Darwinian selection will also cause apparent rate variation among lineages, by producing a general over-dispersion of the molecular clock over the entire phylogeny (Takahata, 1987, 1991). This second source of rate variation among lineages may be better modeled by uncorrelated relaxed clock models (Drummond et al., 2006), which make no assumption about the autocorrelation of rates between ancestral and descendent branches. Published analyses have provided strong evidence supporting the uncorrelated relaxed clock model (e.g., Salemi et al., 2008; Worobey et al., 2008) over the strict clock model.

As well as estimating the age of ancestral divergences, it is also of interest to estimate the time of cross-species transmission if the

disease is zoonotic in origin. One method of identifying the time of the host-switch is by applying non-homogeneous substitution models. The motivation of non-homogeneous substitution models is to acknowledge possible differences in pattern of substitution in the virus within different host species, which violates the assumptions of homogeneity and stationarity underlying the standard substitution models. Therefore it may be more appropriate to apply different substitution models to different parts of the tree (Forsberg and Christiansen, 2003). Non-homogeneous substitution models permit the equilibrium frequencies, and hence the model parameters, to change on a branch and all the descendant lineages from the point of change are assumed to have different equilibrium base frequencies to the lineages prior to that point. This technique has been used to suggest that the immediate ancestral population of 1918 Influenza A virus resided in a mammalian host (Reis et al., 2009). However, it does not indicate whether the most recent common ancestor of the swine Influenza virus and the 1918 virus resided in humans or other mammals.

2.2.2. Interpretation and accuracy of divergence time estimates

Interpretation of estimated divergence times can be difficult. There may be direct ancestors that are more ancient, but the lineages that would reveal them have not been sampled or did not survive to the present due to processes such as genetic drift. Therefore, the estimated t_{MRCA} may not answer the question of interest. For epidemics that resulted from a zoonotic transmission, the host switch event is of paramount interest, but estimating the t_{MRCA} of the epidemic strain does not directly estimate the time of the transmission, and only serves as a lower bound. Likewise, if there have been processes causing a loss of genetic diversity in the past or the sampling is not comprehensive, then the estimated t_{MRCA} could be substantially younger than the age of the viral lineage. An obvious example of the former occurs in seasonal influenza due to seasonal population fluctuations and also strong positive Darwinian selection caused by immune surveillance (Fitch et al., 1991; Bush et al., 1999), leading to rapid lineage turnover and a recent common ancestor of any single-season sample.

Similarly, the analysis by Worobey et al. (2008) shows that the t_{MRCA} of HIV-1 group M seems to have been pushed back due to the inclusion of an additional pre-epidemic sample from 1960 which is highly divergent to the 1959 sequence (ZR59). In general the inclusion of older samples can increase the estimated age of root by (i) revealing previously unsampled lineages that are outgroup to the t_{MRCA} estimated without them, or (ii) simply because more temporal sampling breaks up long internal branches as well as potentially revealing ancient evidence of variants that were assumed modern, resulting in a slower estimated rate and therefore older estimated root height.

Finally, it is likely that current techniques alone cannot always recover accurate divergence dates in the distant past, as illustrated by recent analyses suggesting a much deeper history of SIV (Worobey et al., 2010) than previously suggested (Sharp et al., 2000; Wertheim and Worobey, 2009). Fig. 2 illustrates the problem with three estimated viral time-trees that have vastly different inferred ages of their most recent common ancestor. We would expect the greatest confidence in the inferred age of the human influenza A time-tree where the sample period is a large fraction of the total age of the time tree, and the least confidence in the inferred age of the Hepatitis C time-tree in which the sampling period is a small fraction of the inferred age of greater than 1000 years.

So, apart from better models of rate variation across lineages (see Guindon et al., 2004, for early steps in this direction), future research in divergence time dating will likely focus on models that more accurately account for purifying selection and its role in maintaining the structure and function of the encoded genes. The impact of Darwinian selection is expressed both in distortions of the genealogy (O'Fallon, 2010; O'Fallon et al., 2010) and the substitution process (e.g., Bloom et al., 2007; Cartwright et al., 2011) from neutral expectations. Consideration of the action of pervasive purifying selection is especially important in viral genomes prone to clonal interference and which are compact, information rich and subject to great levels of functional and structural constraint in their evolutionary trajectories, especially when considering long time periods. Beyond that there is also a need for more statistically rigorous methods of incorporating diverse sources of calibration information, such as biogeography, archaeology and paleontological evidence. Bayesian statistical frameworks are uniquely suited for this sort of integration of multiple sources of information.

2.3. Genealogy-based population dynamics

Genealogy-based population genetics can be used to infer demographic parameters including population size, rate of growth or decline, and population structure. When the characteristic time scale of demographic fluctuations are comparable to the rate of accumulations of substitutions then past population dynamics are "recorded" in the substitution patterns of molecular sequences. Coalescent theory can therefore be combined with temporal information in heterochronous sequences to uncover past epidemiological events and pinpoint them on a calendar time scale.

Kingman's coalescent (Kingman, 1982) describes the relationship between the coalescent times in a sample genealogy and the population size assuming an idealized Wright–Fisher population (Fisher, 1930; Wright, 1931). The original formulation was for a constant population, but the theory has since been generalized to any deterministically varying function of population size for which the integral $\int_{t_0}^{t_1} N(t)^{-1} dt$ can be computed (Griffiths and Tavaré, 1994). Parametric models with a pre-defined population function, such as exponential growth, expansion model and logistic growth models can easily be used in a coalescent framework (see Fig. 3 and Box 1 for details). For example a "piecewise-logistic" population model was employed in a Bayesian coalescent framework to estimate the population history of HCV genotype 4a infections in Egypt (Pybus et al., 2003). This analysis demonstrated a rapid expansion of HCV in Egypt between 1930-1955, consistent with the hypothesis that public health campaigns to administer anti-schistosomiasis injections had caused the expansion of an HCV epidemic in Egypt.

The coalescent process is highly variable, so sampling multiple unlinked loci (Felsenstein, 2006; Heled and Drummond, 2008) or increasing the temporal spread of sampling times (Seo et al., 2002) can both be used to increase the statistical power of coalescent-based methods and improve the precision of estimates of both population size and substitution rate (Seo et al., 2002). However in many virus species, the entire genome acts as a single locus, or undergoes recombination only when the opportunity arises through superinfection. The lack of independent loci therefore places an upper limit on the precision of estimates of population history.

In many situations the precise functional form of the population size history is unknown, and simple population growth functions may not adequately describe the population history of interest. Non-parametric coalescent methods provide greater flexibility by estimating the population size as a function of time directly from the sequence data and can be used for data exploration to guide the choice of parametric population models for further analysis. These methods first cut the time tree into segments, then estimate the population size of each segment separately according to the coalescent intervals within it.

The main differences among these methods are (i) how the population size function is segmented along the tree, (ii) the statistical estimation technique employed and (iii) in Bayesian methods, the form of the prior density on the parameters governing the population size function. In the 'classic skyline plot' (Pybus et al., 2000)

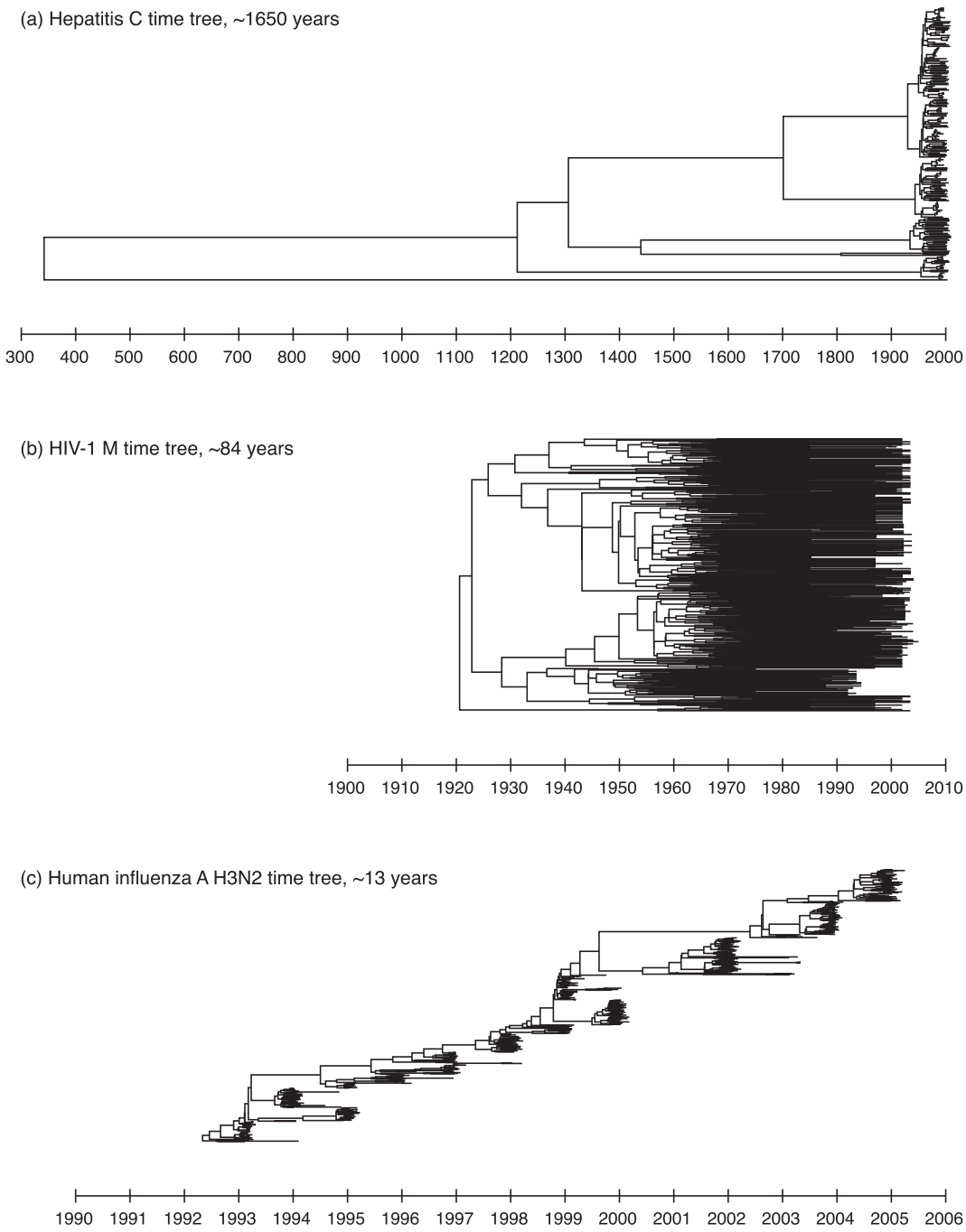

Fig. 2. Three time-trees estimated using BEAST. Notice the different orders of magnitude of time spanned and the different proportion of the tree spanned by samples. (a) A phylogeny of Hepatitis C spanning all major genotypes: the sampling interval spans 32 years [1977,2009] but represents a very small fraction of the estimated root height ($\approx 0.019 t_{MRCA}$), and this root height estimate could be severely underestimated and very misleading. (b) A phylogeny of HIV-1 M group: the sampling interval spans 27 years [1978,2005] and represents a significant fraction ($\approx 0.32 t_{MRCA}$) of the overall tree height, but still small enough that the estimated root should be viewed with caution. (c) A phylogeny of human Influenza A subtype H3N2: the sampling interval spans 12.2 years [1993.1,2005.3] and represents almost the full height of the tree ($\approx 0.94 t_{MRCA}$), and all divergence times are likely to be quite accurately estimated, since interpolation between many known sample times is inherently less error prone than extrapolation to ancient divergence times.

each coalescent interval is treated as a separate segment, so a tree of n taxa has $n-1$ population size parameters. However, the true number of population size changes is likely to be substantially fewer, and the generalized skyline plot (Strimmer and Pybus, 2001) acknowledges this by grouping the intervals according to the small-sample Akaike information criterion (AIC_c) (Burnham and Anderson, 2002). The epidemic history of HIV-2 was investigated using the generalized skyline plot (Strimmer and Pybus, 2001), indicating the population size was relatively constant in the early history of HIV-2 subtype A in Guinea-Bissau, before expanding more recently (Lemey et al., 2003). Using this information, the authors then employed a piecewise expansion growth model, to estimate the time of expansion to a range of 1955–1970.

While the generalized skyline plot is a good tool for data exploration, and to assist in model selection (e.g., Pybus et al., 2003; Lemey et al., 2004), it infers demographic history based on a single input tree and therefore does not account for sampling error produced by phylogenetic reconstruction nor for the intrinsic stochasticity of the coalescent process. This shortcoming is overcome by implementing the skyline plot method in a Bayesian statistical

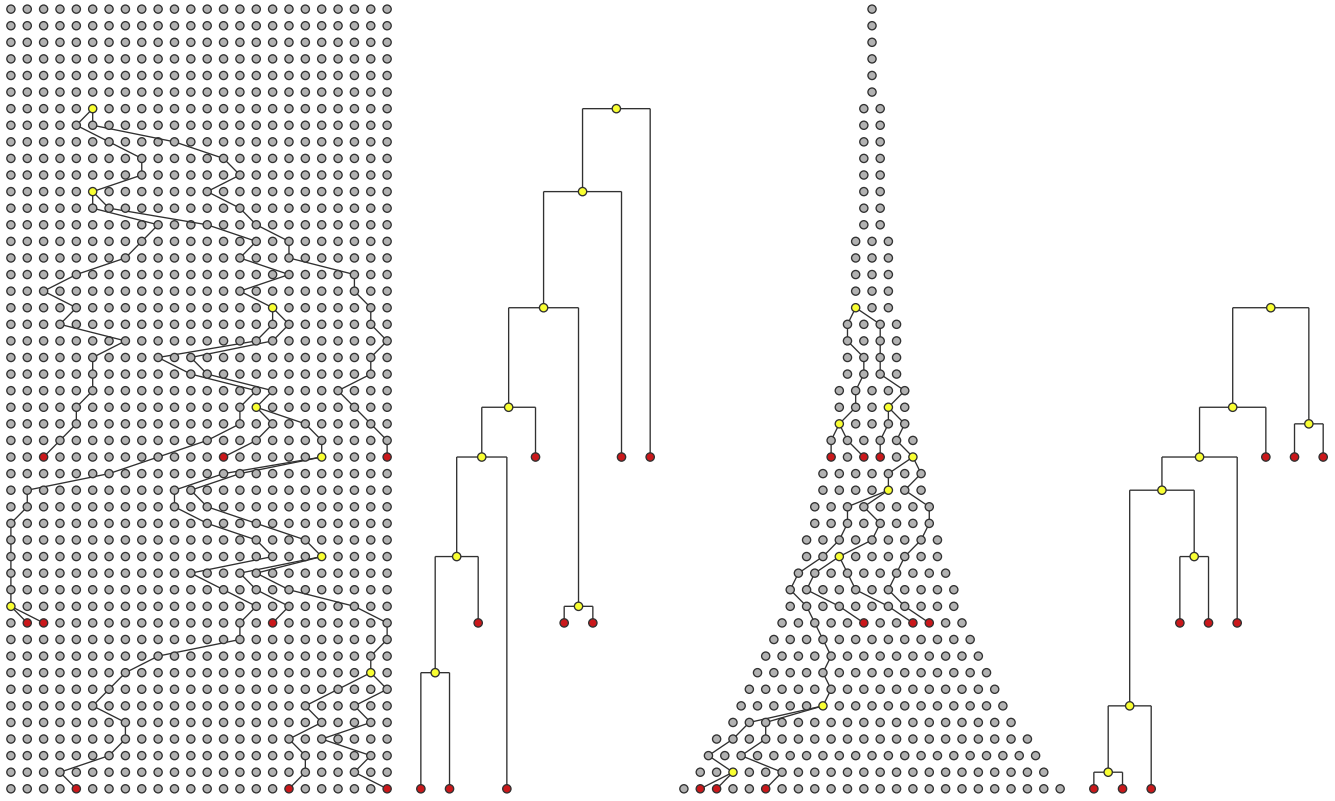

Fig. 3. The underlying Wright–Fisher population and serially-sampled genealogies from two populations. The first population has a constant population size over the history of the genealogy, while the second population has been exponentially growing. The coalescent likelihood calculates the probability of a genealogy given a particular background population history (e.g., constant or exponentially growing) and can therefore be employed to estimate the population history that best reflects the shape of the co-estimated phylogeny.

framework, which simultaneously infers the sample genealogy, the substitution parameters and the population size history. Further extensions of the generalized skyline plot include modeling the population size by a piecewise-linear function instead of a piecewise-constant population, allowing continuous changes over time rather than sudden jumps. The Bayesian skyline plot (Drummond et al., 2005) has been used to suggest that the effective population size of HIV-1 group M may have grown at a relatively slower rate in the first half of the twentieth century, followed by much faster growth (Worobey et al., 2008). On a much shorter time scale, the Bayesian skyline plot analysis of a dataset collected from a pair of HIV-1 donor and recipient was used to reveal a substantial loss of genetic diversity following virus transmission (Edwards et al., 2006). Further analysis with a constant-logistic growth model estimated that more than 99% of the genetic diversity of HIV-1 present in the donor is lost during horizontal transmission. This has important implications as the process underlying the bottleneck determines the viral fitness in the recipient host.

One disadvantage of the Bayesian skyline plot is that the number of changes in the population size has to be specified by the user *a priori* and the appropriate number is seldom known. One solution is provided by methods that perform Bayesian model averaging on the demographic model utilizing either Reversible jump MCMC (Opgen-Rhein et al., 2005) or Bayesian variable selection (Heled and Drummond, 2008), and in which case the number of population size changes is a random variable estimated as part of the model.

The methods for demographic inference discussed so far assume no subdivision within the population of interest. Like changes in the size, population structure can also have an effect on the pattern of the coalescent interval sizes, and thus the reliability of results can be questioned when population structure exists (Pybus et al., 2009). In the next section we will discuss approaches to phylogeographic inference, including coalescent approaches to population structure.

2.4. Statistical phylogeography and coalescence in structured populations

Phylogeography is a field that studies the evolution and dispersal process that has given rise to the observed spatial distribution of population or taxa. Phylogeographic methods can be divided into two approaches. The first performs post-tree-reconstruction analysis to answer phylogeographic questions, while the second jointly estimates the phylogeny and phylogeographic parameters of interest. When treating geographic location as discrete states, the former approach has been popular in the past couple of decades. It has the advantage of being less computationally intensive, but the outcome of the analysis depends on the input tree. Due to its simplicity, the most popular method for inferring ancestral locations has been maximum parsimony (Slatkin and Maddison, 1989; Swofford, 2003; Maddison and Maddison, 2005; Wallace et al., 2007), however this method does not allow for any probabilistic assessment of the uncertainty associated with the reconstruction of ancestral locations.

2.4.1. Mugration models

A *mugration model* is a mutation model used to analyze a migration process. A recent study of Influenza A H5N1 virus introduced a fully probabilistic 'mugration' approach by modeling the process of geographic movement of viral lineages via a continuous time Markov process where the state space consists of the locations from which the sequences have been sampled (Lemey et al., 2009b). This

facilitates the estimation of migration rates between pairs of locations. Furthermore, the method estimates ancestral locations for internal nodes in the tree and employs Bayesian variable selection (BVS) to infer the dominant migration routes and provide model averaging over uncertainty in the connectivity between different locations (or host populations). This method has helped with the investigation of the Influenza A H5N1 origin and the paths of its global spread, and also the reconstruction of the initial spread of the novel H1N1 human Influenza A pandemic (Lemey et al., 2009b). However, a shared limitation of models for discrete location states is that ancestral locations are limited to sampled locations. As demonstrated by the analysis of the data set on rabies in dogs in West and Central Africa, absence of sequences sampled close to the root can hinder the accurate estimation of viral geographic origins (Lemey et al., 2009b). Phylogeographic estimation is therefore improved by increasing both the spatial density and the temporal depth of sampling. However, dense geographic sampling leads to large phylogenies and computationally intensive analyses.

2.4.2. The structured coalescent

The *structured coalescent* (Hudson, 1990) can also be employed to study phylogeography. The structured coalescent has also been extended to heterochronous data (Ewing et al., 2004), thus allowing the estimation of migration rates between demes in calendar units. The *serial structured coalescent* was first applied to an HIV dataset with two demes to study the dynamics of subpopulations within a patient (Ewing et al., 2004), but the same type of inference can be made at the level of the host population. Further development of the model allowed for the number of demes to change over time (Ewing and Rodrigo, 2006a). MIGRATE (Beerli and Felsenstein, 2001) also employs the structured coalescent to estimate subpopulation sizes and migration rates in both Bayesian and maximum likelihood frameworks and has recently been used to investigate spatial characteristics of viral epidemics (Bedford et al., 2010). Additionally, some studies have focused on the effect of ghost demes (Beerli, 2004; Ewing and Rodrigo, 2006b), however no models explicitly incorporating population structure, heterochronous samples and nonparametric population size history are yet available.

One *ad hoc* solution involves modeling the migration process along the tree in a way that is conditionally independent of the population sizes estimated by the skyline plot (Lemey et al., 2009a). Thus, given the tree, the migration process is considered independent of the coalescent prior. However this approach does not capture the interaction between migration and coalescence that is implicit in the structured coalescent, since coalescence rates should depend on the population size of the deme the lineages are in. As we will see in the following section, statistical phylogeography is one area where the unification of phylogenetic and mathematical epidemiological models looks very promising.

2.4.3. Phylogeography in a spatial continuum

In some cases it is more appropriate to model the spatial aspect of the samples as a continuous variable. The phylogeography of wildlife host populations have often been modeled in a spatial continuum by using diffusion models, since viral spread and host movement tend to be poorly modeled by a small number of discrete demes. One example is the expansion of geographic range in eastern United States of the raccoon-specific rabies virus (Biek et al., 2007; Lemey et al., 2010). Brownian diffusion, via the comparative method (Felsenstein, 1985; Harvey and Pagel, 1991), has also been utilized to model the phylogeography of Feline Immunodeficiency Virus collected from the cougar (*Puma concolor*) population around western Montana. The resulting phylogeographic reconstruction was used as proxy for the host demographic history and population structure, due to the predominantly vertical transmission of the virus (Biek et al., 2006). However, one of the assumptions of Brownian diffusion is rate homogeneity on all branches. This assumption can be relaxed by extending the concept of relaxed clock models to the diffusion process (Lemey et al., 2010). Simulations show that the relaxed diffusion model has better coverage and statistical efficiency over Brownian diffusion when the underlying process of spatial movement resembles an over-dispersed random walk.

Like their *mugration model* counterparts, these models ignore the interaction of population density and geographic spread in shaping the sample genealogy. However there has been progress in the development of mathematical theory that extends the coalescent framework to a spatial continuum (Barton et al., 2002, 2010a,b), although no methods have yet been developed providing inference under these models.

Box 1: The anatomy of a Bayesian coalescent analysis using MCMC

Bayesian phylogenetic inference by Markov Chain Monte Carlo (MCMC) (Yang and Rannala, 1997; Mau et al., 1999) involves the simulation of the joint posterior distribution of substitution model parameters (ϕ) and the phylogenetic tree given the sequence data (D). By restricting the phylogenetic model to time-trees (see Fig. 1) and coupling the phylogenetic likelihood with a coalescent prior, the parameters (θ) of the population history, $N_\theta(t)$, can also be estimated simultaneously by sampling from the posterior probability distribution (Drummond et al., 2002):

$$f_{\theta g \phi}(\theta, g, \phi|D) = \frac{1}{\Pr(D)} \Pr(D|g,\phi) f_G(g|\theta) f_\Theta(\theta) f_\Phi(\phi). \quad (2)$$

The term $\Pr(D|g,\phi)$ is often referred to as the phylogenetic likelihood, and is the probability of the data given the time-tree g and substitution model parameters. It can be computed by the pruning algorithm (Felsenstein, 1981), which efficiently sums over all ancestral sequence states at the internal nodes of the tree. An extension of the likelihood accommodates heterogeneity across sites (Yang, 1994). If the time-tree g relates a heterochronous sample of sequences, then the substitution parameters ϕ also includes the overall substitution rate μ, and this can be estimated from the heterochronous data, so that the population history is estimated on a calendar scale. The normalizing constant $\Pr(D)$ is also known as the partition function or marginal likelihood and its magnitude provides a measure of model support, although its estimation requires advanced MCMC techniques (e.g., thermodynamic integration or transdimensional MCMC).

Coalescent models come into play when determining the prior density for the time-tree topology and coalescent/divergence times. The coalescent provides a probability distribution, $f_G(g|\theta)$, conditional on a deterministic model of population size history, $N_\theta(t)$. Its parameters (θ) can in turn be estimated as hyperparameters. Given a time-tree $g = \{E_g, \mathbf{t}\}$ of n contemporaneous samples composed of an edge graph E_g and coalescent times $\mathbf{t} = \{t_n = 0, t_{n-1}, \ldots, t_2, t_1\}$ the coalescent density is:

$$f_G(g|\theta) = \prod_{i=1}^{n-1} \left[N_\theta(t_i)^{-1} \exp\left(- \int_{t_i}^{t_{i+1}} \frac{\binom{n}{2}}{N_\theta(t)} dt \right) \right]. \quad (3)$$

The prior distributions $f_\Theta(\theta)$ and $f_\Phi(\phi)$ are usually selected from standard univariate or multivariate distributions.

3. Evolutionary models combining epidemiological and genomic data

In the previous section we have seen that phylogenetics can be used to infer the date of an outbreak, its source population and the viral transmission history, directly from time-stamped genomic data. Whereas phylogenetic models mainly address questions about evolutionary history, dynamical models are often used to make predictions about the future. Predictive models are important because they provide the possibility of anticipating certain aspects of the outcome of emerging epidemics and assessing the risk of pandemics, and the potential effects of planned intervention.

Phylogenetic inference is based on genetic data such as sampled DNA sequences from infected hosts. Current models using such data to infer information about the past often require simplifying assumptions about the population size e.g., to be constant or to be subject to pure exponential growth. Epidemiologists, on the other hand, fit their models to prevalence or incidence data. Standard epidemiological models are described by sets of ordinary differential equations tracking the (often non-linear) changes in numbers of susceptible and infected individuals. Consequently, the simple prior assumptions for the population sizes (of infected individuals) used in phylogenetics appear inadequate from an ecological perspective.

Epidemiological models play a major role in deciding which measures of disease control are taken to avoid or stop viral outbreaks. The effects of isolation, vaccination and other measures are estimated through model simulations, serving as a basis for decisions on which public health policies to institute and actions to take. However, knowledge of the phylogenetic history of viral outbreaks can be vital in reconstructing transmission pathways which contributes to effective management and future prevention efforts (e.g., Cottam et al., 2008).

The epidemiological and ecological processes determining the diversity of fast evolving RNA viruses act on the same time scale as that on which mutations arise and are fixed in the population (Holmes, 2004). This implies that genetic sequence data can provide independent evidence on transmission histories. Whereas epidemiological data typically provides information about who was infected and when, it generally does not provide positive evidence about transmission history. Thus the combination of these sources of information should open the way to more detailed epidemiological inference, including Bayesian estimation of contact networks and transmission histories (Welch et al., 2011).

3.1. Standard epidemiological models and their stochastic analogues

Standard epidemiological models are based on flux between host compartments dividing the host population e.g., into susceptible (S), infected (I) and recovered or removed (R) individuals. Standard models are termed SI, SIS and SIR. The choice of model is based on the characteristics of the considered disease, the existence of a latent period, immunity after infection *et cetera* (see Box 2) (Anderson and May, 1991; Keeling and Rohani, 2008). Restricting the focus to the time evolution of the number of individuals in each compartment, these models grasp the overall progress of an epidemic. Certain disease characteristics require adaptations or extensions of standard models, for example, the inclusion of asymptomatic infections that account for a sampling bias towards symptomatic infections in case the virus of interest does not always cause noticeable symptoms (e.g., Aguas et al., 2008). An important threshold ratio is the basic reproduction ratio R_0, the expected number of secondary infections caused by one primary infection in a completely susceptible population (Diekmann et al., 1990). Based on its value epidemiologists make predictions on the effect of the disease. In classical deterministic epidemiological models, *if the basic reproduction ratio is larger than one, an epidemic is expected*.

Box 2: **Compartmental models for infectious diseases** (Keeling and Rohani, 2008)

Let S, E, I and R be the fractions of **s**usceptible, **e**xposed, **i**nfected and **r**ecovered/**r**emoved individuals in the host population. The left hand side of each equation block gives the model equations, the right hand side the (non-trivial) endemic equilibria, which are only obtainable for $R_0 > 1$. The basic reproduction ratio R_0 depends on the corresponding model. Apart from the SI model, the overall population is assumed to be constant, such that the sum of fractions for each model equals one. Under the assumption of homogeneous mixing in the population the transmission term $\beta S\, I$ can be derived, which determines the total rate of new infections.

SI model. Fatal infections, eventually killing the infected, can be modeled with only two compartments: susceptible and infected. Assume a fixed birth rate ν and death rate μ. The endemic equilibrium (S^*,I^*) is obtainable for $R_0 = \frac{\beta}{\mu+\gamma} > 1$.

$$\dot{S} = \nu - \beta SI - \mu S \qquad S^* = \tfrac{\nu}{\beta-\gamma}$$
$$\dot{I} = \beta SI - (\gamma+\mu)I \qquad I^* = \tfrac{\nu(\beta-\gamma-\mu)}{(\beta-\gamma)(\gamma+\mu)}$$

SIR model. Transmission of the disease to susceptibles leads to a period of illness until recovery, which in turn implies immunity. Demography is described by the birth and death rate μ and recovery is obtained at rate γ; its reciprocal $1/\gamma$ is the mean infectious period. Here, $R_0 = \frac{\beta}{\mu+\gamma}$. The last equation is redundant since $S + I + R = 1$.

$$\dot{S} = \mu - \beta SI - \mu S \qquad S^* = \tfrac{1}{R_0}$$
$$\dot{I} = \beta SI - \gamma I - \mu I \qquad I^* = \tfrac{\mu}{\beta}(R_0 - 1)$$
$$\dot{R} = \gamma I - \mu R \qquad R^* = 1 - S^* - I^*$$

SIS model. Recovery from infection does not imply immunity. Instead, after infection the individuals go back to the susceptible stage. Therefore, the disease can persist even without including newborns in the population. Ignoring demography, the dynamics are characterized by coupled differential equations $\dot{S} = \gamma I - \beta SI$ and $\dot{I} = \beta S\, I - \gamma I$. Since $S = 1 - I$, they can be replaced by one equation.

$$\dot{I} = (\beta - \beta I - \gamma)I \qquad S^* = \tfrac{1}{R_0}, \quad I^* = 1 - \tfrac{1}{R_0}$$

SEIR model. In order to account for a latent period with assumed average duration $1/\sigma$, the SIR model can be extended by including exposed individuals composing a fraction E of the population. Exposed individuals are infected, but not yet infectious. The differential equations for S (and R) are as in the SIR model. Dynamics in E and I are described as follows.

$$\dot{E} = \beta SI - (\mu+\sigma)E \qquad S^* = \tfrac{1}{R_0}, \text{ with } R_0 = \tfrac{\beta\sigma}{(\mu+\gamma)(\mu+\sigma)}$$
$$\dot{I} = \sigma E - (\gamma+\mu)I \qquad E^* = \tfrac{\mu(\mu+\gamma)}{\beta\gamma}(R_0 - 1)$$
$$I^* = \tfrac{\mu}{\beta}(R_0 - 1), \quad R^* = 1 - S^* - E^* - I^*$$

Further models are SIRS, SEIS, MSIR, MSEIR, MSEIRS, etc., where M denotes passively immune infants, allowing for diseases where an individual can be born with a passive immunity from its mother.

Typically, epidemiologists fit a suitable set of deterministic differential equations to empirical data, often the number of infections or related hospitalizations in a population. Consequently, the model can be used to estimate if an epidemic can be kept under control by measures such as (i) vaccination and (ii) antiviral prophylaxis for susceptible individuals, (iii) treatment of infected individuals or (iv) isolation of infected individuals from susceptible individuals. Decisions on public health policies are often based on these estimates.

The simplest epidemiological models assume homogeneous mixing within a population. In many cases this assumption is not valid. Due to host contact dynamics viral infections spread easily within social units such as schools, cities and farms, less so among them. Integration of population structure is therefore essential. However, even within subpopulations individual dynamics might differ stochastically (see Fig. 4). Such randomness can be accounted for by considering stochastic models (see e.g., survey by Britton, 2010).

3.1.1. Stochastic models

Before introducing stochastic compartmental models thoroughly, we illustrate them based on a stochastic SIR model simulation. We simulate the spread of a virus strain in a population divided into n subpopulations which are connected by comparatively rare migration events. Let $\mathcal{L} = \{0, \ldots, n-1\}$ denote the set of locations. A single infected individual initiates the epidemic in one of the n completely susceptible populations. After an exponentially distributed waiting time one of the following events happens:

- Infection at mass action infection rate β.
- Migration at migration rate m_{ik} for $i \neq k \in \mathcal{L}$.
- Recovery of an infected individual at recovery rate γ.
- Birth of a susceptible individual at rate μ.
- Death of an individual at rate μ.

Fig. 5 shows a realization of the simulated dynamics for $n = 3$ populations. The epidemic starts in population 1 (blue) and many individuals get infected before the first individuals in population 2 (yellow) and eventually population 3 (red) get infected.

Let S_k, I_k and R_k be the fractions of individuals in each subpopulation $k \in \mathcal{L}$. The sum $S_k + I_k + R_k$ equals one for every $k \in \mathcal{L}$. The deterministic analogue of our model can be described with the following differential equations:

$$\dot{S}_k = \mu - \beta S_k I_k - \mu S_k + \sum_{l \neq k} (m_{lk} S_l - m_{kl} S_k),$$
$$\dot{I}_k = \beta S_k I_k - \gamma I_k - \mu I_k + \sum_{l \neq k} (m_{lk} I_l - m_{kl} I_k),$$
$$\dot{R}_k = \gamma I_k - \mu R_k + \sum_{l \neq k} (m_{lk} R_l - m_{kl} R_k), \quad k, l \in \mathcal{L}.$$

However it is important to realize that this set of differential equations cannot capture all of the behaviors of its stochastic counterpart. In fact, starting from a deterministic representation like this, there are multiple stochastic Markov processes that exhibit the same deterministic limit, but can potentially have exponentially different behavior in their stochastic properties, such as the time to extinction (e.g., Drummond et al., 2010).

Formally, two distinct sources of variance can be considered in stochastic models of populations (Engen et al., 1998). The first is *environmental stochasticity* and is often modeled by admitting temporal variation in the parameters of the population model. The second is *demographic stochasticity* and describes the stochasticity of fluctuations in populations of finite size due to the inherent unpredictability of individual outcomes.

To model demographic stochasticity (also known as internal stochasticity; Chen and Bokka, 2005) in the absence of environmental (external) stochasticity, the time-evolution of an epidemic can be represented by a jump process and its corresponding master equation (Gardiner, 2009). The master equation describes the time evolution of the probability distribution over the discrete state space. For the closed SIR model (Kermack and McKendrick, 1927) the master equation for the numbers of individuals in each of the three compartments (n_S, n_I, n_R) is:

$$\begin{aligned}\dot{P}_{n_S,n_I,n_R}(t) = {} & \beta(n_S+1)(n_I-1)P_{n_S+1,n_I-1,n_R}(t) \\ & + \gamma(n_I+1)P_{n_S,n_I+1,n_R-1}(t) \\ & - (\beta n_S n_I + \gamma n_I)P_{n_S,n_I,n_R}(t)\end{aligned} \tag{4}$$

A single realization of this epidemic jump process is described by a sequence of timed transition events (individual infection or recovery events). In the closed SIR model, the waiting or sojourn time between a pair of sequential events is exponentially distributed (i.e., the transition process is memoryless), and thus the process is a continuous-time Markov process.

Stochastic models of this form can also be viewed in terms of their reaction kinetics. For the closed stochastic SIR model above the two 'reactions' are infection and recovery:

$$\begin{aligned} I + S &\xrightarrow{\beta} 2I \\ I &\xrightarrow{\gamma} R, \end{aligned} \tag{5}$$

indicating that a susceptible contacts an infectious individual and gets infected at reaction rate β whereas an infected recovers at reaction rate γ. More precisely, the time (τ) an individual spends in the susceptible and infected compartments are exponentially distributed with rates βI and γ, respectively. It is the binary infection reaction that leads to the non-linear dynamics of the system.

For stochastic models $R_0 > 1$ does not necessarily imply an outbreak of the disease. Instead, a higher basic reproduction ratio suggests a higher probability of an outbreak, but the precise

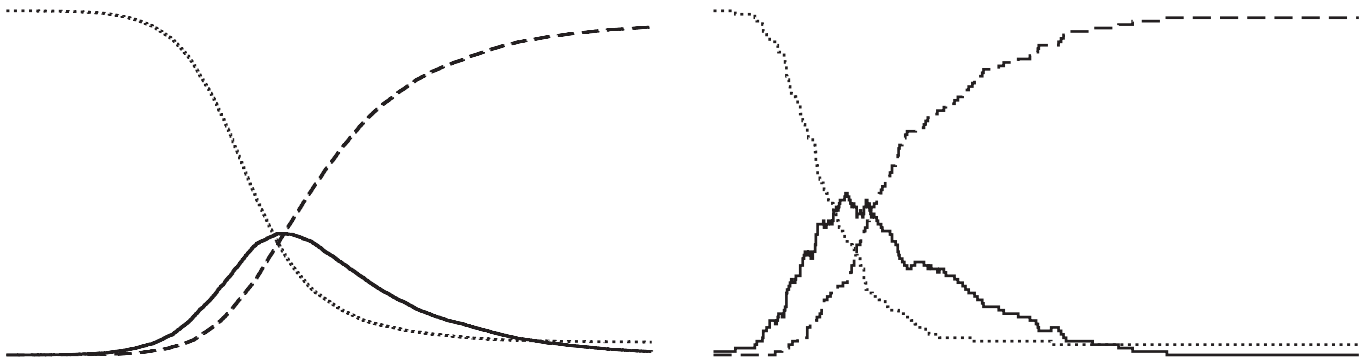

Fig. 4. Realization of an SIR model showing the dynamics of susceptible (dotted line), infected (solid line) and recovered (dashed line) individuals over time. The stochastic version leads to less "smooth" dynamics (right hand side).

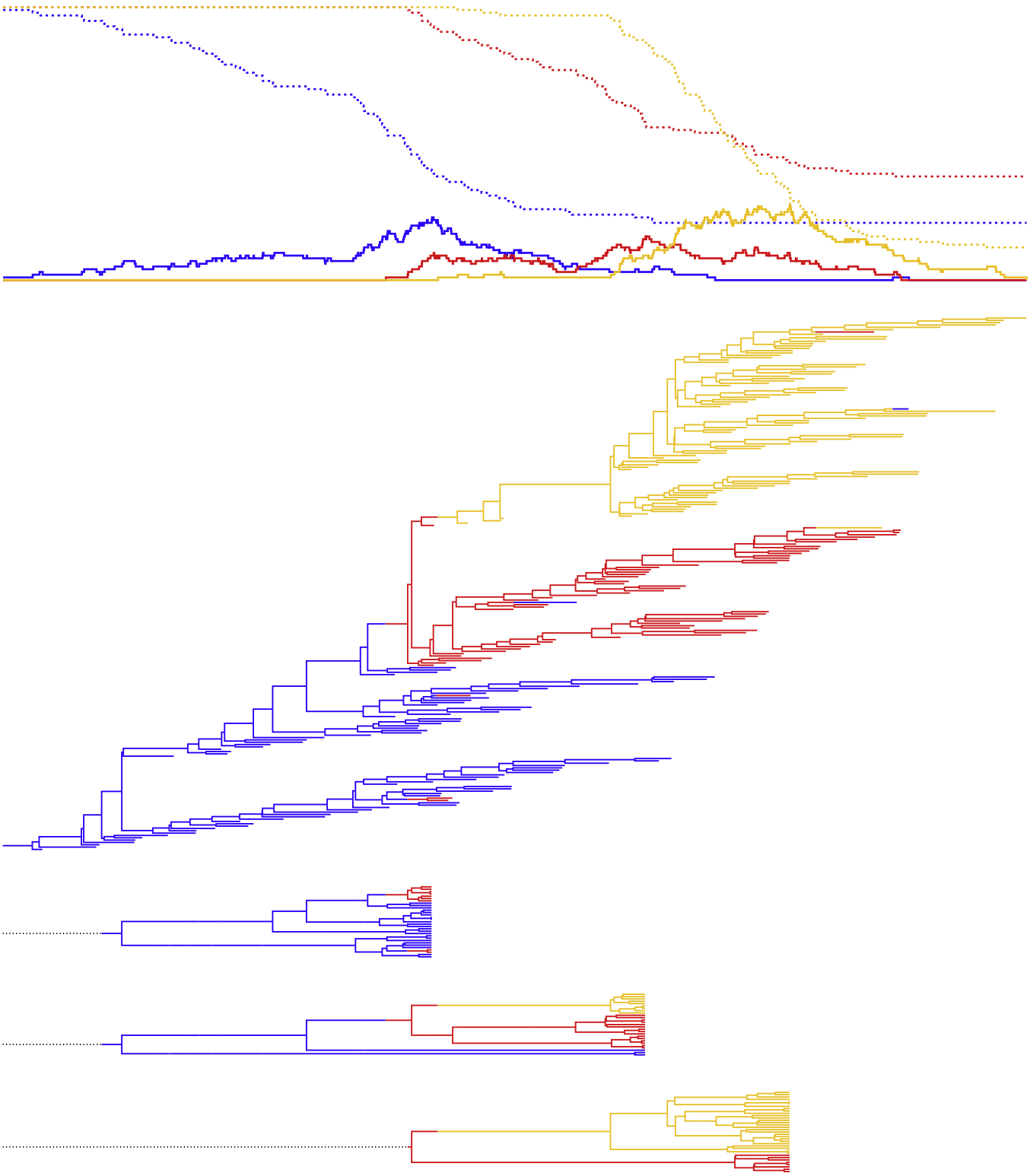

Fig. 5. Realization of a stochastic SIR model in a structured population. Simulated viral dynamics in $n = 3$ subpopulations, each color denoting one of them. The numbers of susceptibles (dotted lines) are plotted against the numbers of infected individuals (solid lines), a full infection tree and three sample trees at different times throughout the epidemic.

relationship depends on the specific model considered and the initial condition.

Algorithms have been developed that allow exact and approximate simulation of coupled reactions such as the closed SIR (Bartlett, 1957; Gillespie, 1976, 2001). Fig. 6 shows simulated viral outbreaks under a stochastic SIR and SIS model with $R_0 \approx 2.3$ in a population divided into three distinct subpopulations. Note that there is no outbreak in (3) although $R_0 > 1$.

3.1.2. Relating epidemic models to genealogies

Deterministic epidemic models can be derived from the underlying jump process, and can represent useful macroscopic laws of

motion in the appropriate limit. However such approaches are not adequate for modeling systems in which small numbers of individuals are frequently involved. For a similar reason, it is awkward to reconcile large-limit deterministic models with the small sample genealogies that are obtained with molecular phylogenetic approaches. Therefore, stochastic continuous-time discrete-state formulations of epidemic models may be more suited to forming connections between the two disciplines. The forward simulations of a stochastic epidemic model introduced with Fig. 5 demonstrate the relationship between epidemic models and genealogies. Knowing the exact parameters and resulting dynamics throughout the simulated outbreak, we can build a full transmission history for the outbreak (which is not unique given only the time evolution of the number of infected individuals, since at each event the infected individuals involved are chosen randomly). An infection event in the forward simulation corresponds to a bifurcation in the transmission tree. Restricting the full tree to a "sample genealogy" that only includes the individuals that were infectious at a specific sampling time yields very different results for different times during the outbreak, which underlines the importance of sampling methods (see e.g., Stack et al., 2010).

As we can see in the simulations, virus transmission often depends on spatial structure. The interaction among humans living in the same city, for example, differs from among-city interaction, which is important whenever viral transmission exceeds city borders. There are many other social and spatial units this concept applies to: households, schools, or on a larger scale, regions, countries and continents. In fact, most phylogenetic and epidemiological studies model the dynamics of spatially distributed systems, albeit many of them ignore spatial structure for the sake of simplicity. Durrett and Levin demonstrate that models ignoring spatial structure yield qualitatively different results than spatial models (Durrett and Levin, 1994).

3.2. Phylogenetic epidemiology and phylodynamics

Phylodynamics is a term used to describe a synthetic approach to the study of rapidly evolving infectious agents that considers the action (and interaction) of both evolutionary and ecological processes. The term phylodynamics was introduced by Grenfell et al. (2004) to describe the "melding of immunodynamics, epidemiology, and evolutionary biology" that is required to analyse the interacting evolutionary and ecological processes especially of rapidly evolving viruses for which both processes have the same time scale.

Two distinct pursuits have been labeled phylodynamics by recent studies. The first relies on the idea that ecological processes and population dynamics can effectively be tracked by neutral genetic variation, such that past ecological and population events are "imprinted" in genetic variation within populations and can be reconstructed along with the reconstruction of evolutionary history. The idea is sound for truly neutral variation, but the compact genomes of rapidly evolving viruses are not simple recording devices. Instead they are packed with functional information and mutations play an active role in population and ecological processes through the action of Darwinian selection. Hence, the more challenging second phylodynamic pursuit is the analysis of the inevitable interaction of evolutionary and ecological processes that requires the joint analysis of both. We will call the former pursuit *phylogenetic epidemiology*, and reserve the term *phylodynamics* for approaches that aspire to model the interaction of ecological and evolutionary processes. The effect of novel mutations on population dynamics through their interaction with the immune system or anti-viral drugs are examples of phylodynamics in this stricter sense.

3.2.1. Phylogenetic epidemiology

The focus of many studies aspiring to combine population genetic and epidemiological approaches is the basic reproduction ratio R_0, estimates of which are used to develop containment strategies for emerging pandemics. Such estimates can be obtained from phylogenetic analysis, e.g., through estimating population growth rates (Pybus et al., 2001). Another popular way to infer population dynamic information from genomic data is the application of parametric and non-parametric coalescent models (Strimmer and Pybus, 2001; Drummond et al., 2005; Minin et al., 2008).

Phylogenetic methods can be used to estimate R_0, which can then be used to investigate transmission patterns and the number of generations of transmission. Depending on the distribution of the generation time (i.e., the duration of infectiousness) the relationship between R_0 and the growth rate r of the population can be used to compute the basic reproduction number (Wallinga and Lipsitch, 2007). Little is known about generation time distributions, the usual approach is to fit the epidemic models to the observed data. Wallinga and Lipsitch list the resulting equations for R_0 for exponential, normal, or delta distributions of generation time. They show that without knowledge of the generation time distribution an upper bound for the reproductive number can still be estimated. Others obtain R_0 estimates based on coalescent theory, as for example (Rodrigo et al., 1999) who estimated it in vivo for HIV-1.

In a recent study on the Influenza A (H1N1) outbreak in 2009 both epidemiological and Bayesian coalescent approaches for the computation of R_0 were applied (Fraser et al., 2009). Whereas the epidemic approaches gave estimates of 1.4–1.6 for R_0, the Bayesian coalescent approach yielded a posterior median of 1.22. All estimates are larger than one, correctly indicating that the virus spreads successfully, rather than dying out. However, an age-dependent heterogeneous epidemic model best fits the data and results in an estimate of $R_0 = 1.58$.

Structures determining host interaction are often modeled as contact networks (Welch et al., 2011). The transmission of foot and mouth disease virus is highly dependent on the interaction among farms and the detection of infected farms is essential. A plausible approach is to consider each farm as an individual in a contact network. Through phylogenetic analysis of consensus sequences (one sequence for each farm) contacts between farms can be traced in order to find infected but non-detected farms such that contacts between farms can be traced in order to find infected but non-detected farms (Cottam et al., 2008).

Changes in effective population size estimated through phylogenetic analyses can indicate past changes in population size. Therefore, many recent studies infer the demographic history of a virus using Bayesian skyline plot models (Drummond et al., 2005). For example, (Siebenga et al., 2010) are interested in the epidemic expansion of norovirus GII.4 which they investigate by reconstructing the changes in population structure using Bayesian skyline plots. Similarly, (Hughes et al., 2009) explore the heterosexual HIV epidemic in the UK. Analyses of the genomic and epidemiological dynamics of human Influenza A virus explore the sink-source theory and investigate the spatial connections of a seasonal global epidemic (Rambaut et al., 2008; Lemey et al., 2009b; Bedford et al., 2010).

Coalescent theory has also been adapted to fit an epidemic SIR model to sequence data (Volz et al., 2009). Frost and Volz (2010) provide an overview on how appropriate interpretation of coalescent rates differs among the different population dynamic approaches it is being used with. Interpretation of the coalescent-based skyline plots must be made with caution. As opposed to generation times referring to durations of infection in epidemiological theory, for coalescent approaches being applied to infectious diseases the generation times usually describe times between transmission events. Accordingly, although prevalence

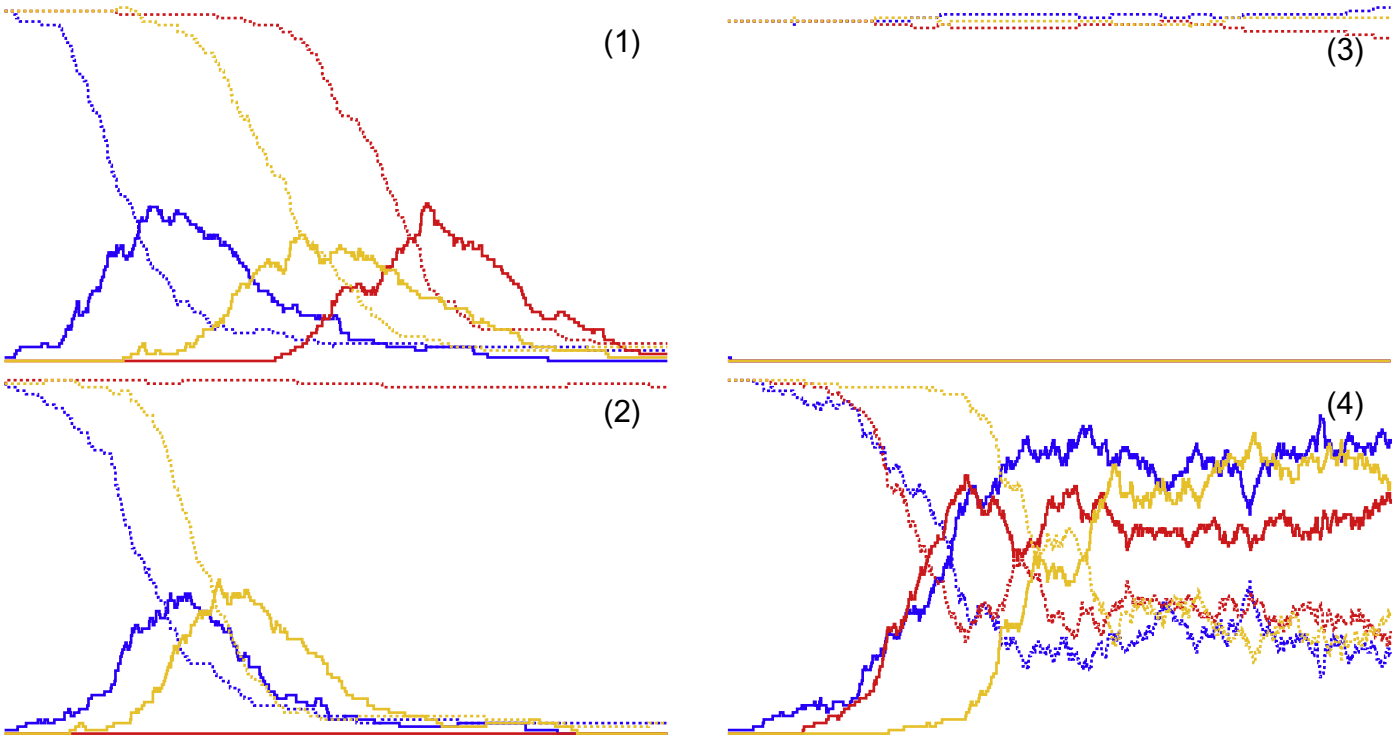

Fig. 6. Simulated viral outbreak under stochastic SIR (1–3) and SIS (4) model among three populations (denoted by blue, yellow and red curves). The initial condition is a single infected individual in the blue population. In (3) the disease does not break out (numbers of susceptibles in dotted lines and infected in solid lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

does affect phylogenetic reconstruction through sampling, the population dynamic patterns are mainly determined by incidence (Frost and Volz, 2010).

3.2.2. Phylodynamics sensu stricto

One early attempt to integrate dynamical and population genetic models used coupled differential equations and Markov chain theory to model the within-host time evolution of viral genetic diversity under basic dynamic models of a persistent infection (Kelly et al., 2003). The main focus was the impact of the dynamical model on the variance in the number of replication cycles, as this is a key determinant of the rate of genetic divergence and thus potential for adaptation. Interestingly, the model reveals that multiple cell type infections can decrease viral evolutionary rates and increase the likelihood of persistent infection.

Genetic diversity within hosts is closely related to between host dynamics: Gordo and Campos (2007) develop structured population genetic models, explicitly incorporating epidemiological parameters to analyze the relationship between genetic variability and epidemiological factors. A simple SIS model is simulated based on two different models of host contact structure, the island model and a scale free contact network. For low clearance rates and low intrahost effective population size, levels of genetic variability turn out to be maximal when transmission levels are intermediate, independent of the host population structure. In a scale free contact network the population consists of many low-connectivity hosts and very few high-connectivity hosts, a common pattern for sexually transmitted diseases (e.g., Lloyd and May, 2001; Liljeros et al., 2001). In this setting genetic variation appears to be lower in highly connected than in weakly connected hosts. With their study Gordo and Campos (2007) underline that an integration of population genetics and epidemiology can have important implications for public health policies.

In a deterministic framework Day and Gandon (2007) model the interaction of evolutionary and ecological processes by coupling SIS host dynamics with viral evolution. The interaction of evolution and ecology is incorporated through the fitness of each virus strain. For strain i they define a fitness $r_i = \beta_i n_S - \mu - \upsilon_i - \gamma$, where β_i is the strain-specific transmission rate per susceptible, υ_i is the strain-specific virulence (determining the increase in mortality rate due to infection), μ is the baseline mortality rate and γ is the recovery rate. The evolutionary dynamics of strain frequencies are tracked quantitatively and the evolutionary dynamics of strain frequencies are intimately linked with the overall infection dynamics of the host population via the strain-specific virulence and transmission rates. Their analysis provides insight into the mechanistic laws of motion connecting genetic evolution with the evolution of virulence and transmission rates.

An exceptional feature of Influenza viruses is the limited genetic diversity which appears to contradict the viruses' high mutation rate. Integrating single virus strain features and host immunity into a stochastic transmission model Ferguson et al. (2003) search an explanation for this. Although epidemiological factors play a role in limiting Influenza diversity, strain-transcendent immunity must be relevant as well.

Through a phylodynamic analysis of interpandemic Influenza in humans Koelle et al. (2006) underline the importance of the viral structure for antigenicity and the immune recognition dynamics of Influenza epitopes. They consider clusters that contain strains with similar conformations of HA epitopes such that there is high cross-immunity of strains within each cluster. A genotype–phenotype model that implements neutral networks (the clusters) is coupled with an epidemiological transmission model in which the number of susceptible, infected and recovered individuals in each cluster are modeled. Model simulations result in time series of infected cases that agree with the typical annual outbreaks in temperate regions and empirical dominance of certain antigenic clusters. According to this model, years in which a formerly dominant cluster is replaced by a new one have the highest numbers of infections. In the following year there are particularly few infec-

tions, presumably due to higher host immunity caused by the previous year's outbreak. Thereafter follow "average" years until the next cluster-transition occurs, i.e., until another cluster becomes dominant again.

Another natural explanation of the contradiction between high mutation rates and constant genetic diversity is the fixation of many deleterious mutations that leads to the extinction of the respective strains. Recent population genetic models account for population dynamics e.g., in order to enhance the understanding of allele fixation processes and the importance of demographic stochasticity (Parsons and Quince, 2007; Champagnat and Lambert, 2007; Parsons et al., 2010).

Structured models do not only allow for more realistic dynamics, they can also bridge the gap to phylogenetic/-geographic methods since most of them are sample-based, ideally, with each sample representing one infected individual. Modeling coupled host–virus dynamics Welch et al. (2005) embed an epidemic population model into a branching and coalescent structure, producing a scaled coalescent process that describes the inter-host dynamics given a virus sample genealogy. Their simulations show that, for large sample sizes, the model provides accurate estimates of the contact rate and the selection parameter.

Overall, phylodynamic methods have been developed and proven useful for the analysis of various viruses. However, phylogenetic reconstruction is still quite restricted by coalescent assumptions. An alternative to the coalescent for cases in which sample sizes are big compared to the overall population is the birth–death with incomplete-sampling model (Gernhard, 2008; Stadler, 2009), and this framework has recently been extended to include heterochronous data (Stadler, 2010), opening the way for an alternative approach to phylodynamic inference from time-stamped virus data.

4. Outlook

Bayesian phylogenetic inference has led to an explosion of analyses of rapidly evolving viruses in recent years. While this explosion has been fruitful in elucidating the manifold variation in origin, transmission routes and evolutionary rates underlying the present diversity of infection agents, there is a nascent field that promises to extend the conceptual reach of molecular sequence data, through a unification of phylogenetics and mathematical epidemiology. This new field of phylodynamics encompasses both inference of classical epidemiological parameters using phylogenetics as well as exciting new approaches that aim to investigate the consequences of the inevitable interaction between evolutionary (mutation, drift, Darwinian selection) and ecological (population dynamics and ecological stochasticity) processes. The research being pursued has broader consequences for evolutionary biology and molecular ecology. This interaction of evolution and ecology will occur whenever a population contains genotypes with different intrinsic dynamical properties (e.g., virulence, transmission rates, recovery rates). Whereas this condition is almost always met in real populations and frequently definitive in its role in shaping outcomes, the mathematical and theoretically analysis of Darwinian selection within epidemiological models is the most challenging and least studied area within the emerging field of phylodynamics. It is thus ripe for future research.

In the meantime, it is likely that phylodynamic research will rapidly develop new methods for statistical phylogeography and structured population dynamics.

References

Aguas, R., White, L.J., Snow, R.W., Gomes, M.G.M., 2008. Prospects for malaria eradication in sub-Saharan Africa. Plos One 3.

Anderson, R.M., May, R.M., 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Aris-Brosou, S., Yang, Z., 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Systematic Biology 51, 703–714.

Bartlett, M., 1957. Measles periodicity and community size. Journal of the Royal Statistical Society. Series A (General) 120, 48–70.

Barton, N.H., Depaulis, F., Etheridge, A.M., 2002. Neutral evolution in spatially continuous populations. Theoretical Population Biology 61, 31–48.

Barton, N.H., Etheridge, A.M., Veber, A., 2010a. A new model for evolution in a spatial continuum. Electronic Journal of Probability 15, 162–216.

Barton, N.H., Kelleher, J., Etheridge, A.M., 2010b. A new model for extinction and recolonization in two: dimensions quantifying phylogeography. Evolution 64, 2701–2715.

Bedford, T., Cobey, S., Beerli, P., Pascual, M., 2010. Global migration dynamics underlie evolution and persistence of human influenza a (h3n2). PLoS Pathogens 6, e1000918.

Beerli, P., 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. Molecular Ecology 13, 827–836.

Beerli, P., Felsenstein, J., 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences of the United States of America 98, 4563–4568.

Biek, R., Drummond, A.J., Poss, M., 2006. A virus reveals population structure and recent demographic history of its carnivore host. Science 311, 538–541.

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. Proceedings of the National Academy of Sciences 104, 7993–7998.

Bloom, J.D., Raval, A., Wilke, C.O., 2007. Thermodynamics of neutral protein evolution. Genetics 175, 255–266.

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

Britton, T., 2010. Stochastic epidemic models: a survey. Mathematical Biosciences 225, 24–35.

Burnham, K., Anderson, D., 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer, Verlag.

Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J., Fitch, W.M., 1999. Predicting the evolution of human influenza a. Science 286, 1921–1925.

Cartwright, R.A., Lartillot, N., Thorne, J.L., 2011. History can matter: non-markovian behavior of ancestral lineages. Systematic Biology. doi:10.1093/sysbio/syr012.

Champagnat, N., Lambert, A., 2007. Evolution of discrete populations and the canonical diffusion of adaptive dynamics. Annals of Applied Probability 17, 102–155.

Chen, W.Y., Bokka, S., 2005. Stochastic modeling of nonlinear epidemiology. Journal of Theoretical Biology 234, 455–470.

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. Proceedings. Biological Sciences 275, 887–895.

Day, T., Gandon, S., 2007. Applying population-genetic models in theoretical evolutionary epidemiology. Ecology Letters 10, 876–888.

Diekmann, O., Heesterbeek, J., Metz, J., 1990. On the definition and the computation of the basic reproduction ratio r0 in models for infectious-diseases in heterogeneous populations. Journal of Mathematical Biology 28, 365–382.

Drummond, A., Pybus, O., Rambaut, A., Forsberg, R., Rodrigo, A., 2003a. Measurably evolving populations. Trends in Ecology & Evolution 18, 481–488.

Drummond, A., Pybus, O.G., Rambaut, A., 2003b. Inference of viral evolutionary rates from molecular sequences. Advances in Parasitology 54, 331–358.

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

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

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2003c. Genealogies from time-stamped sequence data, in: Buck, C.E., Millard, A.R. (Eds.), Tools for constructing chronologies: crossing disciplinary boundaries. Springer. Volume 177 of Lecture Notes in Statistics, pp. 149–174 (Chapter 7).

Drummond, A.J., Rambaut, A., 2007. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7, 214.

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution 22, 1185–1192.

Drummond, P.D., Vaughan, T.G., Drummond, A.J., 2010. Extinction times in autocatalytic systems. Journal of Physical Chemistry a 114, 10481–10491.

Durrett, R., Levin, S., 1994. The importance of being discrete (and spatial). Theoretical Population Biology.

Edwards, A., Cavalli-Sforza, L., 1965. A method for cluster analysis. Biometrics, 362–375.

Edwards, C.T.T., Holmes, E.C., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. BMC Evolutionary Biology 6, 28.

Engen, S., Ø., B., Islam, A., 1998. Demographic and environmental stochasticity-concepts and definitions. Biometrics 54, 840–846.

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.
Ewing, G., Rodrigo, A., 2006a. Coalescent-Based estimation of population parameters when the number of demes changes over time. Molecular Biology and Evolution 23, 988–996.
Ewing, G., Rodrigo, A., 2006b. Estimating population parameters using the structured serial coalescent with bayesian MCMC inference when some demes are hidden. Evolutionary Bioinformatics 2, 227–235 (PMID: 19455215 PMCID: 2674663).
Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17, 368–376.
Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist 125, 1–15.
Felsenstein, J., 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Molecular Biology and Evolution 23, 691–700.
Ferguson, N.M., Galvani, A.P., Bush, R.M., 2003. Ecological and immunological determinants of influenza evolution. Nature 422, 428–433.
Fisher, R., 1930. Genetical Theory of Natural Selection. Clarendon Press, Oxford.
Fitch, W.M., Leiter, J.M., Li, X.Q., Palese, P., 1991. Positive darwinian evolution in human influenza a viruses. Proceedings of the National Academy of Sciences of the United States of America 88, 4270–4274.
Forsberg, R., Christiansen, F.B., 2003. A codon-based model of host-specific selection in parasites, with an application to the influenza a virus. Molecular Biology and Evolution 20, 1252–1259.
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., WHO Rapid Pandemic Assessment Collaboration, 2009. Pandemic potential of a strain of influenza a (H1N1): early findings. Science 324, 1557–1561.
Frost, S.D.W., Volz, E.M., 2010. Viral phylodynamics and the search for an 'effective number of infections'. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 365, 1879–1890.
Gao, F., Bailes, E., Robertson, D.L., Chen, Y., Rodenburg, C.M., Michael, S.F., Cummins, L.B., Arthur, L.O., Peeters, M., Shaw, G.M., Sharp, P.M., Hahn, B.H., 1999. Origin of HIV-1 in the chimpanzee pan troglodytes troglodytes. Nature 397, 436–441.
Gao, F., Yue, L., White, A.T., Pappas, P.G., Barchue, J., Hanson, A.P., Greene, B.M., Sharp, P.M., Shaw, G.M., Hahn, B.H., 1992. Human infection by genetically diverse SIVSM-related HIV-2 in West Africa. Nature 358, 495–499.
Gardiner, C.W., 2009. Stochastic methods: a handbook for the natural and social sciences. Springer series in synergetics, fourth ed. Springer, Berlin.
Gernhard, T., 2008. The conditioned reconstructed process. Journal of Theoretical Biology 253, 769–778.
Gibbs, M., Gibbs, A., 2006. Molecular virology: was the 1918 pandemic caused by a bird flu? Nature 440, E8.
Gilbert, M.T.P., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the americas and beyond. Proceedings of the National Academy of Sciences of the United States of America 104, 18566–18570.
Gillespie, D., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22, 403–434.
Gillespie, D., 2001. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of Chemical Physics 115, 1716.
Gordo, I., Campos, P.R.A., 2007. Patterns of genetic variation in populations of infectious agents. BMC Evolutionary Biology 7, 116.
Grassly, N., Holmes, E., 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Molecular Biology and Evolution 14, 239–247.
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.
Griffiths, R.C., Tavaré, S., 1994. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society B: Biological Sciences 344, 403–410.
Guindon, S., Rodrigo, A.G., Dyer, K.A., Huelsenbeck, J.P., 2004. Modeling the site-specific variation of selection patterns along lineages. Proceedings of the National Academy of Sciences of the United States of America 101, 12957–12962.
Harvey, P.H., Pagel, M.D., 1991. The comparative method in evolutionary biology. Oxford University Press, Oxford.
Heled, J., Drummond, A., 2008. Bayesian inference of population size history from multiple loci. BMC Evolutionary Biology 8, 289.
Hirsch, V.M., Olmsted, R.A., Murphey-Corb, M., Purcell, R.H., Johnson, P.R., 1989. An African primate lentivirus (SIVsmclosely) related to HIV-2. Nature 339, 389–392.
Holmes, E., Worobey, M., Rambaut, A., 1999. Phylogenetic evidence for recombination in dengue virus. Molecular Biology and Evolution 16, 405–409.
Holmes, E.C., 2004. The phylogeography of human viruses. Molecular Ecology 13, 745–756.
Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., George, K.S., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J., Taubenberger, J.K., 2005. Whole-Genome analysis of human influenza a virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. PLoS Biology 3, e300.
Hudson, R., 1990. Gene genealogies and the coalescent process. In: Futuyma, D., Antonovics, J. (Eds.), Oxford Surveys in Evolutionary Biology, vol. 7. Oxford University Press, Oxford, pp. 1–44.
Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310–2314.
Hughes, G.J.UK HIV Drug Resistance Collaboration, Fearnhill, E.UK HIV Drug Resistance Collaboration, Dunn, D.UK HIV Drug Resistance Collaboration, Lycett, S.J.UK HIV Drug Resistance Collaboration, Rambaut, A.UK HIV Drug Resistance Collaboration, Leigh Brown, A.J.UK HIV Drug Resistance Collaboration, 2009. Molecular phylodynamics of the heterosexual HIV epidemic in the united kingdom. PLoS Pathogens 5, e1000590.
Huson, D., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14, 68.
Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. Journal of Molecular Evolution 54, 156–165.
Keele, B.F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., Loul, S., Ngole, E.M., Bienvenue, Y., Delaporte, E., Brookfield, J.F.Y., Sharp, P.M., Shaw, G.M., Peeters, M., Hahn, B.H., 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313, 523–526.
Keeling, M.J., Rohani, P., 2008. Modeling infectious diseases in humans and animals. Princeton University Press, Princeton.
Kelly, J.K., Williamson, S., Orive, M.E., Smith, M.S., Holt, R.D., 2003. Linking dynamical and population genetic models of persistent viral infection. The American Naturalist 162, 14–28.
Kermack, W., McKendrick, A., 1927. A contribution to the mathematical theory of infections. Proceedings of the Royal Society of London. Series A 115, 700–721.
Kingman, J., 1982. The coalescent. Stochastic Processes and Their Applications 13, 235–248.
Kishino, H., Thorne, J.L., Bruno, W.J., 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Molecular Biology and Evolution 18, 352–361.
Koelle, K., Cobey, S., Grenfell, B., Pascual, M., 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza a (H3N2) in humans. Science 314, 1898–18903.
Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796.
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.
Lemey, P., Pybus, O.G., Wang, B., Saksena, N.K., Salemi, M., Vandamme, A., 2003. Tracing the origin and history of the HIV-2 epidemic. Proceedings of the National Academy of Sciences of the United States of America 100, 6588–6592.
Lemey, P., Rambaut, A., Drummond, A.J., Suchard, M.A., 2009a. Bayesian phylogeography finds its roots. PLoS Computational Biology 5, e1000520.
Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in continuous space and time. Molecular Biology and Evolution.
Lemey, P., Suchard, M., Rambaut, A., 2009b. Reconstructing the initial global spread of a human influenza pandemic: a bayesian spatial–temporal model for the global spread of H1N1pdm. PLoS Currents. Influenza, RRN1031 (PMID: 20029613).
Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J., Wang, H., Crameri, G., Hu, Z., Zhang, H., et al., 2005. Bats are natural reservoirs of SARS-like coronaviruses. Science 310, 676.
Li, W., Wong, S., Li, F., Kuhn, J., Huang, I., et al., 2006. Animal origins of the severe acute respiratory syndrome coronavirus: insight from ACE2–S-protein interactions. Journal of Virology 80, 4211.
Liljeros, F., Edling, C., Amaral, L., Stanley, H., Aberg, Y., 2001. The web of human sexual contact. Nature, 907–908.
Lindstrom, S.E., Cox, N.J., Klimov, A., 2004. Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957–1972: evidence for genetic divergence and multiple reassortment events. Virology 328, 101–119.
Lloyd, A., May, R., 2001. How viruses spread among computers and people. Science, 1316–1317.
Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., Ingersoll, R., Sheppard, H.W., Ray, S.C., 1999. Full-Length human immunodeficiency virus type 1 genomes from subtype C-Infected seroconverters in india, with evidence of intersubtype recombination. The Journal of Virology 73, 152–160.
Maddison, D., Maddison, W., 2005. MacClade 4.08. Massachusetts, Sinauer Associates, Sunderland.
Markov, P., Pepin, J., Frost, E., Deslandes, S., Labbe, A., Pybus, O., 2009. Phylogeography and molecular epidemiology of hepatitis C virus genotype 2 in Africa. Journal of General Virology 90, 2086.
Mau, B., Newton, M.A., Larget, B., 1999. Bayesian phylogenetic inference via markov chain monte carlo methods. Biometrics 55, 1–12.
Minin, V.N., Bloomquist, E.W., Suchard, M.A., 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution 25, 1459–1471. Available from: <http://mbe.oxfordjournals.org/content/25/7/1459.full.pdf+html>.

Motomura, K., Kusagawa, S., Lwin, H.H., Thwe, M., Kato, K., Oishi, K., Yamamoto, N., Zaw, M., Nagatake, T., Takebe, Y., 2003. Different subtype distributions in two cities in myanmar: evidence for independent clusters of HIV-1 transmission. AIDS 17.

Nakajima, K., Desselberger, U., Palese, P., 1978. Recent human influenza A (H1N1) viruses are closely related genetically to strains isolated in 1950. Nature 274, 334–339.

Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Holmes, E.C., 2009. The origin and global emergence of adamantane resistant A/H3N2 influenza viruses. Virology 388, 270–278.

Nelson, M.I., Viboud, C., Simonsen, L., Bennett, R.T., Griesemer, S.B., George, K.S., Taylor, J., Spiro, D.J., Sengamalay, N.A., Ghedin, E., Taubenberger, J.K., Holmes, E.C., 2008. Multiple reassortment events in the evolutionary history of H1N1 influenza a virus since 1918. PLoS Pathogens 4, e1000012.

Nicholls, G., Gray, R., 2008. Dated ancestral trees from binary trait data and their application to the diversification of languages. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 545–566.

O'Fallon, B.D., 2010. A method to correct for the effects of purifying selection on genealogical inference. Molecular Biology and Evolution 27, 2406–2416.

O'Fallon, B.D., Seger, J., Adler, F.R., 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Molecular Biology and Evolution 27, 1162–1172.

Opgen-Rhein, R., Fahrmeir, L., Strimmer, K., 2005. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evolutionary Biology 5, 6.

Paraskevis, D., Deforche, K., Lemey, P., Magiorkinis, G., Hatzakis, A., Vandamme, A., 2005. SlidingBayes: exploring recombination using a sliding window approach based on bayesian phylogenetic inference. Bioinformatics 21, 1274–1275.

Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., Daszak, P., 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiology and Molecular Biology Reviews 72, 457–470.

Parsons, T.L., Quince, C., 2007. Fixation in haploid populations exhibiting density dependence I: the non-neutral case. Theoretical Population Biology 72, 121–135.

Parsons, T.L., Quince, C., Plotkin, J.B., 2010. Some consequences of demographic stochasticity in population genetics. Genetics 185, 1345–1354.

Pybus, O.G., Barnes, E., Taggart, R., Lemey, P., Markov, P.V., Rasachak, B., Syhavong, B., Phetsouvanah, R., Sheridan, I., Humphreys, I.S., Lu, L., Newton, P.N., Klenerman, P., 2009. Genetic history of hepatitis c virus in East Asia. Journal of Virology 83, 1071–1082.

Pybus, O.G., Charleston, M.A., Gupta, S., Rambaut, A., Holmes, E.C., Harvey, P.H., 2001. The epidemic behavior of the hepatitis c virus. Science 292, 2323–2325.

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. Molecular Biology and Evolution 20, 381–387.

Pybus, O.G., Rambaut, A., 2009. Evolutionary analysis of the dynamics of viral infectious disease. Nature Reviews. Genetics 10, 540–550.

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.

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

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.

Rannala, B., Yang, Z., 2007. Inferring speciation times under an episodic molecular clock. Systematic Biology 56, 453–466.

Reis, M., Hay, A.J., Goldstein, R.A., 2009. Using Non-Homogeneous models of nucleotide substitution to identify host shift events: application to the origin of the 1918 'Spanish' influenza pandemic virus. Journal of Molecular Evolution 69, 333–345.

Robertson, D.L., Hahn, B.H., Sharp, P.M., 1995. Recombination in AIDS viruses. Journal of Molecular Evolution 40, 249–259.

Rodrigo, A.G., Shpaer, E.G., Delwart, E.L., Iversen, A.K., Gallo, M.V., Brojatsch, J., Hirsch, M.S., Walker, B.D., Mullins, J.I., 1999. Coalescent estimates of HIV-1 generation time in vivo. Proceedings of the National Academy of Sciences of the United States of America 96, 2187–2191.

Ryder, R.J., Nicholls, G.K., 2011. Missing data in a stochastic dollo model for binary trait data, and its application to the dating of proto-indo-european. Journal of the Royal Statistical Society: Series C (Applied Statistics) 60, 71–92.

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.

Salminen, M., Carr, J., Burke, D., McCutchan, F., 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Research and Human Retroviruses 11, 1423–1425.

Sanderson, M., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution 14, 1218–1231.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Molecular Biology and Evolution 19, 101–109.

Santiago, M.L., Range, F., Keele, B.F., Li, Y., Bailes, E., Bibollet-Ruche, F., Fruteau, C., Noe, R., Peeters, M., Brookfield, J.F.Y., Shaw, G.M., Sharp, P.M., Hahn, B.H., 2005. Simian immunodeficiency virus infection in Free-Ranging sooty mangabeys (*Cercocebus atys atys*) from the Tai forest, cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. Journal of Virology 79, 12515–12527.

Santiago, M.L., Rodenburg, C.M., Kamenya, S., Bibollet-Ruche, F., Gao, F., Bailes, E., Meleth, S., Soong, S., Kilby, J.M., Moldoveanu, Z., Fahey, B., Muller, M.N., Ayouba, A., Nerrienet, E., McClure, H.M., Heeney, J.L., Pusey, A.E., Collins, D.A., Boesch, C., Wrangham, R.W., Goodall, J., Sharp, P.M., Shaw, G.M., Hahn, B.H., 2002. SIVcpz in wild chimpanzees. Science 295, 465.

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

Shapiro, B., Ho, S., Drummond, A., Suchard, M., Pybus, O., Rambaut, A., 2010. A Bayesian phylogenetic method to estimate unknown sequence ages. Molecular Biology and Evolution.

Sharp, P., Bailes, E., Gao, F., Beer, B., Hirsch, V., Hahn, B., 2000. Origins and evolution of AIDS viruses: estimating the time-scale. Biochemical Society Transactions 28, 275–282.

Siebenga, J.J., Lemey, P., Kosakovsky Pond, S.L., Rambaut, A., Vennema, H., Koopmans, M., 2010. Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. PLoS Pathogens 6, e1000884.

Slatkin, M., Maddison, W.P., 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123, 603–613.

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.

Smith, J., 1992. Analyzing the mosaic structure of genes. Journal of Molecular Evolution 34.

Stack, J.C., Welch, J.D., Ferrari, M.J., Shapiro, B.U., Grenfell, B.T., 2010. Protocols for sampling viral sequences to study epidemic dynamics. Journal of the Royal Society Interface 7, 1119–11127.

Stadler, T., 2009. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 261, 58–66.

Stadler, T., 2010. Sampling-through-time in birth–death trees. Journal of Theoretical Biology 267, 396–404.

Strimmer, K., Pybus, O.G., 2001. Exploring the demographic history of dna sequences using the generalized skyline plot. Molecular Biology and Evolution 18, 2298–2305.

Swofford, D., 2003. PAUP*: phylogenetic analysis using parsimony (* and other methods). version 4. Massachusetts, Sinauer Associates, Sunderland.

Takahata, N., 1987. On the overdispersed molecular clock. Genetics 116, 169–179.

Takahata, N., 1991. Statistical models of the overdispersed molecular clock. Theoretical Population Biology 39, 329–344.

Takehisa, J., Kraus, M.H., Ayouba, A., Bailes, E., Van Heuverswyn, F., Decker, J.M., Li, Y., Rudicell, R.S., Learn, G.H., Neel, C., Ngole, E.M., Shaw, G.M., Peeters, M., Sharp, P.M., Hahn, B.H., 2009. Origin and biology of simian immunodeficiency virus in Wild-Living Western gorillas. Journal of Virology 83, 1635–1648.

Thorne, J., Kishino, H., Painter, I., 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15, 1647–1657.

Van Heuverswyn, F., Li, Y., Neel, C., Bailes, E., Keele, B.F., Liu, W., Loul, S., Butel, C., Liegeois, F., Bienvenue, Y., Ngolle, E.M., Sharp, P.M., Shaw, G.M., Delaporte, E., Hahn, B.H., Peeters, M., 2006. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444, 164.

Vanden Haesevelde, M., Peeters, M., Jannes, G., Janssens, W., Van Der Greon, G., Sharp, P., Saman, E., 1996. Sequence analysis of a highly divergent HIV-1-related lentivirus isolated from a wild captured chimpanzee. Virology 221, 346–350.

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

Wallace, R., HoDac, H., Lathrop, R., Fitch, W., 2007. A statistical phylogeography of influenza A H5N1. Proceedings of the National Academy of Sciences 104, 4473.

Wallinga, J., Lipsitch, M., 2007. How generation intervals shape the relationship between growth rates and reproductive numbers. Proceedings. Biological Sciences 274, 599–604.

Welch, D., Bansal, S., Hunter, D., 2011. Statistical inference to advance network models in epidemiology. Epidemics 3, 38–45.

Welch, D., Nicholls, G.K., Rodrigo, A., Solomon, W., 2005. Integrating genealogy and epidemiology: the ancestral infection and selection graph as a model for reconstructing host virus histories. Theoretical Population Biology 68, 65–75.

Wertheim, J., 2010. The re-emergence of H1N1 influenza virus in 1977: a cautionary tale for estimating divergence times using biologically unrealistic sampling dates. PLoS One 5, e11184.

Wertheim, J.O., Kosakovsky Pond, S.L., 2011. Purifying selection can obscure the ancient age of viral lineages. Molecular Biology and Evolution. Available from: <http://mbe.oxfordjournals.org/content/early/2011/06/22/molbev.msr170.full.pdf+html>.

Wertheim, J.O., Worobey, M., 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Computational Biology 5, e1000377.

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J., Kabongo, J.M., Kalengayi, R.M., Marck, E.V., Gilbert, M.T.P., Wolinsky, S.M., 2008. Direct evidence of extensive diversity of HIV-1 in kinshasa by 1960. Nature 455, 661–664.

Worobey, M., Telfer, P., Souquiere, S., Hunter, M., Coleman, C.A., Metzger, M.J., Reed, P., Makuwa, M., Hearn, G., Honarvar, S., Roques, P., Apetrei, C., Kazanji, M., Marx, P.A., 2010. Island biogeography reveals the deep history of SIV. Science 329, 1487.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39, 306–314.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution 14, 717.

Zimmer, S., Burke, D., 2009. Historical perspective–Emergence of influenza A (H1N1) viruses. New England Journal of Medicine.