JOURNAL OF THE ROYAL SOCIETY Interface

Protocols for sampling viral sequences to study epidemic dynamics

J. Conrad Stack, J. David Welch, Matt J. Ferrari, Beth U. Shapiro and Bryan T. Grenfell

J. R. Soc. Interface published online 10 February 2010
doi: 10.1098/rsif.2009.0530

JOURNAL OF THE ROYAL SOCIETY Interface

FirstCite® e-publishing

J. R. Soc. Interface
doi:10.1098/rsif.2009.0530
Published online

Protocols for sampling viral sequences to study epidemic dynamics

J. Conrad Stack[1,3,*], J. David Welch[2,3], Matt J. Ferrari[1,3], Beth U. Shapiro[1] and Bryan T. Grenfell[3,4,5]

[1]*Department of Biology,* [2]*Department of Statistics, and* [3]*Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA, USA*
[4]*Fogarty International Center, National Institutes of Health, Bethesda, MD, USA*
[5]*Department of Ecology and Evolutionary Biology and Woodrow Wilson School, Princeton University, Princeton, NJ 08544, USA*

With more emphasis being put on global infectious disease monitoring, viral genetic data are being collected at an astounding rate, both within and without the context of a long-term disease surveillance plan. Concurrent with this increase have come improvements to the sophisticated and generalized statistical techniques used for extracting population-level information from genetic sequence data. However, little research has been done on how the collection of these viral sequence data can or does affect the efficacy of the phylogenetic algorithms used to analyse and interpret them. In this study, we use epidemic simulations to consider how the collection of viral sequence data clarifies or distorts the picture, provided by the phylogenetic algorithms, of the underlying population dynamics of the simulated viral infection over many epidemic cycles. We find that sampling protocols purposefully designed to capture sequences at specific points in the epidemic cycle, such as is done for seasonal influenza surveillance, lead to a significantly better view of the underlying population dynamics than do less-focused collection protocols. Our results suggest that the temporal distribution of samples can have a significant effect on what can be inferred from genetic data, and thus highlight the importance of considering this distribution when designing or evaluating protocols and analysing the data collected thereunder.

Keywords: phylodynamics; epidemics; simulation

1. INTRODUCTION

Over the last decade, there has been significant progress in the application of coalescent theory to the reconstruction of past population dynamics using genetic sequence data (Pybus *et al.* 2000; Strimmer & Pybus 2001; Drummond *et al.* 2005; Hein *et al.* 2005; Drummond & Rambaut 2007; Minin *et al.* 2008). Work by Pybus *et al.* (2000) introduced the skyline plot, a non-parametric estimate of demographic history calculated from the phylogeny of related genetic sequences (Pybus *et al.* 2000). Originally conceived as a backwards projection from a single point in time, the skyline plot provides researchers with a quick and easy way to visualize the past genetic diversity of a population. Under coalescent theory, if the mutation rate within a population is known, genetic diversity can be translated into effective population size, providing a graphical depiction of how a population's size has fluctuated over time (Hein *et al.* 2005).

Subsequent improvements to this model include, most notably, nesting the skyline plot method within a Bayesian Markov chain Monte Carlo (MCMC) framework. This framework allows population size parameters to be estimated in various ways from a posterior sample of phylogenetic trees instead of the single tree used in the traditional skyline plot reconstruction. While this approach requires considerably more computational effort, the resulting Bayesian skyline plot provides a smoother estimate of population history than does the original skyline plot and, crucially, reflects the statistical uncertainty inherent in the inference of phylogenetic trees (Drummond *et al.* 2005). This method was incorporated into a program called BEAST (Drummond & Rambaut 2007).

The development of these phylodynamic methods has been of particular interest in the study of RNA virus epidemiology. Owing to their important public health impact, a number of surveillance programmes have been established to monitor RNA virus activity on global and local scales (Hanon *et al.* 2003; Gnaneshan *et al.* 2008). Sequence collection is rapidly becoming a routine part of these monitoring protocols and provides an increasingly large pool of genetic data from which valuable epidemiological information can be gleaned. For example, samples taken from consecutive epidemics have provided important insights into

*Author for correspondence (stack@psu.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2009.0530 or via http://rsif.royalsocietypublishing.org.

Received 3 December 2009
Accepted 20 January 2010

both the spatio-temporal dynamics of viral populations (Real *et al.* 2005; Nelson *et al.* 2008; Rambaut *et al.* 2008) and the details of strain interactions (Zhang *et al.* 2007; Rambaut *et al.* 2008).

However, the strong nonlinearity inherent in recurrent epidemics such as those caused by acute RNA viruses (ARVs) like measles and influenza strains the capabilities of coalescent-based reconstruction techniques. These methods were not designed to accommodate complex epidemiological dynamics. Instead, the skyline plot approximates arbitrarily complex dynamics using a multiple-change point model, where each inferred phylogenetic tree is partitioned into a fixed number of sets of coalescent events. Within each set, the population size can either vary according to a simple, pre-defined demographic model (i.e. linear or exponential growth) or stay constant. Given this, the complex demographic pattern exhibited by many ARVs (Earn *et al.* 2000; Finkenstadt *et al.* 2002; frequent epidemics followed by viral population bottlenecks in the form of epidemic troughs) provides a serious complication. A population bottleneck effectively causes all lineages in a tree to coalesce down to only a few lineages, making a skyline plot projection beyond that bottleneck considerably less reliable (figure 1; electronic supplementary material, figure S2). The practical consequence of this is that any samples taken at any time during an epidemic can only be expected to accurately depict an estimate of past population size change back to the last severe bottleneck.

While there is an increasing abundance of viral sequence data from surveillance programmes, very little work has shown how the timing of sample collection may affect the reconstruction of epidemic histories. These data may either represent a deliberate attempt to monitor a specific aspect of disease epidemiology (vaccine development, strain interaction and spread) or simply constitute a convenience sample. In this context, two important open questions are (i) how might the temporal distribution of sampling (i.e. the sampling protocol) affect the efficiency of the estimators of population size? and (ii) what sampling design can best be employed to reconstruct population histories using the Bayesian skyline plot? In this study, we investigate how, with finite resources allocated for disease surveillance programmes, we can optimize the sequence sampling of ARVs so as to best capture the underlying dynamics of a viral population.

To assess the influence of sampling protocol on the inference of past population dynamics, we use a simulation that captures realistic nonlinear epidemic dynamics and individual host-level evolutionary changes in the viral genotype. As coalescent methods typically make the explicit assumption that sequence variation is neutral, the directional selection to escape host immunity common in some ARVs (i.e. flu, Restif & Grenfell 2007) could significantly obscure the relationship between the sampling protocol and the reconstruction of past population dynamics. To avoid this complication, we model our simulations on the ARV measles, which causes remarkably strong and long-lived immunity against all viral variants within the host. In this case, the observed viral variation is considered to be effectively neutral, at least with respect to adaptive immunity. As such, selection is not included in our simulation.

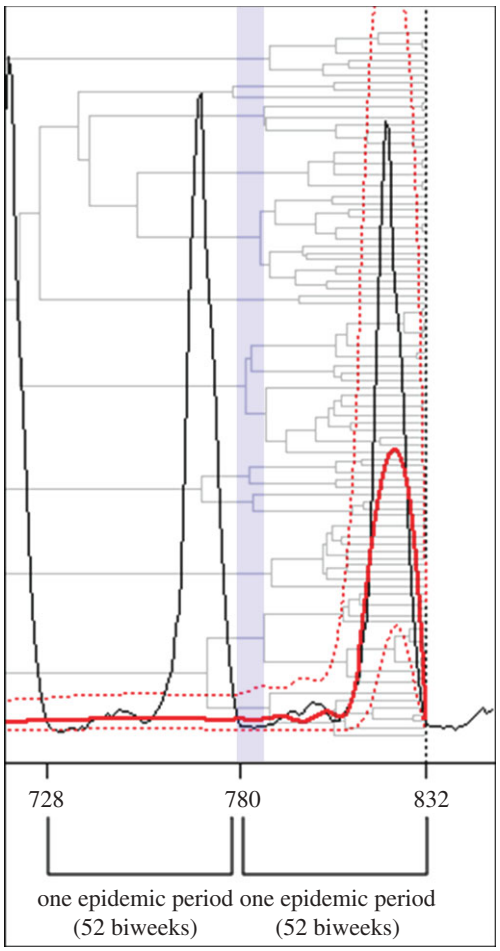

Figure 1. The thin black curve shows the population dynamics of the TSIR model or the actual time series. The thick red curve shows an estimation of the actual time series as inferred by BEAST, with the two dotted red lines above and below representing the 95% confidence interval. The grey binary tree is the actual genealogy relating the sequences used in the BEAST reconstruction. The sequences themselves are from a short-term point sample analysis of 100 samples collected from generation 832. Note that the BEAST estimate (the skyline plot) does not project accurately past the population bottleneck (indicated by the blue-shaded region) and that most of the lineages in the tree have coalesced by that point. When reading the tree right to left, branch lengths are long during the epidemic phase—indicating population expansion—and then rapidly begin to coalesce—indicating a sharp contraction in the number of infecteds, and thus the bottleneck (electronic supplementary material, figure S1).

A further benefit to basing our simulation on measles virus is that its epidemiological dynamics are relatively well-studied and understood (Bolker & Grenfell 1995; Earn *et al.* 2000; Bjornstad *et al.* 2002; Grenfell *et al.* 2002; Glass *et al.* 2003; Xia *et al.* 2004). Measles in developed countries in the pre-vaccine era regularly appeared in multi-annual epidemics, which could vary in amplitude over several orders of magnitude and which were generally preceded by epidemic troughs (population bottlenecks). These recurrent epidemics

occurred because of the strong herd immunity imparted by measles infections, such that new cohorts of susceptible individuals had to accumulate to fuel the next epidemic. Further, there is significant seasonal variation in the infection rate of measles due in large part to the aggregation and dissipation of school-age children during school terms and holidays, respectively (Grenfell *et al.* 1995). The result of combining these two factors is a mix of annual and biennial dynamics where epidemics are triggered seasonally, but vary significantly in size owing to the relatively slow build up of susceptibles (figure 1; electronic supplementary material, figure S1). A time-series SIR (TSIR) model embodying these characteristic dynamics was developed and validated by fitting it to England and Wales case count data in the pre-vaccination era (Bjornstad *et al.* 2002; Grenfell *et al.* 2002).

In an ARV epidemic system, Holmes points out (Holmes 2008), the quality of any inference of population dynamics will be largely affected by the timing and design of sampling protocols. He suggests that with frequent and periodic epidemics comes the need to account for the 'graininess', or temporal distribution, of the sampling. We combine this TSIR model (parametrized using pre-vaccination London data) with an agent-based component (Ferguson *et al.* 2003) to serve as a basis for simulating the viral genetic variation underlying measles epidemics. In our model, each infected individual is assigned a bitstring to represent the consensus viral genome of their infection. This provides us with a framework from which we can take samples of simulated viral isolates across epidemic cycles, thus allowing us to analyse the effect of sampling on the coalescent-based reconstruction of realistic ARV epidemic dynamics. As this framework provides all the epidemiological information for each individual (i.e. genetic sequence and transmission history), we are limited only by the substantial computational time demands of multiple MCMC inferences. Our aim is to improve the predictive capacity of genetic data collected as a part of surveillance programmes by seeking the most efficient way to sample epidemics. With our simulation framework, this key question can be addressed: what is the optimal strategy of viral sampling to capture recent epidemic history?

2. MATERIAL AND METHODS

2.1. Simulating the epidemic and evolutionary dynamics of measles

2.1.1. Demographic model. Dynamics of ARV (measles) epidemics were simulated using a TSIR model in which hosts move from susceptible (S) to infected (I) to recovered (R) classes in that order and then are assumed to leave the system (Bjornstad *et al.* 2002). Mathematically, the model is specified as follows:

$$S_{t+1} = S_t - I_{t+1} + b_t, \tag{2.1}$$

$$\lambda_t = \frac{\beta_t S_t I_t^{\alpha}}{N_t} \tag{2.2}$$

and

$$I_{t+1} = \text{Poisson}\,(\lambda t). \tag{2.3}$$

Individuals in the system are assumed to be interacting at random. Thus, at each time step, $(t \rightarrow t+1)$, individuals in the susceptible class are infected at a rate that reflects both the number of susceptibles and the number of infecteds. Each time step is two weeks long, reflecting the average infectious period of measles virus. The parameter α is a conversion factor from continuous to discrete time and is slightly less than one (Glass *et al.* 2003). β is the transmission rate, or the likelihood of infection of a person in S_t by a person in I_t during 'contact'. At the end of each time step, individuals in the infected class (I_t) are assumed to gain lifelong immunity (or die) and the susceptible class (S_{t+1}) is replenished with new births (b_t) that are kept constant in our simulation.

2.1.2. Evolutionary model. The transcending immunity generated by measles virus means there is little to no selection for immune evasion (Grenfell *et al.* 2004), so we use a model of neutral evolution. To simulate neutral evolution, individuals in the infected class were assigned a 1000 base-pair sequence that was mutated over time according to the HKY85 model of nucleotide substitution (Hasegawa *et al.* 1985). The HKY model was chosen for simplicity; we attempted to minimize the runtime of the MCMC analyses, which, even with this simpler model, ranged from a few hours to a few days. During a time step, susceptible individuals who were recruited into the infected class randomly drew a 'parent' from current infected individuals and inherited a copy of their sequence. As the sequence is passed from parent to child (representing the change from generation t to generation $t+1$), nucleotides mutated with an independent, binomial probability of $2.2\text{e}{-4}$ per site per generation favouring transitions to transversions 6.4 to 1. This model and its parametrization are consistent with the genetic variability of measles virus seen in the wild (Khne *et al.* 2006).

2.2. Simulations

The population dynamic component of the TSIR model was run with initial values $S = 135\,300$ and $I = 300$ (taken from pre-vaccination London data (Bjornstad *et al.* 2002; Grenfell *et al.* 2002)), up to 'generation 0', when consistent biennial cycles were achieved (figure 1; electronic supplementary material, figure S1). Following this burn-in period, all infected individuals were initially assigned the same consensus sequence, and these individual sequences were then allowed to mutate over the course of the simulation. The simulation was run for an additional 1000 biweeks (roughly 38.5 years) beyond the burn-in period. The first 200 of these 1000 generations were then discarded as a second burn-in, allowing the evolutionary dynamics to stabilize. Expected births per time step, b_t, was set to 2150: an average value obtained from the same London data (Bjornstad *et al.* 2002). The depth (or severity) of population bottlenecks in the TSIR model is largely a function of the initial S and b_t values. In an attempt to control for the size of the bottleneck, we ran additional simulations that scaled the number of births and the parameter starting

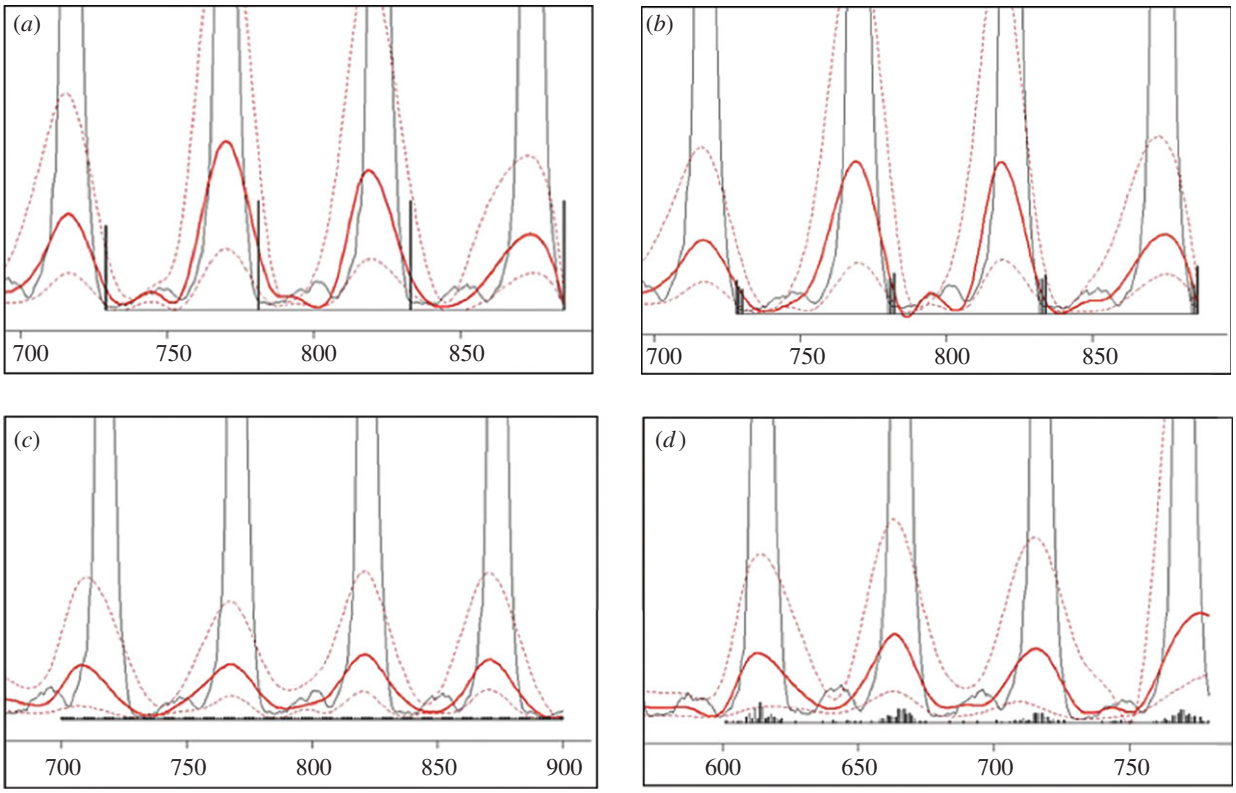

Figure 2. Examples of the four sampling protocols described in §2. The thin black curves show the TSIR time series and the thick red curves show example BEAST reconstructions. The thick vertical lines on the x-axis show the relative number of sequences taken at that generation. In this example, 400 samples were taken under each protocol (long term). (a) Point sampling, (b) fuzzy sampling, (c) serial sampling and (d) convenience sampling.

values up by 2.2 and down by 1/2, respectively, to generate a slightly more irregular epidemic pattern.

2.3. Sampling protocols

Fundamentally, one viral sample refers to the sequence state of an infected individual chosen at random from a specified generation (a time span of roughly two weeks). A sampling set comprises a variable number of viral samples collected over time according to a specific sampling protocol. These sets are further categorized as short or long term depending on whether the temporal span of the samples covered less than one or more than one epidemic cycle, respectively. We tested four different sampling protocols (point sampling, fuzzy point sampling, serial sampling and convenience sampling) in both temporal categories, all of which are described in detail below.

To identify which regions in an epidemic cycle provide the most information about the past population dynamics, single generation sampling (point sampling), in which individual sampling sets consisted of 100 samples taken from a single generation, t, was performed on each generation in the epidemic cycle (52 generations total; figure 1). Taking more than 100 samples did not significantly improve the reconstruction. This sampling protocol was a good starting point for our investigation in part because coalescent methods were originally developed to deal only with contemporaneous samples. To compensate for the impracticality of point sampling in the field, similar analyses were also performed using sampling sets constructed in a less precise, more realistic way, to account for natural logistical difficulties when collecting a large number of samples over a short time span. Similar to point sampling, what we call fuzzy point sampling was done by sampling randomly from around a target generation, with sampling points ultimately encompassing three generations $\{t-1,\ t,\ t+1\}$.

Long-term point and fuzzy point sampling sets were compiled by combining multiple short-term point or fuzzy point sampling sets, respectively. These larger sampling sets spanned multiple epidemic cycles, keeping constant the region of the epidemic cycle from which samples were taken (figure 2*a*,*b*). In this way, we aimed to test how well consistently sampling specific regions of the epidemic cycle affected coalescent-based reconstruction over a longer term.

As a baseline strategy, we considered uniform or 'serial' sampling in which a set number of sequences were taken from each consecutive generation. To make this strategy comparable to point sampling, we initiated a trade-off between the frequency of sampling and the number of samples per generation, keeping the total number of samples similar. Intuitively, this approach to sampling might reflect the changing viral diversity over time in finer detail and thus capture all or most of the population fluctuations. Long-term serial sampling sets were constructed by combining multiple short-term serial sampling sets (figure 2*c*).

These three highly structured surveillance protocols, in which sampling is deliberately targeted at certain points in time, are in distinct contrast to our last sampling protocol, convenience sampling. This protocol is meant to reflect the manner in which samples might be collected in the field, where demand for samples is driven more by availability than by a set quota. For example, as a natural reaction to increasing case counts and rising public health concern about a novel genotype, more samples might be collected, resulting in frequency-dependent sampling sets where the number of cases and the number of samples rise and fall roughly together (figure 2*d*). Short-term sampling sets again usually consisted of 100 viral samples collected at random, with replacement, from all the generations within an epidemic cycle, with each generation having a probability of being chosen proportional to the number of infected individuals in that generation.

Overall, our goals were to find regions of the epidemic curve that had a positive influence on the accuracy of coalescent-based reconstruction and also to investigate the relative efficacy of different sampling protocols.

2.4. Coalescent analysis

The MCMC analyses were run for between 30 million and 100 million iterations, depending on the number of sequences used. The skyline plot method requires that the number of change points (group size) is specified *a priori*. Group sizes ranging from 10 to 30 were tested for both short- and long-term analyses with no discernable difference. Larger group sizes were not used owing to a large increase in the computational time required, and it is unclear how this might have affected the results. The program Tracer (Rambaut & Drummond 2007) was used to check for convergence and to confirm that a sufficient number (100+) of approximately independent samples were collected. Skyline plots were generated in Tracer, as well. To further check convergence, multiple MCMC chains were run for each analysis.

2.5. Comparing skyline plots with true time series

To test how well the BEAST reconstructions captured the epidemic dynamics of the simulated TSIR time series, the sum of squared differences (SSDs) between the two over a certain time span was calculated. When appropriate, the entire posterior distribution of trees and population sizes from a BEAST analysis was used to calculate a corresponding distribution of SSDs. For each comparison of estimated versus actual time series, the time span along which the SSD value was calculated comprised all generations spanned by the earliest and latest samples of the set plus one epidemic cycle (e.g. generations 832 to 780 in figure 1; see the electronic supplementary material, figures S17–S20, for more information on this choice). Visual inspection of all the time-series comparisons confirmed that lower SSD values were indicative of a more accurate fit between the Bayesian skyline estimation and the true time series. The correlation between the actual time series and a BEAST estimate was also routinely calculated, showing similar results (electronic supplementary material, figure S8). Other metrics such as absolute differences were tried, but yielded similar results to the SSDs, and are thus omitted. It is also worth noting that BEAST reconstructs the effective population size. In most practical situations, this measure will be an underestimate of the actual population size, but in our simulations the two are equivalent.

For long-term point and fuzzy point sampling and for some short-term serial sampling analyses we calculated the correlation between the normalized SSD values and proportion of samples in a sampling set that fell into one of eight regions of the epidemic cycle (electronic supplementary material, figure S9). The normalized SSD values are effectively the average SSD per generation—in other words, to normalize them, we divided the SSD values by the time span of the comparison. A significant negative correlation between the normalized SSD values and the number of samples collected from a particular region indicated that concentrated sampling in that region yielded a skyline plot with a better fit to the true time series. A significant positive correlation indicated the opposite.

2.6. Software

All coalescent-based reconstructions were done in BEAST. All simulations and other statistical tests were done in R, using the *ape* package (Paradis *et al.* 2004) to handle phylogenetic and nucleotide data.

3. RESULTS

To identify the regions in an epidemic cycle that provided the most information about the past population dynamics, short-term point and fuzzy sampling analyses were performed, targeting every generation over a span of two epidemic periods (2×52 generations) resulting in 208 total analyses. The results clearly show that when a sample set is composed of samples exclusively collected as a major epidemic subsides or directly following a major epidemic, there is a significant drop in the SSD value between the BEAST reconstruction and the actual time series (figure 3). This experiment was repeated on the scaled control simulations described above with similar results, indicating that initial values and bottleneck magnitude do not affect the results (electronic supplementary material, figures S4–S7).

Based on this point sampling survey of the epidemic curve, we then hypothesized that concentrating our sampling in the major epidemic down slope and in the trough area directly following would significantly improve BEAST reconstructions for long-term fuzzy and point sampling sets. Long-term point and fuzzy sampling sets comprising either 200 samples (two target generations, 100 samples each) or 400 samples (four target generations, 100 samples each) showed that consistently sampling the trough region following

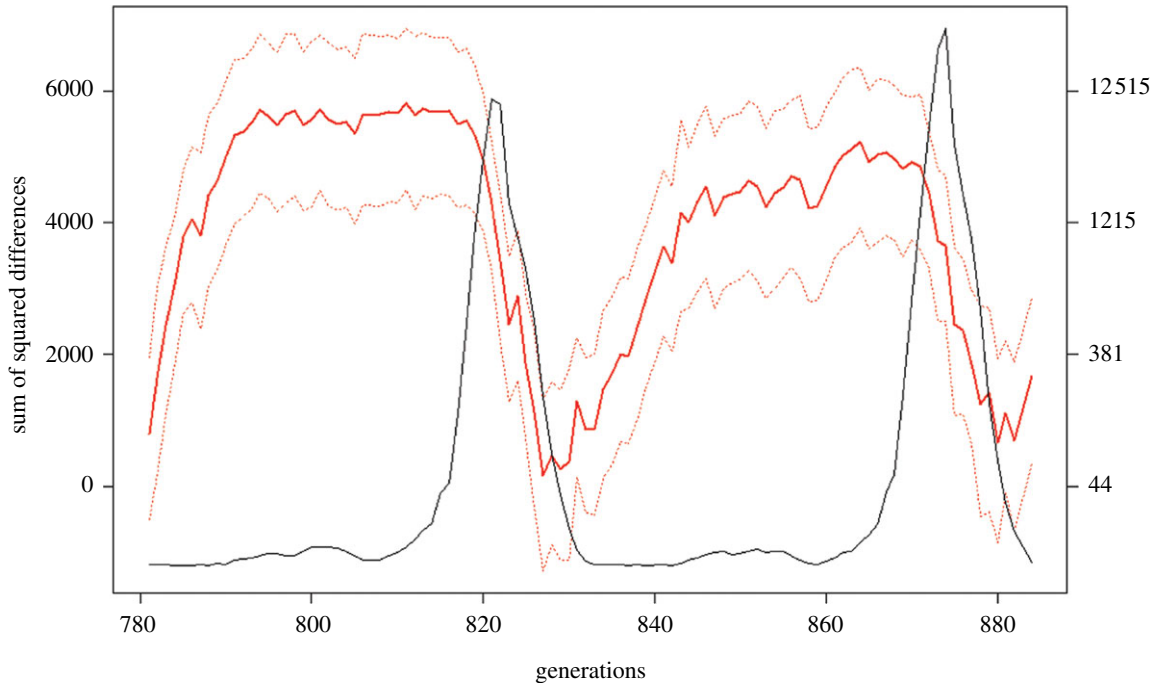

Figure 3. Point sampling analyses (§2), each consisting of 100 samples, were done for every generation between 781 and 886. The solid black line is the true time series. The solid red line shows the mean SSD values between the BEAST posterior reconstruction and the TSIR time series as a function of generation. The dotted red lines indicate the upper and lower 95% posterior density SSD values for each generation. Note that the SSD value drops significantly when samples are collected from the region immediately following a major epidemic (see electronic supplementary material, table S1, for more information).

a major epidemic was correlated with a significant drop in the SSDs between the BEAST reconstruction and the actual TSIR time series (figure 4; electronic supplementary material, figure S9). No other region showed this significant negative correlation. Visual inspection was helpful in confirming that sampling in this region provided a better picture of the dynamics (electronic supplementary material, figures S14–S16).

Short-term serial sampling SSDs also showed a very strong negative correlation with sampling that was chiefly sampled from the two regions immediately following a major epidemic peak (electronic supplementary material, figure S10). Short-term convenience sampling SSDs showed neither positive nor negative correlations with sampling in any of the eight regions. Doubling the number of samples for both of these protocols did not have a significant effect on these results. Long-term serial and convenience sampling, where the density of samples collected from each region was fixed across all analyses, reliably captured the biennial pattern of the TSIR model, but generally failed to capture the annual pattern (figure 2*c*,*d*; electronic supplementary material, figures S11 and S12).

A quantitative comparison of the different protocols (figure 4) shows that the ranges of SSD values for serial and convenience sampling protocols fall broadly between the SSD value ranges of both the best and worst fuzzy and point sampling protocols. For the two regions directly following a major epidemic (f and g), the range of SSD values for fuzzy and point sampling sets is lower than all other regions and lower than the ranges of both serial and convenience sampling. Only region g, covering the end of a major epidemic and the beginning of an epidemic trough, was significantly lower than all other protocols, however (figure 4).

3.1. Breaking up long-term analyses

During the course of our sampling study, we observed that none of the long-term sampling protocols were able to fully capture the true biennial dynamics of the TSIR model. To explore why this was the case, we broke up long-term sampling sets into short-term series. The shorter-term analyses were run separately and the results concatenated to form a new long-term reconstruction that was then compared with the actual time series. For all sampling protocols, the concatenated smaller analyses fit the actual time series as well as if not substantially better than did the full, longer term analysis (electronic supplementary material, figure S13).

4. DISCUSSION

It is understood that disease surveillance programmes have various objectives. Our research suggests that, if a goal is to reconstruct population histories, the temporal distribution of sampling can strongly influence inference. As a corollary, it is worth considering the potential biases if the inference of population history is to be made based on data collected for another purpose. Surveillance sampling protocols need not be rewritten based on our findings, but our findings can be helpful in deciding whether and when extra samples are worth collecting and also in choosing which data to analyse and, potentially, how to best subdivide them.

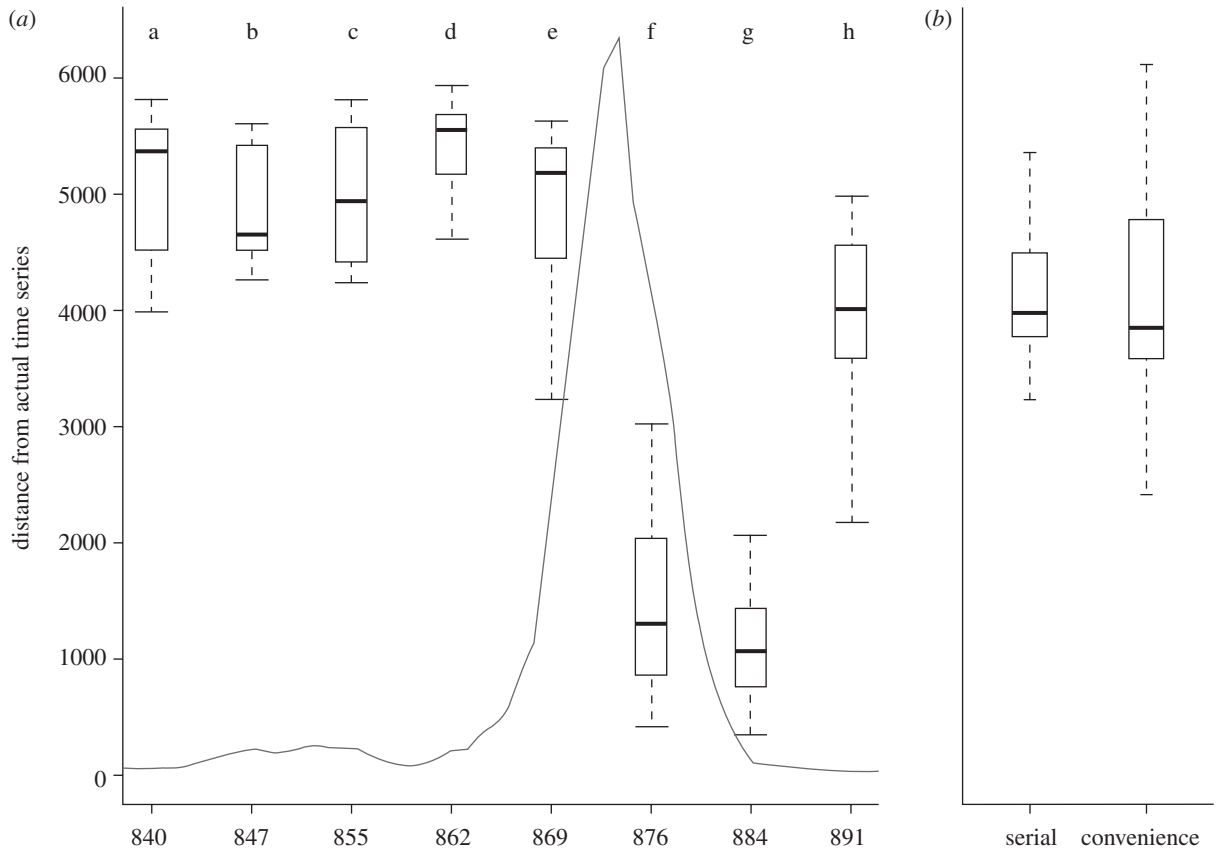

Figure 4. (*a*) Comparison between fuzzy and point sampling and (*b*) serial and convenience sampling. (*a*) Range of normalized SSD values (calculated between skyline plot means and the actual time series only) above the regions of the epidemic curve (electronic supplementary material, figure S3) they represent. The faint grey line shows which region the point and fuzzy sampling sets came from. Letters a–f above each region have been included for ease of reference. (*b*) For comparison, the normalized SSD values (means) for aggregated serial and convenience sampling analyses. Point and fuzzy sampling offers no real advantage over serial and convenience sampling except when point and fuzzy samples are taken exclusively from the regions following the peak of a major epidemic.

In the context of reconstructing population history, the results of our simulations suggest a specific strategy to maximize the effectiveness of viral sampling or post-sampling analysis. Because very little demographic information survives through strong bottlenecks (figure 1; electronic supplementary material, figure S20), point or fuzzy sampling just before a bottleneck maximizes the amount of information that can be retrieved about the past population dynamics up to that point. This can be seen visually (figure 1; electronic supplementary material, figures S14–S16) and is also reflected in the lower SSD values associated with sampling in this region (figure 4). Since the epidemic model under consideration is characterized by a biennial cycle of alternating major and minor epidemics with strong bottlenecks in between, the most we can expect to reconstruct is one full epidemic period (2 years). Specifically, samples taken at the end of a major epidemic and early into the following bottleneck are the most informative for this purpose, as there we find lineages that fall into two critical classes: those that rapidly die out (transients that only exist because of the rapid population expansion experienced during an epidemic) and those that persist throughout the major epidemic period (electronic supplementary material, figure S20). Combined, these two lineage types help to indicate population growth and decline, in the case of the former, and a low level of population persistence, in the case of the latter. Sampling concentrated in other, non-ideal regions can be informative, but often omits samples from crucial parts of the epidemic cycle. Additionally, an accurate reconstruction of the past population dynamics implies the accurate inference of other phylogenetically useful quantities such as evolution rate and divergence times. Even in cases where the dynamics of a virus are well known beforehand, sampling the tail ends of major epidemics could still be justified on this basis—that doing so better estimates the parameters upon which the population history reconstruction is based (Seo *et al.* 2002).

Interestingly, BEAST analyses of many long-term sampling sets within each class of the sampling protocol tended not to capture the full annual and biennial dynamics of the TSIR model as well as did a concatenation of the skyline plots from short-term analyses. Indeed, breaking up the best-fitting of our long-term analyses provided appreciably better or comparable estimations (electronic supplementary material, figure S13). This is important to consider when the temporal range of samples is long. Additional information about population dynamics might be gained from a dataset where the samples have a high degree of temporal dispersion

simply by subdividing the data into subsets as our results suggest: that is, by aggregating samples taken in regions suspected of being on the tail end of major outbreaks.

Generally, we believe that the inability of the skyline method to capture both short- and long-term dynamics equally is due to insufficient prior probabilities being placed on the population size within BEAST. There is a definite need for new priors that specifically incorporate the seasonal dynamics of many infectious diseases. Much work has been done along these lines (Minin *et al.* 2008) and we, too, are in the process of testing new priors. It is also worth noting that a method to improve population history reconstruction through the use of multiple loci has been developed (Heled & Drummond 2008), although its impact on viral population history reconstruction is unclear given that many RNA viruses are too small to provide independent loci.

4.1. Implications for viral sampling and data analysis

The utility of virological surveillance is well established. In this study, we considered how sampling times can affect the inference of past population dynamics, which can be thought of as proxies for other phylogenetic parameters such as evolution rates and divergence times (Seo *et al.* 2002). While inferring past population dynamics is not always the primary goal of surveillance programmes, observing these changes can play an important auxiliary role for other monitoring objectives such as tracing strain interaction over seasons (Rambaut *et al.* 2008) or measuring the effectiveness of control programmes (Zhang *et al.* 2007; Rota *et al.* 2009). These other objectives of temporal monitoring are outside the scope of this study, but our findings provide a way to incorporate these objectives with the accurate recovery of past population dynamics.

With a view towards reconstructing epidemic dynamics, we make the following recommendations based on our study. When regions in the epidemic cycle cannot be even roughly identified, then heavy serial sampling should be aimed for. This should be a way to sample frequently enough so as to capture lineages from all regions of an epidemic and to get a broad view of the epidemic dynamics. It would also not risk over-sampling in a non-ideal region of the epidemic curve that would likely result in a worse reconstruction (figures 3 and 4). The most cost-effective and efficacious overall strategy (or supplement to an existing protocol) according to our results is point or fuzzy point sampling as an epidemic begins to subside. The feasibility of this protocol, however, would be tied to how well the end of an epidemic can be predicted (through a proxy like hospital case counts) and how effectively agents in the field could respond to the demand for samples. The quality of this prediction would need to be judged case-by-case, as incorrectly identifying an ideal target generation can negatively affect coalescent-based estimations (electronic supplementary material, figure S21). With good predictive information about when the end of an epidemic might occur, however, a systematic approach to sampling is generally preferable to convenience or haphazard sampling. Our results indicate that the information contained in these datasets more often provided a poor estimate of the past population history of a virus (figure 4). Interestingly, in the context of seasonal influenza surveillance, which focuses on samples from the end of influenza seasons to identify new antigenic variants (Smith *et al.* 2004), our analysis indicates that this provides the added benefit of relatively accurate reconstruction of recent epidemic dynamics. By contrast, any form of 'frequency-dependent' sampling during other viral epidemics will be less successful in this respect.

The limitations of our simulation model also need to be considered. While measles dynamics may be archetypal for a large number of ARVs (those for which seasonality and high acquired immunity drive recurrent epidemics and the effect of natural selection is relatively weak), the evolutionary model used here is too naive for ARVs on whose genome strong selection is acting. Because our study focuses exclusively on sampling by ignoring any effect of natural selection, this model will be less successful for non-measles-like ARVs.

This simulation framework can be easily expanded to include spatial dynamics, immune selection and other forms of natural selection. Ultimately, we aim to explore how selection-driven viral genetic diversity propagates across an epidemic metapopulation, thereby coming closer to developing a full phylodynamic model for RNA viruses (Grenfell *et al.* 2004). That is, we intend to form a more complete picture of how various population-level processes affect the underlying viral genealogies. With this information, genetic sequence data will become vastly more useful in predicting statistics such as migration rates among infected populations, a statistic which is currently difficult to predict with accuracy, especially in rural or underdeveloped areas. Ultimately, our goal is to better understand disease dynamics so as to effectively control their inevitable outbreak and spread.

J.C.S., B.S. and B.G. were supported by the National Institutes of Health (R01 GM083983-01) as was D.W. (R01 GM083603-01). B.G. and M.F. also were supported by the Bill and Melinda Gates Foundation, and the RAPIDD programme of the Science and Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

REFERENCES

Bjornstad, O. N., Finkenstadt, B. F. & Grenfell, B. T. 2002 Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecol. Monogr.* **72**, 169–184.

Bolker, B. & Grenfell, B. 1995 Space, persistence and dynamics of measles epidemics. *Phil. Trans. R. Soc. Lond. B* **348**, 309–320. (doi:10.1098/rstb.1995.0070)

Drummond, A. J. & Rambaut, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)

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

Earn, D. J., Rohani, P., Bolker, B. M. & Grenfell, B. T. 2000 A simple model for complex dynamical transitions in epidemics. *Science* **287**, 667–670. (doi:10.1126/science.287.5453.667)

Ferguson, N. M., Galvani, A. P. & Bush, R. M. 2003 Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433. (doi:10.1038/nature01509)

Finkenstadt, B. F., Bjornstad, O. N. & Grenfell, B. T. 2002 A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* **3**, 493–510. (doi:10.1093/biostatistics/3.4.493)

Glass, K., Xia, Y. & Grenfell, B. T. 2003 Interpreting time-series analyses for continuous-time biological models—measles as a case study. *J. Theor. Biol.* **223**, 19–25. (doi:10.1016/S0022-5193(03)00031-6)

Gnaneshan, S., Brown, K. E., Green, J. & Brown, D. W. 2008 On-line global/WHO-European regional measles nucleotide surveillance. *Euro Surveill.* **13**, 18 861.

Grenfell, B. T., Bolker, B. M. & Kleczkowski, A. 1995 Seasonality and extinction in chaotic metapopulations. *Proc. Biol. Sci.* **259**, 97–103. (doi:10.1098/rspb.1995.0015)

Grenfell, B. T., Bjornstad, O. N. & Finkenstadt, B. F. 2002 Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecol. Monogr.* **72**, 185–202.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L., Daly, J. M., Mumford, J. A. & Holmes, E. C. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)

Hanon, F. X., John, J. S., Steven, S. W. & Emiroglu, N. 2003 WHO European region's strategy for elimination of measles and congenital rubella infection. *Euro Surveill.* **8**, 129–130.

Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174. (doi:10.1007/BF02101694)

Hein, J., Schierup, M. H. & Wiuf, C. 2005 *Gene genealogies, variation and evolution: a primer in coalescent theory.* Oxford, UK: Oxford University Press.

Heled, J. & Drummond, A. 2008 Bayesian inference of population size history from multiple loci. *BMC Evol. Biol.* **8**, 289. (doi:10.1186/1471-2148-8-289)

Holmes, E. C. 2008 Comparative studies of RNA virus evolution. In *Origin and evolution of viruses* (eds E. Domingo, C. R. Parrish & J. J. Holland), pp. 119–134, 2nd edn. London, UK: Academic Press.

Khne, M., Brown, D. W. G. & Jin, L. 2006 Genetic variability of measles virus in acute and persistent infections. *Infect. Genet. Evol.* **6**, 269–276.

Minin, V. N., Bloomquist, E. W. & Suchard, M. A. 2008 Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471. (doi:10.1093/molbev/msn090)

Nelson, M. I. *et al.* 2008 Molecular epidemiology of a/h3n2 and a/h1n1 influenza virus during a single epidemic season in the United States. *PLoS Pathog.* **4**, e1000133. (doi:10.1371/journal.ppat.1000133)

Paradis, E., Claude, J. & Strimmer, K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)

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. & Drummond, A. J. 2007 Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.

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

Real, L. A. *et al.* 2005 Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proc. Natl Acad. Sci. USA* **102**, 12 107–12 111. (doi:10.1073/pnas.0500057102)

Restif, O. & Grenfell, B. T. 2007 Vaccination and the dynamics of immune evasion. *J. R. Soc. Interface* **4**, 143–153. (doi:10.1098/rsif.2006.0167)

Rota, P. A., Featherstone, D. A. & Bellini, W. J. 2009 Molecular epidemiology of measles virus. In *Measles: pathogenesis and control.* Current Topics in Microbiology and Immunology, vol. 330, pp. 129–150. Berlin, Germany: Springer.

Seo, T.-K., 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. (doi:10.1093/bioinformatics/18.1.115)

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E. & Fouchier, R. A. M. 2004 Mapping the antigenic and genetic evolution of influenza virus. *Science* **305**, 371–376. (doi:10.1126/science.1097211)

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

Xia, Y., Bjornstad, O. N. & Grenfell, B. T. 2004 Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.* **164**, 267–281. (doi:10.1086/422341)

Zhang, Y. *et al.* 2007 Molecular epidemiology of measles viruses in China, 1995–2003. *Virol. J.* **4**, 14. (doi:10.1186/1743-422X-4-14)