The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data

Sebastián Duchêne,*[1] David Duchêne,[2] Edward C. Holmes,[1,3] and Simon Y.W. Ho[1]

[1]School of Biological Sciences, University of Sydney, Sydney, NSW, Australia

[2]Research School of Biology, Australian National University, Canberra, ACT, Australia

[3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, Sydney Medical School, University of Sydney, Sydney, NSW, Australia

***Corresponding author:** E-mail: sebastian.duchene@sydney.edu.au.

Associate editor: Thomas Leitner

Abstract

Rates and timescales of viral evolution can be estimated using phylogenetic analyses of time-structured molecular sequences. This involves the use of molecular-clock methods, calibrated by the sampling times of the viral sequences. However, the spread of these sampling times is not always sufficient to allow the substitution rate to be estimated accurately. We conducted Bayesian phylogenetic analyses of simulated virus data to evaluate the performance of the date-randomization test, which is sometimes used to investigate whether time-structured data sets have temporal signal. An estimate of the substitution rate passes this test if its mean does not fall within the 95% credible intervals of rate estimates obtained using replicate data sets in which the sampling times have been randomized. We find that the test sometimes fails to detect rate estimates from data with no temporal signal. This error can be minimized by using a more conservative criterion, whereby the 95% credible interval of the estimate with correct sampling times should not overlap with those obtained with randomized sampling times. We also investigated the behavior of the test when the sampling times are not uniformly distributed throughout the tree, which sometimes occurs in empirical data sets. The test performs poorly in these circumstances, such that a modification to the randomization scheme is needed. Finally, we illustrate the behavior of the test in analyses of nucleotide sequences of *cereal yellow dwarf virus*. Our results validate the use of the date-randomization test and allow us to propose guidelines for interpretation of its results.

Key words: **molecular clock, date-randomization test, tip calibrations, virus evolution, Bayesian phylogenetics, time-structured sequence data.**

Introduction

The evolutionary rates and timescales of viruses can be estimated from molecular sequence data. This usually involves phylogenetic methods based on molecular clocks, which make assumptions about patterns of rate variation among lineages (recenly reviewed by Heath and Moore 2014 and Ho and Duchêne 2014). A key step is the scaling of the divergence times in the phylogeny into units of absolute time, a procedure known as "calibrating" the molecular clock. Calibrations typically involve constraining the age of one or more nodes in the tree using independent temporal data, such as dates from the fossil record. In viruses and some bacteria, however, nucleotide substitutions occur rapidly and appreciable amounts of genetic change can accumulate over the course of years or even months (Rodrigo and Felsenstein 1999; Drummond and Rodrigo 2000). In these cases, samples can be collected at different points in time to produce time-structured, or "heterochronous," data sets. The sampling dates can then be used to fix the ages of the tips in the tree, known as tip-calibrations (Rambaut 2000; Drummond et al. 2001, 2002; Drummond, Pybus, and Rambaut 2003; Ewing et al. 2004).

Phylogenetic analyses of time-structured sequence data have been critical in improving our understanding of virus evolution and emergence. For example, they have revealed that the substitution rates of some DNA viruses overlap with those of RNA viruses (Shackelton et al. 2005; Duffy and Holmes 2008, 2009). This challenges the expectation that DNA viruses should evolve more slowly than their RNA counterparts, because the latter depend on an error-prone polymerase to replicate without the assistance of proof-reading mechanisms (Furió et al. 2005). Tip-calibrations are also valuable for inferring the epidemiological dynamics of infectious disease (Pybus and Rambaut 2009), including such key measures as the basic reproductive number (R_0) and population growth rate (Stadler et al. 2012), as well as identifying chains of transmission (Lewis et al. 2008; Leventhal et al. 2014).

The tip-calibrations are the most important factor affecting the reliability of estimates of viral evolutionary rates and timescales. Clearly, the sampling dates need to have sufficient temporal spread to capture a measurable amount of evolutionary change (Drummond, Pybus, Rambaut et al. 2003). Tip-calibrations appear to be effective in analyses of empirical data and of data simulated under simple evolutionary scenarios (Seo et al. 2002; Ho et al. 2007; Molak et al. 2013). However, the reliability of the resulting estimates of rates and divergence times can be affected by the presence of population structure and fluctuations in population size (Navascués and Emerson

Article

2009), the distribution of sampling times (Ho et al. 2007, 2011; Firth et al. 2010), the information content of the data set (Debruyne and Poinar 2009; Ho et al. 2011), and the level of phylogenetic tree imbalance (Duchêne D, Duchêne S, et al. 2014). This points to an important role for methods of evaluating the reliability of inferences from time-structured data.

A simple method of validating estimates of rates and timescales is to fit a linear regression of the number of substitutions from the root of the tree to the tips as a function of the sampling times (Fitch et al. 1991). This method is often used as a diagnostic of the reliability of rate estimates, where the slope coefficient corresponds to the substitution rate under the assumption of a strict molecular clock, and R^2 indicates the degree to which sequence evolution has been clocklike (Korber et al. 2000). An important limitation of this method, however, is that the root-to-tip distances do not represent statistically independent samples. In addition, this method can produce spurious estimates of rates and timescales if the number of samples is small, such that the regression is based on few data points. Moreover, even if the regression points to a substantial departure from a strict molecular clock, the use of relaxed-clock models might satisfactorily accommodate rate variation among branches (Firth et al. 2010).

A more general method of validating estimates of rates and timescales is to investigate the extent of temporal signal within a data set, which is typically done using the date-randomization test (Ramsden et al. 2008). This test involves randomly reassigning the sampling times of the sequences, which effectively breaks the association between substitutions and time. This procedure is repeated a number of times, generating an expectation of rate estimates in the absence of temporal signal in the data. A widely used criterion for determining whether there is sufficient temporal signal is to verify that the mean rate estimated with the correct sampling times is not contained within any of the 95% credible intervals of those estimated from the date-randomized data sets (Firth et al. 2010). A more stringent criterion is to verify that there is no overlap between the 95% credible interval of the original rate estimate and any of those from the date-randomized data sets (Duffy and Holmes 2009; Ramsden et al. 2009). We refer to these criteria as CR1 and CR2, respectively. The number of date randomizations used for the test varies between studies, from 5 (Silva et al. 2012) to 20 (Kerr et al. 2012), but a large number of randomizations is probably needed for the test to be reliable.

Although the date-randomization test has been widely adopted, its statistical properties are not well understood and its performance has not been rigorously evaluated, such that its error rate is unknown. The efficacy of the test can be understood in terms of type I and type II errors; a type I error occurs when a data set with no temporal signal passes the test, whereas a type II error occurs when a data set with sufficient temporal signal fails the test. Generally, an ideal test should have low rates of type I and type II errors. In the context of molecular dating, however, type I errors are of special concern because they can mislead interpretations of estimates of substitution rates and timescales.

The performance of the test can depend on the number of date-randomizations and on different characteristics of the data, such as the number of variable sites, the sampling timeframe (calibration window), the level of rate variation among branches, and the temporal and spatial structure of the sampling times. For example, consider a data set consisting of samples that were collected at different times from two distinct subpopulations, such that the ages of samples are the same within subpopulations but differ between subpopulations. The random assignment of dates from different subpopulations to each of the samples will produce rate estimates that are considerably different from those obtained with the correct sampling times, regardless of whether the data have sufficient temporal signal.

Here, we investigate the performance of the date-randomization test using data generated by simulation. Using a simulation framework allows a broad range of conditions to be explored while controlling for the confounding effects of model misspecification. Furthermore, the true parameter values are known with certainty, making the calculation of error rates straightforward. We also illustrate the behavior of the date-randomization test in a case study of nucleotide sequences from *cereal yellow dwarf virus* (CYDV). For this data set, we show the effect of reducing the width of the calibration window and the number of variable sites. Based on the results of our analyses, we provide guidelines and improvements for the practical use and interpretation of the date-randomization test.

RESULTS

To assess the performance of the date-randomization test, we simulated the evolution of sequences of 2,000 nt along trees with root-node ages of 100 years. We generated the data under nine rate treatments using the uncorrelated lognormal relaxed-clock model (Drummond et al. 2006). The treatments involved three values for the mean rate: 1×10^{-3} (high), 5×10^{-4} (medium), and 1×10^{-4} (low) subs/site/year. We chose these values of the rates to represent the range of estimates from typical RNA virus data sets (Duchêne, Holmes, et al. 2014). The treatments also involved three levels of among-linage rate variation: 0%, 5%, and 20% of the mean. In addition, we specified different ages for the tips to vary the width of the calibration window, from narrow (0–5 years) to wide (60 years). We generated 40 data sets under each rate treatment. All of the data sets were analyzed using a Bayesian phylogenetic method implemented in BEAST v2.1 (Bouckaert et al. 2014). For each synthetic data set, we conducted 20 date-randomizations. We used two criteria for the date-randomization test, CR1 and CR2 (described above), to determine whether the rate estimates passed or failed the test. To assess the performance of the test, we calculated the number of type I and type II errors for both criteria. In our simulation framework, we consider that a data set has sufficient temporal signal if the estimate of the substitution rate is accurate.

Accuracy of Rate Estimates in Simulations

The substitution rate used in our simulations was an important determinant of the reliability of rate estimates. Analyses of the data sets simulated with high substitution rates (1×10^{-3} subs/site/year) consistently produced accurate rate estimates (fig. 1, panels *A*1–*A*3), with 95% credible intervals that always contained, or were within one standard deviation of, the true mean rate. In contrast, analyses of the data simulated with substitution rates that were medium (5×10^{-4} subs/site/year) or low (1×10^{-4} subs/site/year) values often yielded overestimates of the rate (fig. 1, panels *B*1–*B*3 and *C*1–*C*3). The mean numbers of variable sites for the data simulated with high, medium, and low rates were 1,200 (60%), 700 (35%), and 176 (8.8%), respectively.

Analyses of the data sets with narrow calibration windows produced estimates that were less accurate and precise than those with wide calibration windows (fig. 1). This pattern was most pronounced for the data simulated using medium and low substitution rates. In these cases, analyses of the data sets with narrow calibration windows produced mean rate estimates that were up to eight times as high as those used to generate the data (fig. 1, columns B and C).

The degree of rate variation among branches did not appear to have a strong impact on the accuracy of the estimates, suggesting that our choice of a lognormal relaxed-clock model performed well under the simulation conditions. For example, analyses of the data simulated with a high rate always yielded estimates with 95% credible intervals that contained, or were within one standard deviation of, the true mean rate, regardless of the level of rate variation among branches (fig. 1, panels *A*1–*A*3).

Performance of the Date-Randomization Test

The performance of the date-randomization test depended on the criterion used. Many data sets with no temporal signal, and with correspondingly inaccurate rate estimates, passed the test according to CR1, with a large number of type I errors. CR2 was preferable in the sense that it resulted in few type I errors. However, the improved detection of inaccurate estimates when using CR2 came at the expense of many accurate estimates failing the test, with higher numbers of type II errors than when using CR1.

Our simulation study illustrates the trade-off between the two types of errors for CR1 and CR2. We report the number of errors per rate treatment, each of which includes the rate estimates from 40 data sets, each with 20 randomizations for the date-randomization test. In the simulations with low mean rate and medium rate variation, six inaccurate rate estimates passed the test for CR1, whereas all of the inaccurate estimates failed the test according to CR2. In contrast, only one accurate estimate failed the test according to CR1, but six accurate estimates failed the test according to CR2 (fig. 1, panel *C*2). We do not consider type I errors for the data simulated with high rates because the estimates always included the true rate (fig. 1, panels *A*1–*A*3).

In most of our analyses of simulated data, the use of CR2 led to zero type I errors. The exceptions were three data sets simulated with low mean rate and zero rate variation (fig. 1, panel *C*1), and one data set simulated with a low mean rate and high rate variation (fig. 1, panel *C*3). The mean rate estimated from these data sets was up to 7.5 times as high as that used to generate the data, which is consistent with their narrow calibration window (<10 years).

The largest number of type II errors was 14, for CR2 in the data simulated with a high mean rate and high rate variation (fig. 1, panel *A*3), and for those with a medium mean rate and medium rate variation (fig. 1, panel *B*2). The smallest number of type II errors was 1, for CR1 in the simulations with low mean rate and medium rate variation (fig. 1, panel *C*2). This result echoes the trade-off between type I and type II errors for CR1 and CR2. Importantly, both types of errors were low when the calibration window was very wide. In all simulations, analyses of the data sets with calibration windows of at least 30 years ($\sim 10^{1.5}$ in fig. 1) produced accurate rate estimates, most of which passed the date-randomization test.

We used 20 randomizations for each data set. However, the number of randomizations needed to detect inaccurate estimates was highly variable. In figure 2, we show the rate estimates with the correct sampling times and those from the randomizations for a subset of ten data sets that produced inaccurate rate estimates. In some cases, the 95% credible interval of the rate with the correct sampling times overlapped with those of 17 randomized replicates (e.g., replicate 10, fig. 2). However, in other cases it did not overlap with those of any of the date-randomized replicates, such that 20 randomizations were not sufficient to detect the lack of temporal signal in these data (e.g., replicate 9, fig. 2). These results indicate that using a large number of randomizations and preferring CR2 over CR1 is critical for reducing type I errors (fig. 3). In particular, conducting as many as 20 date-randomizations might be insufficient in some cases.

Effect of Nonuniform Temporal Sampling on the Date-Randomization Test

We evaluated whether the performance of the date-randomization test was affected by nonuniform distribution of sampling times. We simulated phylogenetic trees with a root-node age of 100 years. The sampling times could take five values: 0, 1, 2, 3, and 4 years. We simulated a total of 20 trees. For ten of these trees, we assumed that the samples with the same sampling times were also closely related, forming monophyletic clusters. This situation represents the expected pattern when there is a strong association between spatial and temporal structure. For the remaining ten trees, we relaxed this assumption: the tips could also take any of the five sampling times, but they did not form monophyletic groups. We simulated these data under an uncorrelated lognormal relaxed clock, using a low mean substitution rate (1×10^{-4} subs/site/year) and a standard deviation of 5% of the mean rate. We simulated the evolution of sequences of 2,000 nt along these trees. The combination of a low mean substitution rate and a narrow calibration window is expected to produce data sets with no temporal signal, leading to

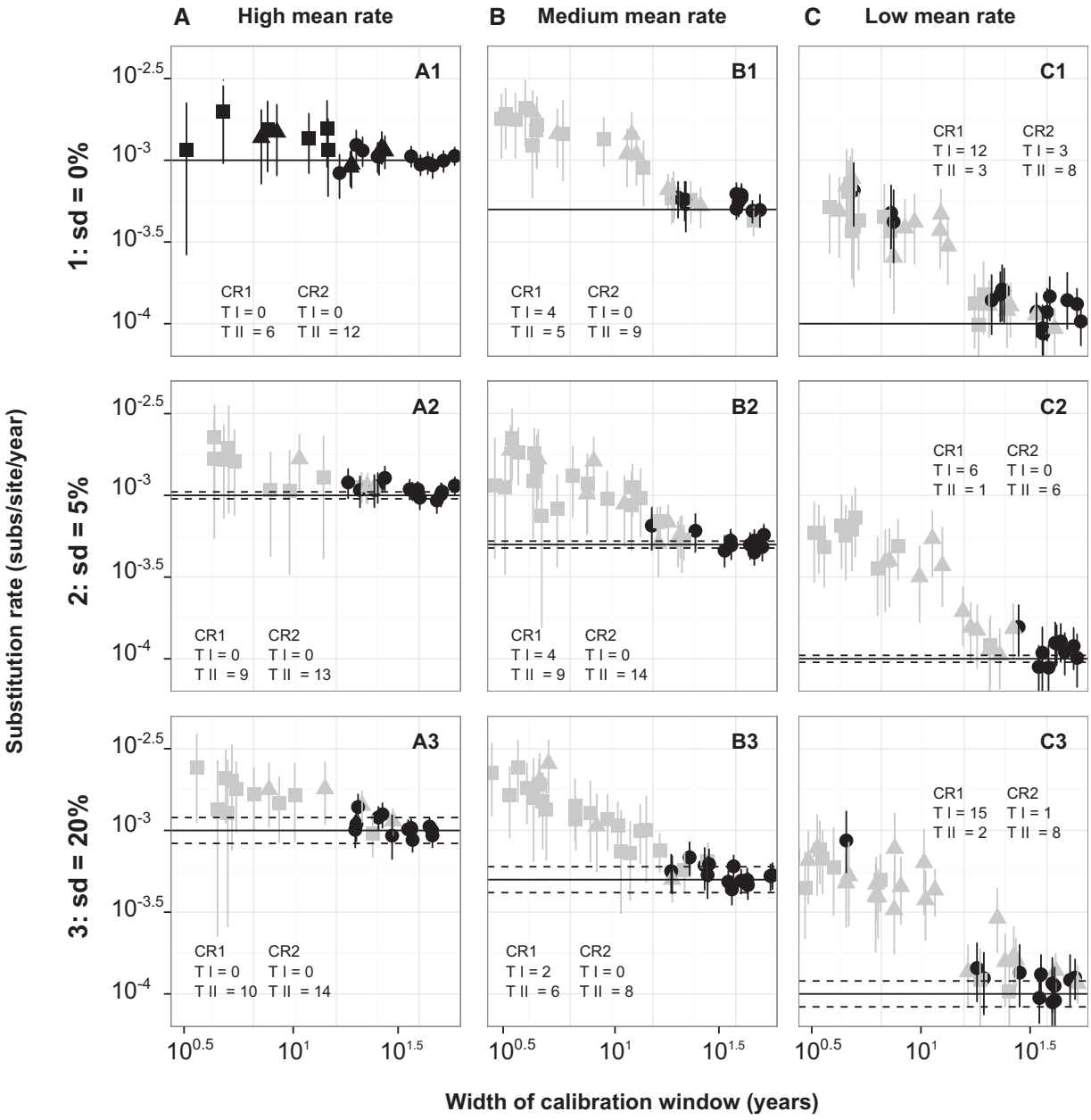

FIG. 1. Estimates of the substitution rate (subs/site/year) against the width of the calibration window (years) for different rate treatments. The axes are in $\log_{10}$ scale. The columns (*A*–*C*) represent data simulated using different mean substitution rates: high, 1×10^{-3} (*A*); medium, 5×10^{-4} (*B*); and low, 1×10^{-4} (*C*). The rows correspond to different levels of rate variation among branches used in the simulations, defined as the standard deviation of rates in the lognormal relaxed-clock model: 0% (strict clock) (1), 5% (2), and 20% (3) of the mean. Solid horizontal lines indicate the mean rate used in the simulations, whereas the dashed horizontal lines correspond to 1 standard deviation on either side of the mean. Symbols represent the mean rate estimate for each simulation, with the error bars showing the 95% credible intervals. Symbols in black represent rate estimates that passed the date-randomization test according to criteria CR1 and CR2. We conducted 20 randomizations for the date-randomization test for all data sets. We consider that rate estimates fail the test according to CR1 if the mean rate estimated with the correct sampling time is contained within the 95% credible interval of that obtained with any of the 20 randomizations. With CR2, rate estimates fail the test if the 95% credible interval with the correct sampling times overlaps with that from any of the 20 randomizations. Rate estimates that failed the date-randomization are shown in gray. Squares denote rate estimates that failed the test according to both CR1 and CR2, whereas triangles denote those that failed according to CR2 only. Numbers of type I and type II errors are shown for each rate treatment. A type I error occurs when the estimate from a data set with no temporal signal passes the test. A type II error occurs when the estimate from a data set with sufficient temporal signal fails the test.

inaccurate rate estimates. To analyze these data, we used the same Bayesian phylogenetic method as that for our simulations described above.

We obtained similar results from the analyses of the 20 simulated data sets with nonuniform distributions of sampling times. The substitution rate was always overestimated by more than one order of magnitude, which is consistent with a lack of temporal signal in these data sets (figs. 4 and 5). We used a modification of the date-randomization test, whereby the sampling times are grouped into clusters.

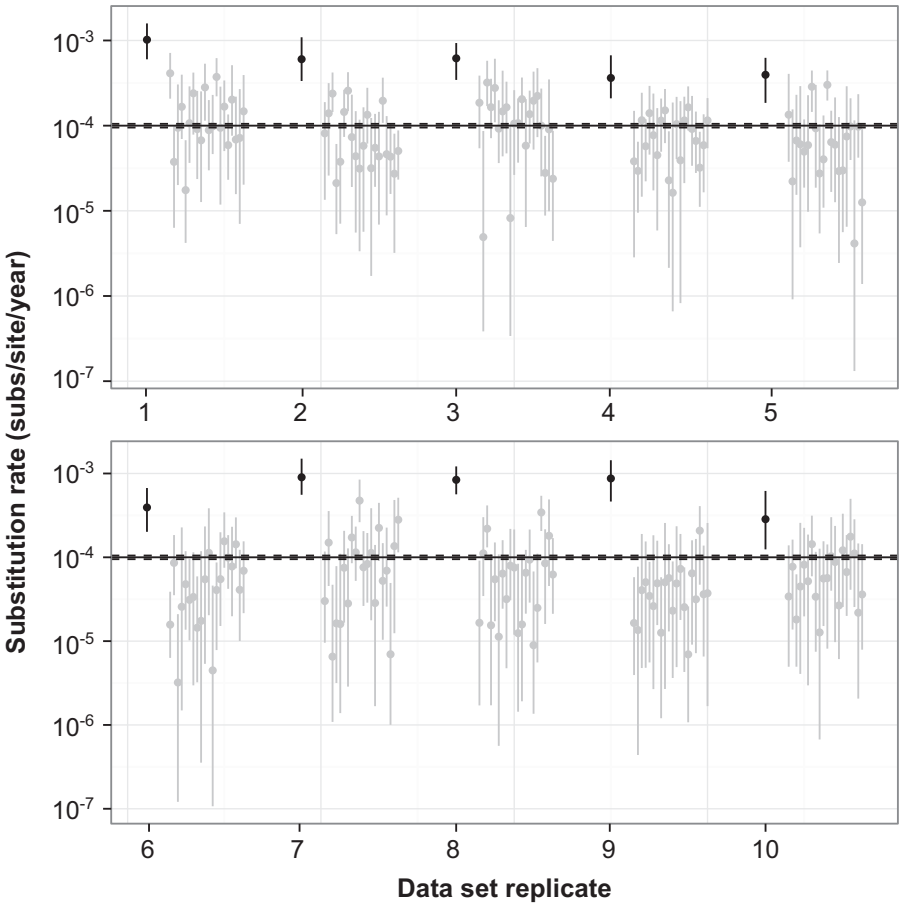

Fig. 2. Estimates of substitution rate (subs/site/year) for a subset of ten data sets that produced inaccurate rate estimates. The *y* axis corresponds to the rate in $\log_{10}$ scale and the *x* axis indicates different replicate data sets. The solid horizontal lines denote the mean rate used to generate the data, and the dashed horizontal lines correspond to 1 standard deviation on either side of the mean. The solid circles represent the mean estimates, and the errors bars show the 95% credible interval. Symbols in black correspond to the estimates obtained with the correct sampling times, whereas those in gray were obtained by randomizing the sampling times, with 20 randomizations per data set.

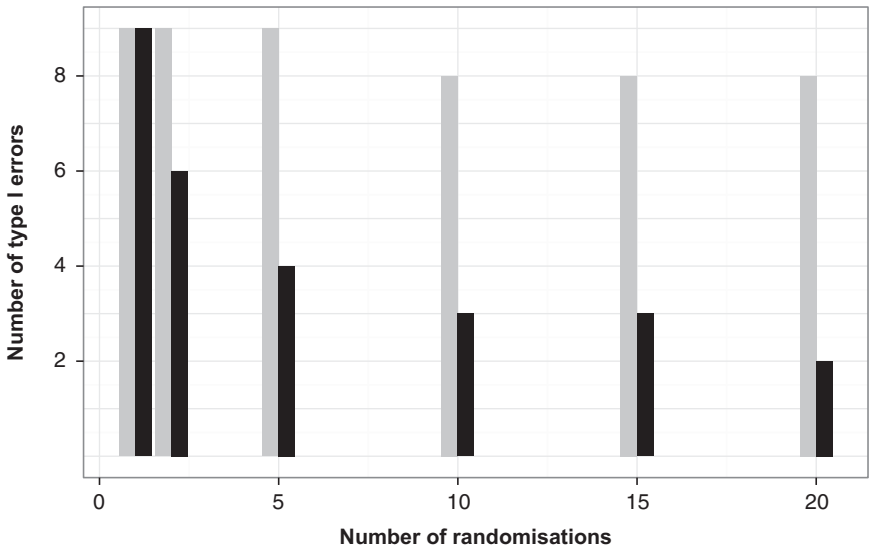

Fig. 3. Number of type I errors plotted against number of randomizations for the data sets shown in figure 2. The gray bars are the number of type I errors according to CR1, whereas the bars in black are those according to CR2. To compute the number of errors as a function of the number of randomizations, we select a given number of randomizations for all data sets and count the errors. For example, we select one randomization for each of the data sets in figure 2 and calculate the number of type I errors according to CR1 and CR2.

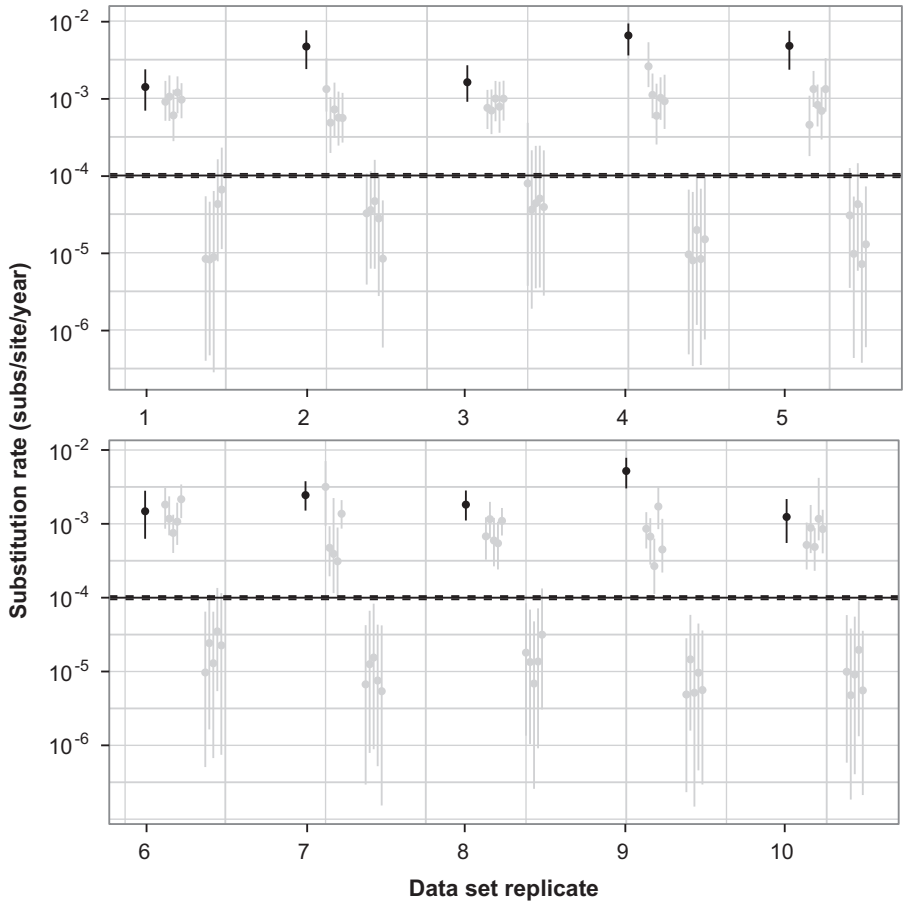

Fig. 4. Estimates of substitution rate (subs/site/year) for ten replicate data sets simulated with nonuniform temporal sampling and for which samples with identical sampling times form monophyletic groups. The *y* axis is in $\log_{10}$ scale. Symbols and error bars in black correspond to the estimates with the correct sampling times, whereas those in gray correspond to the estimates from date-randomized data sets. Each panel contains five simulation replicates. For each simulation replicate, the first five gray points correspond to the estimates obtained using cluster randomizations, while the remaining five were obtained using the standard date-randomization method. The solid black horizontal line is the mean substitution rate used in the simulations, whereas the dashed lines correspond to the standard deviation of 5% on either side of the mean.

We refer to a "sampling cluster" as a set of samples that share the same sampling time. In this respect, each of our simulated data sets consists of five sampling clusters. Our modification to the date-randomization test involved randomizing the sampling times among the sampling clusters, but not for samples within a cluster. For example, the sequences with sampling times of 3 years might all be assigned a sampling time of 1 year. We conducted the date-randomization test with five cluster randomizations. For comparison, we also conducted the test using the standard method of randomizing all of the sampling times. All of the estimates of substitution rates failed the cluster-based date-randomization test according to CR2, and sometimes also according to CR1. In contrast, the standard method of randomizing the sampling times was ineffective; the rate estimates passed the test according to CR1 and CR2 for most of the simulations, resulting in type I errors (figs. 4 and 5). The exception was a data set simulated with nonmonophyletic sampling clusters, for which the rate estimate failed the date-randomization test with both the standard and the cluster-randomization method, according to CR2 (replicate 6, fig. 5).

Case Study of CYDV

We performed a Bayesian phylogenetic analysis of a data set comprising 76 sequences of the coat protein gene from CYDV, sampled from 1925 to 2005. The mean rate estimate was 3.83×10^{-3} subs/site/year, with a 95% credible interval of 9.17×10^{-4} to 1.42×10^{-3} subs/site/year. This estimate passed the date-randomization test according to CR1 and CR2 (fig. 6), which is consistent with the results of previous studies of these data (Pagán and Holmes 2010; Duchêne, Holmes, et al. 2014). We then removed 38 samples to reduce the calibration window from 80 years to 1 year, retaining only the sequences with sampling dates from 2003 to 2004. For comparison, we also removed 38 randomly selected sequences while maintaining the calibration window of 80 years and the original root-node age. The mean rate estimate for the data set with randomly removed samples was similar to that for the complete data set (4.01×10^{-3} subs/site/year). There was greater uncertainty in the estimate, however, with a 95% credible interval of 6.20×10^{-4} to 1.73×10^{-2} subs/site/year. This estimate also passed the date-randomization test according to both CR1 and CR2. In contrast, reducing the

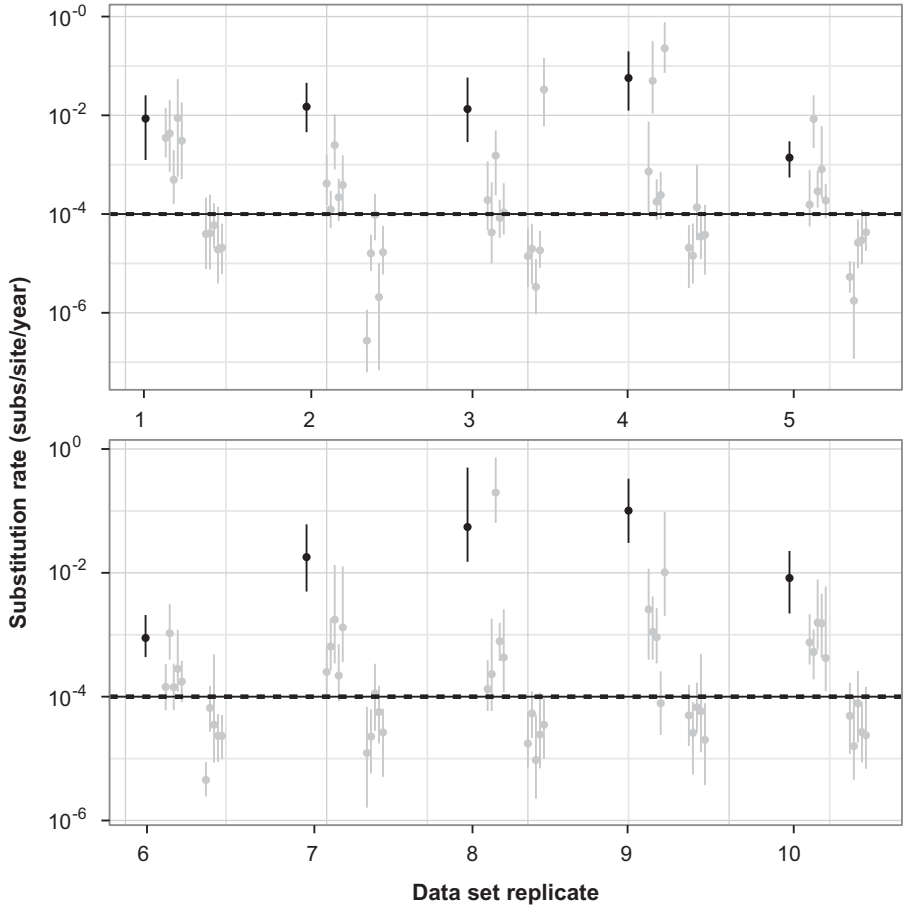

FIG. 5. Estimates of the substitution rate (subs/site/year) with nonuniform temporal sampling, but for which the samples with identical sampling times do not form monophyletic groups. The axes and symbols correspond to those of figure 4.

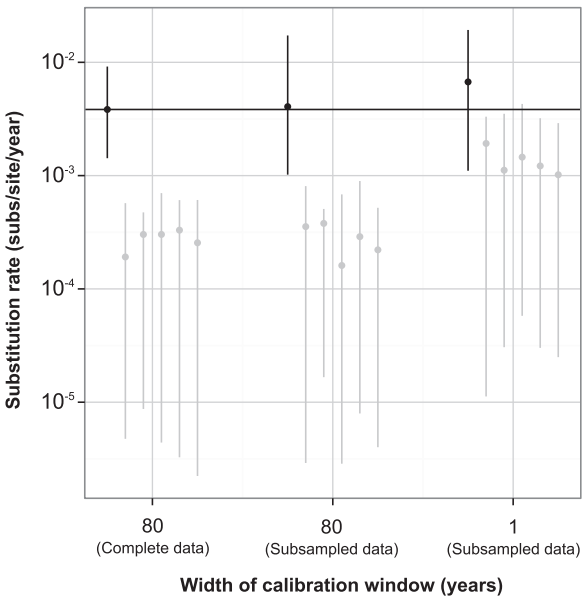

FIG. 6. Estimates of the substitution rate (subs/site/year) for the CYDV data set with different widths for the calibration window. The *y* axis is in $\log_{10}$ scale. Data points in black represent rate estimates obtained with the correct sampling times, whereas those for the date-randomizations are shown in gray. Filled circles represent mean rate estimates and error bars show the 95% credible intervals. The black horizontal line represents the mean rate estimate for the complete data set.

calibration window of the CYDV data set to 1 year resulted in a rate estimate that was higher and had greater uncertainty than that for the complete data. The mean rate estimate was 6.71×10^{-3} subs/site/year, with a 95% credible interval of 1.11×10^{-3} to 1.92×10^{-2} subs/site/year. This rate estimate failed the date-randomization test according to CR2, but not according to CR1 (fig. 6).

We investigated the effect of reducing sequence variation on the outcome of the date-randomization test. To do this, we obtained five data sets from which we removed 117 of the 129 variable sites, resulting in a reduction in sequence variation of 90%. We do not expect the rate estimates from these data sets to match that from the complete data because they have different levels of sequence variation. However, these data sets serve to illustrate the behavior of the date-randomization test when sequence variation is low. To enable comparison of the estimates from these data sets with those from data sets of similar sequence length, we also generated five data sets from which 117 randomly selected sites were removed.

We found considerable variation in the rate estimates from data sets with reduced sequence variation, with mean rate estimates that differed by as much as an order of magnitude. Four of these rate estimates failed the date-randomization test according to both CR1 and CR2 (fig. 7*A*). One rate

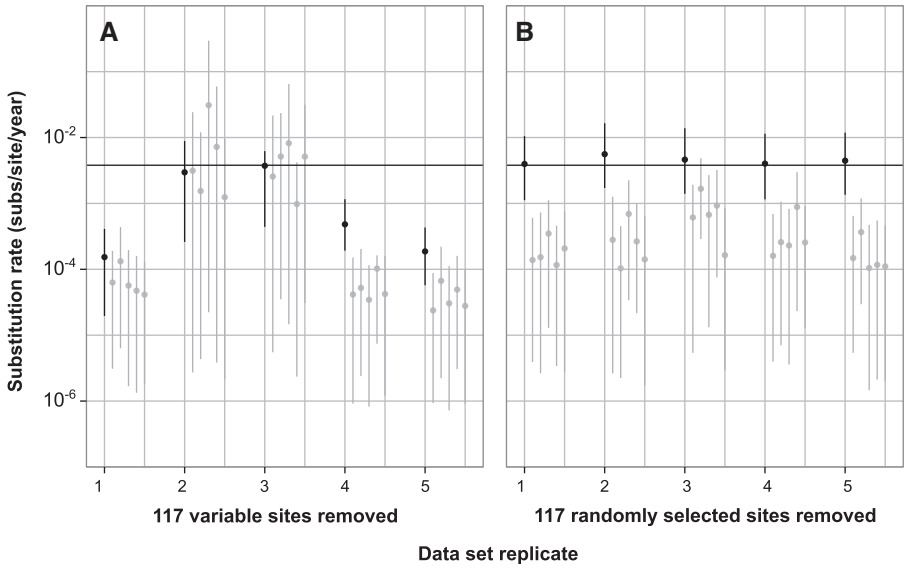

FIG. 7. Estimates of the substitution rate (subs/site/year) for the CYDV data set with reduced numbers of variable sites. The *y* axis is in $\log_{10}$ scale. Panel (*A*) shows five data sets from which 117 variable sites were removed, resulting in lower sequence variation than in the complete data. Panel (*B*) shows the estimates obtained for five data sets from which 117 sites were randomly removed. Data points in black represent rate estimates obtained with the correct sampling times, whereas those in gray represent rate estimates for the date-randomized data sets. Filled circles correspond to mean rate estimates and error bars show the 95% credible intervals. The black horizontal line represents the mean rate estimate for the complete data set.

estimate failed the test according to CR2 only, and its 95% credible interval overlapped with that of only one of the five date-randomizations (fig. 7A, replicate 4). This result illustrates the fact that using only a small number of randomizations can result in type I errors. Conversely, the rate estimates for data sets from which we randomly removed sites were very similar to that of the complete data set, but only two of these data sets passed the date-randomization test according to CR1 and CR2 (fig. 7*B*).

Discussion

Our simulation study shows that tip-calibrations provide a useful and effective source of information for analyses of virus data. This is consistent with previous validations of the use of time-structured data for estimating rates (e.g., Ho et al. 2007, 2011; Firth et al. 2010), but here we have confirmed that reliable rate estimates can be obtained even when there is substantial rate variation among branches. Our results also demonstrate that rates can be estimated reliably when there are small numbers of variable sites, provided that there is a wide span in sampling times. For example, a calibration window of at least 5 years was sufficient for obtaining accurate rate estimates from our data simulated with a high mean substitution rate (1×10^{-3} subs/site/year). For the data simulated with medium (5×10^{-4} subs/site/year) and low (1×10^{-4} subs/site/year) mean substitution rates, calibration windows of 13 and 30 years, respectively, appeared to be sufficient to obtain accurate rate estimates. The expected number of substitutions that accumulated over the timescale of the calibration window can be calculated as the product of the calibration window time, the substitution rate, and the alignment length. In this respect, the simulations that produced accurate estimates had calibration windows that accumulated at least between 5 and 20 substitutions. However, determining whether the samples constitute "measurably evolving populations" (Drummond, Pybus, Rambaut, Forsberg, et al. 2003), for which substitution rates and evolutionary timescales can be reliably estimated, can depend on a combination of other factors. These include the evolutionary timescale of the data set, the number of variable sites, and the degree of among-lineage rate variation. For this reason, the date-randomization test is a valuable tool that can be easily applied to investigate the extent of temporal signal in the data.

In most of our simulations, we found that the data sets that produced inaccurate rate estimates tended to overestimate rather than underestimate the rate. A possible reason for this pattern is that narrow calibration windows are uninformative, leading to a large uncertainty and an upward bias in the estimate of the rate (Ho et al. 2007; Debruyne and Poinar 2009). Our simulation study shows that this pattern of overestimation disappears when the calibration window is sufficiently wide.

The results of our analyses indicate that the more conservative criterion described here, CR2, should be applied when using the date-randomization test. In our simulation study, this criterion allowed us to detect nearly all of the inaccurate estimates of substitution rates. Although CR2 has the disadvantage of rejecting many rate estimates from data sets with sufficient temporal signal, we consider that this is preferable to interpreting estimates from data sets with no temporal signal as being correct, which is more likely to occur when using CR1.

Reducing the number of variable sites had an impact on the performance of the date-randomization test. The only cases in which inaccurate rate estimates passed the test

with CR2 was for a few data sets simulated with low rates, and with correspondingly small numbers of variable sites (fig. 1, panels C1 and C3, and fig. 2, replicate 9). In practice, it is difficult to establish whether a data set has a "small" number of variable sites. For example, there are 129 variable sites in the CYDV data set, which appears to have sufficient temporal signal. In our simulation study, however, some inaccurate rate estimates that passed the date-randomization test had around 176 variable sites. In this respect, a reassuring result is that the test had no type I errors and very few type II errors for data sets with very wide calibration windows, regardless of the number of variable sites or the amount of rate variation among branches. This underscores the importance of drawing samples from as wide a time span as possible.

Conducting large numbers of date-randomizations is important for detecting inaccurate rate estimates. In our simulation study, the test performed well with 20 randomizations, but there were still some inaccurate rate estimates that passed the test. The number of date-randomizations required to eliminate these type I errors depends on the data set, but in general it is advisable to conduct at least 20 date-randomizations to improve the reliability of the test.

A critical aspect of phylogenetic methods that use tip-calibrations is that they typically assume random sampling and the absence of population structure. Severe deviations from these assumptions can have a substantial impact on the estimates of substitution rates and evolutionary timescales (Navascués et al. 2010). In practice, data sets often violate these assumptions because sampling is conducted in a nonrandom manner, resulting in a nonuniform distribution of sampling times. Sometimes this can also result in an association between temporal and phylogenetic clustering. We find that the standard method of randomizing the sampling times is ineffective in these cases, causing the date-randomization test to have a high probability of type I errors. This probably occurs because there are very few distinct sampling times relative to the number of sequences. If the standard randomization method is used, a large portion of sequences will be reassigned their correct sampling time. The rate estimates from these "partial" randomizations will tend to be very different from those obtained with the correct sampling times, regardless of whether they have sufficient temporal signal or not. This problem can be solved by identifying clusters of samples with the same, or very similar, sampling times. The date-randomizations are performed by randomizing the sampling times among, but not within, these clusters. Our guidelines for the date-randomization test are also applicable to the cluster-randomization approach. However, it is important to note that conducting a large number of cluster-randomizations is only possible when there are multiple sampling clusters. For example, in a data set comprising only two sampling clusters, only a single randomization is feasible.

We suggest that the date-randomization test should be applied routinely in studies that use tip-calibrations. This comes with a caution that the performance of the test depends on the degree of temporal clustering. Wider application of the test might reveal that data sets for some rapidly evolving pathogens do not have sufficient temporal signal, indicating that their estimates of rates and timescales might be inaccurate. For rate estimates that fail the test, the sampling strategy should be modified to include older samples, if possible. In a study of *Human immunodeficiency virus* 1, the inclusion of a single molecular sample from 1959 suggested that the substitution rate of this virus was lower than previously reported, implying that its emergence in humans occurred around 1920 (Worobey et al. 2008). Another strategy is to include calibrations for internal nodes, particularly those with deep positions in the tree (Duchêne, Lanfear, et al. 2014). For example, the use of calibrations based on island biogeography revealed that the long-term substitution rate for some *Simian immunodeficiency virus* lineages might have been overestimated previously, such that the timescale for these viruses is at least 30,000 years (Worobey et al. 2010). This result stands in contrast with those obtained using only using tip-calibrations, which yielded estimates of a higher rate and a viral evolutionary timescale of only a few centuries (Wertheim and Worobey 2009).

Our study has focused on Bayesian phylogenetic analysis of time-structured data using popular models of rate variation and demographic history, but these data can be analyzed using a wide range of other phylodynamic models (Kühnert et al. 2011). Accurate rate estimation is important in these complex models (Ho and Shapiro 2011; Hedge et al. 2013), but their sensitivity to the temporal structure in the data is unknown. The date-randomization test is also commonly employed in studies of ancient DNA sequences from animals and plants, which tend to evolve much more slowly than viruses but can include samples from much wider calibration windows (Miller et al. 2009; Ho et al. 2011). These data sets sometimes have large numbers of modern sequences and a small set of ancient sequences, so it has been suggested that it is more appropriate to randomize the ancient sequences only (Miller et al. 2009). Simulation frameworks similar to the one we have presented here will be useful for investigating these questions and different methods for validating estimates of evolutionary rates and time frames.

Materials and Methods

Simulations

We generated phylogenetic trees with 50 taxa and with a root-node age of 100 years. The ages of the tips were uniformly distributed between 0 and 5, 15, 30, or 60 years. This represents a range of calibration windows used in studies of recently emerging RNA viruses, such as influenza viruses (e.g., Smith et al. 2009) and human immunodeficiency virus (e.g., Worobey et al. 2008). We used BEAST to generate the trees for simulating sequence evolution. We fixed the ages of the tips and the root-node and assumed a demographic model with constant population size. We specified these settings in an input file for BEAST and conducted the analyses without sequence data. This allowed us to sample trees from the prior

distribution under the chosen model. These trees are chronograms, in which the branch lengths are expressed in years.

To generate the branch-specific substitution rates for the chronograms, we used the R package NELSI v0.21 (Ho et al. 2014). In this program, the branch lengths of the chronograms are multiplied by the rate to produce phylograms, with branch lengths in units of substitutions per site (subs/site). We chose three values for the mean substitution rate, corresponding to some of those estimated from RNA viruses: 1×10^{-3} (high), 5×10^{-4} (medium), and 1×10^{-4} (low) subs/site/year. Simulations using these rates produce sequence data with different proportions of variable sites. We also specified different levels of rate variation among branches by drawing the branch rates from a lognormal distribution, which is equivalent to the uncorrelated lognormal relaxed molecular clock (Drummond et al. 2006). We set the mean rate to the high, medium, or low values described above, and set the standard deviation as 0% (strict clock), 5%, or 20% of the mean.

We simulated sequence evolution along the trees to produce alignments of 2,000 nt, under the Jukes–Cantor substitution model, using the R package phangorn v1.9 (Schliep 2011). We chose this substitution model to avoid the need to choose a large number of parameter values for more complex substitution models in the simulations, and because the focus of our study does not involve substitution model misspecification. In total, our simulations included nine rate treatments and four different widths for the calibration window. We generated 40 data sets for each rate treatment, with 10 data sets for each calibration width, for a total of 360 data sets.

Phylogenetic Analyses

We analyzed the sequence data using the Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST. We used the uncorrelated lognormal relaxed molecular clock model to accommodate rate variation among branches in our data, calibrated using the ages of the tips. We matched the demographic and substitution models to those used to generate the sequences. We used the "OneOnX" prior for the population size parameter of the constant population size demographic model, which takes the form $f(x) = C/x$, where C is a normalizing constant. We chose this prior because using a more informative prior, or fixing the population size to its true value, can result in an underestimation of the variance of the substitution rate (Debruyne and Poinar 2009; Ho et al. 2011).

We conducted 20 date-randomizations for the date-randomization test for each of the 40 data sets from each rate treatment. The analyses were run with an MCMC chain length of 1.5×10^7 steps, with samples from the posterior distribution drawn every 5×10^3 steps. After discarding the first 10% of steps as burn-in, we assessed sufficient sampling from the posterior by verifying that the effective sample sizes for all parameters were at least 200, using the R package CODA (Plummer et al. 2006).

Statistical Analyses

We compared two different criteria for the date-randomization test. For the first criterion, CR1, a rate estimate fails the test if the mean estimate obtained with the correct sampling times is contained within the 95% credible intervals of any of those obtained with the date-randomized data sets. The second criterion, CR2, is more conservative. A rate estimate fails the test if its 95% credible interval overlaps with those of any of the date-randomized data sets. We evaluated the performance of the test in terms of type I and type II errors. A type I error occurs when a data set with no temporal signal passes the test, and a type II error occurs when a data set with sufficient temporal signal fails the test. In our simulations, we consider that a data set has sufficient temporal signal if it produces an accurate rate estimate, so that its 95% credible interval includes the value used for simulation.

Effect of Nonuniform Temporal Sampling on the Date-Randomization Test

To investigate the effect of nonuniform temporal sampling on the date-randomization test, we generated data sets with low temporal signal. Analyses of these data sets are expected to produce unreliable rate estimates. Accordingly, they are useful for illustrating the probability of type I errors, which is an important concern for estimates of substitution rates and evolutionary timescales. We simulated phylogenetic trees with 50 taxa, a root-node age of 100 years, and a constant population size by sampling from the prior distribution in BEAST. We considered two scenarios of nonuniform temporal sampling, and simulated ten trees in each case. In the first scenario, there is a strong association between sampling time and genetic divergence, such that samples with the same age form monophyletic clusters. To simulate these data, we constrained monophyly for five clusters of tips, with each cluster comprising ten tips. We assigned a single sampling time (0, 1, 2, 3, and 4 years) for all of the tips within each cluster. In the second scenario, we also considered five clusters of ten tips with the same sampling time, but the clusters did not form monophyletic groups. This latter scenario is similar to the "layered" sampling strategy described in previous studies (e.g., Seo et al. 2002; Ho et al. 2007).

The substitution rate varied among branches according to the uncorrelated lognormal relaxed-clock model with a low mean rate (1×10^{-4} subs/site/year) and a standard deviation of 5%. To simulate sequence evolution, we used the same method as that described for our simulations above. We analyzed the data in BEAST, with the sampling times used for calibration, and matched the substitution and demographic models to those used to generate the data.

We conducted the date-randomization test with five randomizations. We also used a modification of the test to account for nonuniform temporal sampling. In the modified test, the units of date-randomization are the sampling clusters, which are groups of sequences that share the same sampling time. Dates are randomized among, but not within, these sampling clusters. We conducted five

cluster-randomizations and used criteria CR1 and CR2 to determine whether the rate estimates passed or failed the test.

Case Study of CYDV

We downloaded a data set of 76 nt sequences from GenBank of the coat protein gene (*CP*) of CYDV, a positive-sense, single-stranded RNA virus. This CYDV data set, previously analyzed by Pagán and Holmes (2010), was convenient for our study because the sequences had a temporal and phylogenetic structure that allowed us to manipulate the calibration window while retaining the original root-node age. The sampling dates ranged from 1925 to 2005, providing a calibration window of 80 years. We aligned the sequences using the Muscle algorithm (Edgar 2004) and visually inspected the alignment. We analyzed the data in BEAST using a Bayesian skyline demographic model (Drummond et al. 2005). We selected the GTR+Γ substitution model according to the Bayesian information criterion using the software ModelGenerator v0.851 (Keane et al. 2006). The settings of the MCMC and assessment of sufficient sampling from the posterior distribution were the same as those used in our simulation study. We conducted a date-randomization test with five randomizations, and verified that the rate estimate passed the test according to CR1 and CR2. We conducted only a small number of randomizations because the rate estimates for this CYDV data set passed the date-randomization test in two previous studies (Pagán and Holmes 2010; Duchêne, Holmes, et al. 2014).

To investigate the effect of the width of the calibration window on the performance of the date-randomization test and on the estimate of the substitution rate, we reduced the calibration window to 1 year (between 2003 and 2004) by removing 38 samples while maintaining the age of the root-node. The estimates obtained with this subsampled data set are not directly comparable to those obtained with the complete set of samples because they differ in the number of sequences. To overcome this limitation, we also removed 38 sequences randomly, while maintaining the original calibration window of 80 years. We analyzed these data with the same method as for the complete data set, conducting five date-randomizations.

We assessed the impact of the number of variable sites on the date-randomization test. The complete CYDV data set contained 129 variable sites. We reduced the number of variable sites by 90% by randomly removing 117, a procedure that we repeated five times. The rate estimates for these subsamples of the data are expected to differ from that of the complete data set because they have different levels of sequence variation; however, the purpose of our experiment was to investigate the behavior of the date-randomization test, rather than to compare the rate estimates with that from the complete data set. For comparison with data sets with similar sequence lengths, we also obtained five data sets from which we removed 117 randomly selected sites. In this case, we expect the rate estimates to match that from the complete data set because the level of sequence variation is similar. We conducted five date-randomizations in each case, and analyzed the data set using the same settings as for the complete data set.

Acknowledgments

The authors thank several anonymous reviewers of previously submitted articles, who pointed out some of the potential limitations of the date-randomization test. S.D. was supported by a Francisco José de Caldas Scholarship from the Colombian government and by a Sydney World Scholars Award from the University of Sydney. D.D. was supported by an Australian National University HDR Merit Scholarship. E.C.H. was supported by a National Health and Medical Research Council Australia Fellowship (AF30). S.Y.W.H. was supported by the Australian Research Council (DP110100383).

References

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 10:e1003537.

Debruyne R, Poinar HN. 2009. Time dependency of molecular rates in ancient DNA data sets, a sampling artifact? *Syst Biol.* 58:348–360.

Drummond AJ, Forsberg R, Rodrigo AG. 2001. The inference of stepwise changes in substitution rates using serial sequence samples. *Mol Biol Evol.* 18:1365–1371.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 161: 1307–1320.

Drummond AJ, Pybus OG, Rambaut A. 2003. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol.* 54:331–358.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends Ecol Evol.* 18:481–488.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Drummond AJ, Rodrigo AG. 2000. Reconstructing genealogies of serial samples under the assumption of a molecular clock using serial-sample UPGMA. *Mol Biol Evol.* 17:1807–1815.

Duchêne D, Duchêne S, Ho SYW. 2014. Tree imbalance causes a bias in phylogenetic estimation of evolutionary timescales using heterochronous sequences. *Mol Ecol Resour.* Advance Access published November 27, 2014, doi: 10.1111/1755-0998.12352.

Duchêne S, Holmes EC, Ho SYW. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc R Soc Lond B Biol Sci.* 281:20140732.

Duchêne S, Lanfear R, Ho SYW. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol Phylogenet Evol.* 78:277–289.

Duffy S, Holmes EC. 2008. Phylogenetic evidence for rapid rates of molecular evolution in the single-stranded DNA begomovirus tomato yellow leaf curl virus. *J Virol.* 82:957–965.

Duffy S, Holmes EC. 2009. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *J Gen Virol.* 90:1539–1547.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

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.

Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A. 2010. Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol Biol Evol.* 27:2038–2051.

Fitch WM, Leiter JM, Li XQ, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. *Proc Natl Acad Sci U S A.* 88: 4270–4274.

Furió V, Moya A, Sanjuán R. 2005. The cost of replication fidelity in an RNA virus. *Proc Natl Acad Sci U S A.* 102:10233–10237.

Heath TA, Moore BR. 2014. Bayesian inference of species divergence times. In: Chen M-H, Kuo L, Lewis PO, editors. Bayesian phylogenetics, methods, algorithms, and applications. Boca Raton (FL): CRC Press. p. 277–318.

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

Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol.* 23:5947–5975.

Ho SYW, Duchêne S, Duchêne D. 2014. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol Ecol Resour.* Advance Access published August 23, 2014, doi: 10.1111/1755-0998.12320.

Ho SYW, Kolokotronis S-O, Allaby RG. 2007. Elevated substitution rates estimated from ancient DNA sequences. *Biol Lett.* 3:702–705.

Ho SYW, Lanfear R, Phillips MJ, Barnes I, Thomas JA, Kolokotronis S-O, Shapiro B. 2011. Bayesian estimation of substitution rates from ancient DNA sequences with low information content. *Syst Biol.* 60: 366–375.

Ho SYW, Shapiro B. 2011. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol Ecol Resour.* 11: 423–434 .

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* 6:29.

Kerr PJ, Ghedin E, DePasse JV, Fitch A, Cattadori IM, Hudson PJ, Tscharke DC, Read AF, Holmes EC. 2012. Evolutionary history and attenuation of myxoma virus on two continents. *PLoS Pathog.* 8:e1002950.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. *Science* 288:1789–1796.

Kühnert D, Wu CH, Drummond AJ. 2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infect Genet Evol.* 11: 1825–1841.

Leventhal GE, Günthard HF, Bonhoeffer S, Stadler T. 2014. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. *Mol Biol Evol.* 31:6–17.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Brown AJL. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 5:e50.

Miller HC, Moore JA, Allendorf FW, Daugherty CH. 2009. The evolutionary rate of tuatara revisited. *Trends Genet.* 25:13–15.

Molak M, Lorenzen ED, Shapiro B, Ho SYW. 2013. Phylogenetic estimation of timescales using ancient DNA: the effects of temporal sampling scheme and uncertainty in sample ages. *Mol Biol Evol.* 30:253–262.

Navascués M, Depaulis F, Emerson BC. 2010. Combining contemporary and ancient DNA in population genetic and phylogeographical studies. *Mol Ecol Resour.* 10:760–772.

Navascués M, Emerson BC. 2009. Elevated substitution rate estimates from ancient DNA: model violation and bias of Bayesian methods. *Mol Ecol.* 18:4390–4397.

Pagán I, Holmes EC. 2010. Long-term evolution of the Luteoviridae: time scale and mode of virus speciation. *J Virol.* 84:6177–6187.

Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* 6:7–11.

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

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

Ramsden C, Holmes EC, Charleston MA. 2009. Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Mol Biol Evol.* 26:143–153.

Ramsden C, Melo FL, Figueiredo LM, Holmes EC, Zanotto PMA. 2008. High rates of molecular evolution in hantaviruses. *Mol Biol Evol.* 25: 1488–1492.

Rodrigo AG, Felsenstein J. 1999. Coalescent approaches to HIV population genetics. In: Crandall KA, editor. The evolution of HIV. Baltimore (MD): Johns Hopkins University Press. p. 233–272.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593.

Seo TK, Thorne JL, 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.

Shackelton LA, Parrish CR, Truyen U, Holmes EC. 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci U S A.* 102:379–384.

Silva G, Marques N, Nolasco G. 2012. The evolutionary rate of citrus tristeza virus ranks among the rates of the slowest RNA viruses. *J Gen Virol.* 93:419–429.

Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, et al. 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* 459:1122–1125.

Stadler T, Kouyos R, von Wyl V, Yerly S, Böni J, Bürgisser P, Klimkait T, Joos B, Rieder P, Xie D, et al. 2012. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol.* 29:347–357.

Wertheim JO, Worobey M. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol.* 5:e1000377.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, Muyembe JJ, Kabongo JM, Kalengayi RM, Van Marck E, et al. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455:661–664.

Worobey M, Telfer P, Souquière S, Hunter M, Coleman CA, Metzger MJ, Reed P, Makuwa M, Hearn G, Honarvar S, et al. 2010. Island biogeography reveals the deep history of SIV. *Science* 329:1487.