Published online 3 September 2018

Nucleic Acids Research, 2018, Vol. 46, No. 22 ***e134***
doi: 10.1093/nar/gky783

Bayesian inference of ancestral dates on bacterial phylogenetic trees

Xavier Didelot[1,*], Nicholas J. Croucher[1], Stephen D. Bentley[2], Simon R. Harris[2] and Daniel J. Wilson[3]

[1]Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK, [2]The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK and [3]Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, UK

Received June 08, 2018; Revised July 26, 2018; Editorial Decision August 17, 2018; Accepted August 21, 2018

ABSTRACT

The sequencing and comparative analysis of a collection of bacterial genomes from a single species or lineage of interest can lead to key insights into its evolution, ecology or epidemiology. The tool of choice for such a study is often to build a phylogenetic tree, and more specifically when possible a dated phylogeny, in which the dates of all common ancestors are estimated. Here, we propose a new Bayesian methodology to construct dated phylogenies which is specifically designed for bacterial genomics. Unlike previous Bayesian methods aimed at building dated phylogenies, we consider that the phylogenetic relationships between the genomes have been previously evaluated using a standard phylogenetic method, which makes our methodology much faster and scalable. This two-step approach also allows us to directly exploit existing phylogenetic methods that detect bacterial recombination, and therefore to account for the effect of recombination in the construction of a dated phylogeny. We analysed many simulated datasets in order to benchmark the performance of our approach in a wide range of situations. Furthermore, we present applications to three different real datasets from recent bacterial genomic studies. Our methodology is implemented in a R package called BactDating which is freely available for download at https://github.com/xavierdidelot/BactDating.

INTRODUCTION

A population evolving sufficiently quickly over a sufficiently long sampling time frame is said to be 'measurably evolving', which means that it is possible to estimate the rates over time at which evolution operates and the dates at which ancestors existed (1). This concept has recently become applicable to bacterial species, following the advent of whole-genome sequencing data, in which the relatively low per site evolutionary rates in bacteria are compensated by long genomes, typically comprising millions of sites (2). Consequently, analytical methods that were previously the hallmark of viral genetics are growing in popularity in bacterial genetics, especially the estimation of dated genealogies through the application of the software BEAST (3–6). In a dated phylogeny (also sometimes known as a time-stamped phylogeny or time-calibrated phylogeny), the branch lengths are measured in unit of time (for example days or years), the leaves are shown at known dates of isolation, and the internal nodes are represented at the dates when common ancestors are estimated to have existed. Such estimation of ancestral dates can often provide direct biological insights, for example to date the emergence of an epidemiologically important lineage, but can also be used as a starting point for further analysis, for example to infer past population size dynamics (7), to reconstruct transmission events between hosts (8), to estimate the parameters of an epidemiological model (9), to investigate geographical range expansion (10) or to study ecological adaptation to host species (11).

The BEAST framework is popular because it includes many models and extensions, and is based on the Bayesian paradigm which enables a complete quantification of uncertainties in date estimates. However, it is sometimes too slow and computationally demanding to be used, especially when large numbers of sequences are involved. Alternatives based on optimization have therefore started to appear, including LSD (12) which uses least-square optimization methods and TempEst (13) which uses a linear regression to explore the temporal structure of the data. A systematic comparison between LSD, TempEst and BEAST reported that they produced highly congruent estimates of evolutionary rates (14). More recently, three new optimization methods have been released based on maximum likelihood, namely node.dating (15), treedater (16) and TreeTime (17). All these methods are faster than BEAST and

*To whom correspondence should be addressed. Tel: +44 2075 943622; Email: x.didelot@imperial.ac.uk

able to deal with larger datasets, in great part due to the fact that they assume that phylogenetic relationships have previously been assessed. Their input data therefore consists of the sampling dates plus an unrooted phylogenetic tree which needs to be built in a separate analytical step using a standard phylogenetic software such as RAxML (18), PhyML (19), FastTree (20) or IQ-TREE (21).

Here, we present a new methodology called BactDating for analyzing dated genetic data in order to estimate evolutionary rates and dated phylogenies in bacterial populations. We use a Bayesian framework for inference as in BEAST, but consider that phylogenetic relationships have been assessed in a previous step as in the optimization and maximum likelihood methods described above. This way we enjoy the benefits of Bayesian inference in ancestor dating (22), such as assessment of uncertainties and flexibility of model choice and comparison, but with a computational scalability and speed comparable to the optimization methods described above. Furthermore, we explore the specific problems posed by application in bacterial genomics, and in particular the disruptive effect that homologous recombination can have on estimates of the temporal signal (23,24). Recombination is well known for disrupting phylogenetic inference, and especially to affect branch lengths estimates so that trees look star-like with abnormally long terminal branches (23,25,26). To account for this, sites detected as recombinant are sometimes removed prior to running BEAST, but this approach is inefficient and can even exacerbate the problem (23). A more principled method is implemented in the Bacter package (27) which incorporates the ClonalOrigin model of bacterial recombination (28) within BEAST2 (5), but this approach is too computationally intensive to be applicable to large genomic datasets. Instead we show how the effect of recombination can be accounted for in the dating of ancestral nodes, by exploiting a scalable phylogenetic method that accounts for bacterial recombination such as ClonalFrameML (29) or Gubbins (30).

We applied BactDating to a large number of datasets simulated under various conditions in order to benchmark its ability to produce correct estimates by comparison with the correct parameter values used during simulation. We also demonstrate the usefulness of BactDating on three case studies based on real datasets from recently published bacterial genomic studies. The first case study used ancient DNA sequencing in order to compare medieval and modern genomes of the leprosy causing pathogen *Mycobacterium leprae* (31). In the second case study a large number of isolates from clonal lineage of *Shigella sonnei* from Vietnam were sequenced and compared to study local emergence and dissemination (32). Finally, in the third case study, a worldwide collection of genomes from a highly recombining lineage of *Streptococcus pneumoniae* were used to investigate its global success and spread (33).

MATERIALS AND METHODS

Overview of Bayesian inference

We consider as input a phylogenetic tree $\mathcal{P}$ previously estimated from a set of n bacterial genomes using a standard phylogenetic method. For ease of presentation, we initially make two simplifying assumptions that will be relaxed later. Firstly, we consider that all the isolation dates of the genomes are known. Secondly, we assume that the tree $\mathcal{P}$ is already rooted, so that it contains $b = 2n - 2$ branches. Our aim is to estimate a dated tree $\mathcal{T}$, which in this case means estimating the dates at which each of the $n - 1$ internal nodes in $\mathcal{P}$ existed. There are two key differences between the input phylogeny $\mathcal{P}$ and the target of inference, the dated or time-calibrated tree $\mathcal{T}$. First, the branch lengths of $\mathcal{P}$ are measured in units of the expected number of substitutions, whereas the branch lengths of $\mathcal{T}$ are measured in calendar time. Second, as a consequence, the 'heights' of all tips and internal nodes in $\mathcal{T}$ are directly interpretable as calendar dates, which is not true of $\mathcal{P}$.

Table 1. Table of symbols

Symbol	Description
$\mathcal{P}$	Input phylogenetic tree
n	Number of leaves in the phylogenetic tree $\mathcal{P}$
b	Number of branches in the phylogenetic tree $\mathcal{P}$
x_i	Length of the ith branch of the phylogenetic tree $\mathcal{P}$
$\mathcal{T}$	Dated phylogeny to be estimated
l_i	Duration of the ith branch of the dated phylogeny $\mathcal{T}$
Θ	Additional parameters to be estimated
α	Coalescent time unit
μ	Mean substitution rate
m_i	Substitution rate for the ith branch of the dated phylogeny $\mathcal{T}$
σ	Standard deviation of the per-branch substitution rates

To estimate the dated tree $\mathcal{T}$ in a Bayesian inferential framework, we need to specify a prior on $\mathcal{T}$ and the likelihood of observing the substitutions in $\mathcal{P}$ given the dated tree $\mathcal{T}$. For this likelihood, we will consider three models of increasing complexity: a strict clock model without recombination, a relaxed clock model without recombination and finally a strict or relaxed clock model with recombination. The main notation is summarized in Table 1.

More formally, we want to jointly infer the dated genealogy $\mathcal{T}$ and some additional model parameters Θ given an estimated phylogeny $\mathcal{P}$, so that the target distribution is:

$$p(\mathcal{T}, \Theta|\mathcal{P}) \propto p(\mathcal{P}|\mathcal{T}, \Theta)p(\Theta)p(\mathcal{T}|\Theta) \quad (1)$$

The first term $p(\mathcal{P}|\mathcal{T}, \Theta)$ is the likelihood, which is described in subsequent sections under various conditions. The second term $p(\Theta)$ represents the prior on the additional parameters in Θ and will also be described later. The third term $p(\mathcal{T}|\Theta)$ is the prior on the dated genealogy $\mathcal{T}$ for which we consider a coalescent model with constant population size (34), which is the genealogical process that corresponds to many forward in time population genetics model such as the standard neutral Wright-Fisher model. The only parameter of this model is the coalescent time unit $\alpha = N_e g$ which is the product of the effective population size N_e and generation time g. The parameter α is included in the vector Θ of parameters that we aim to co-estimate. This prior term $p(\mathcal{T}|\Theta)$ can be computed by considering the ordered list of $2n - 1$ times t_i of both terminal and internal nodes in the dated genealogy, and the values k_i of lineages existing in

each time interval, which gives (35):

$$p(\mathcal{T}|\Theta) = \frac{1}{\alpha^{n-1}} \prod_{i=2}^{2n-1} \exp\left(\frac{-k_i(k_i-1)(t_i - t_{i+1})}{2\alpha}\right) \quad (2)$$

Strict clock model

We break down the likelihood $p(\mathcal{P}|\mathcal{T}, \Theta)$ into the product of the individual likelihoods of the observed number of substitutions, x_i, on each branch $i \in \{1, ..., b\}$ of the input phylogeny $\mathcal{P}$ given the duration, l_i of that branch in the dated tree $\mathcal{T}$. Substitution models typically consider a discrete number of substitutions on each branch. For example in the strict clock model (36) the same rate μ of evolution is applied to all branches, so that the number of substitutions x_i is simply distributed as $x_i \sim \text{Poisson}(\mu l_i)$, where x_i is discrete. However, phylogenetic software typically estimate the branch lengths x_i as a continuous variable, due in particular to the use of non-homogenous mutation models (37) and uncertainties in phylogenetic reconstruction (38). Consequently, we consider here a Gamma distribution, with mean equal to its variance by analogy with the Poisson distribution, so that the likelihood function becomes:

$$p(\mathcal{P}|\mathcal{T}, \Theta) = p(x_{1..b}|l_{1..b}, \mu) = \prod_{i=1}^{b} f_{\text{Gamma}}(x_i|\mu l_i, 1) \quad (3)$$

where the rate μ is included in the vector of parameters Θ. The Gamma distribution used above and throughout this article is parameterized in terms of the shape and scale parameters, respectively.

Relaxed clock model

In practice the assumption of a strict clock rate may be inappropriate, so next we consider an uncorrelated relaxed clock model where each branch has a specific rate m_i sampled from a given distribution (39). For example this distribution could be $m_i \sim \text{Gamma}(k, \theta)$, so that the product of the rate m_i and the branch length l_i is distributed as $m_i l_i \sim \text{Gamma}(k, l_i\theta)$. If we now consider substitution as a Poisson process with rate $m_i l_i$ we find that the number of mutations x_i is discrete and distributed as $x_i \sim \text{NegBin}\left(k, \frac{l_i\theta}{1+l_i\theta}\right)$ which is the relaxed clock model used by treedater (16). More generally, let us consider that the per-branch rates m_i are independent and identically distributed samples from an unspecified distribution with expectation and variance respectively equal to $\mathbf{E}(m_i) = \mu$ and $\mathbf{V}(m_i) = \sigma^2$. We also allow continuous values for x_i and consider, as we did for the strict clock model in (Equation 3), that $x_i \sim \text{Gamma}(m_i l_i, 1)$. By application of the laws of total expectation and variance, we can then deduce the expectation and variance of x_i:

$$\mathbf{E}(x_i) = \mathbf{E}(\mathbf{E}(x_i|m_i l_i)) = \mathbf{E}(m_i l_i) = \mu l_i \quad (4)$$

$$\begin{aligned}\mathbf{V}(x_i) &= \mathbf{E}(\mathbf{V}(x_i|m_i l_i)) + \mathbf{V}(\mathbf{E}(x_i|m_i l_i)) \\ &= \mathbf{E}(m_i l_i) + \mathbf{V}(m_i l_i) = \mu l_i + l_i^2\sigma^2 \quad (5)\end{aligned}$$

By analogy with the case of the strict clock model in (Equation 3), we impose a Gamma distribution with this mean and variance, resulting in the following likelihood function:

$$\begin{aligned}p(\mathcal{P}|\mathcal{T}, \Theta) &= p(x_{1..b}|l_{1..b}, \mu, \sigma) \\ &= \prod_{i=1}^{b} f_{\text{Gamma}}\left(x_i \left| \frac{l_i\mu^2}{\mu + l_i\sigma^2}, 1 + \frac{l_i\sigma^2}{\mu}\right.\right) \quad (6)\end{aligned}$$

where both μ and σ are included in the vector of parameters Θ. We note that the special case where the variance of the branch-specific rates is zero corresponds to the strict clock model, so that setting $\sigma = 0$ in (Equation 6) gives (Equation 3). This relaxed clock model is similar to the uncorrelated lognormal relaxed clock model (39) implemented in BEAST (3), in the sense that both the mean and the variance of the per-branch rates are independent parameters, whereas a model similar to the uncorrelated exponential relaxed clock model (39) could be obtained by setting $\mu = \sigma^2$. Note however that unlike these previous relaxed models we did not specify a distribution for the per-branch rates, but instead we specified a Gamma distribution for the resulting branch lengths in (Equation 6).

Accounting for bacterial recombination

The input phylogeny to be dated may be the output from phylogenetic software that accounts for the effect of bacterial recombination, for example ClonalFrameML (29) or Gubbins (30). In this case, the output contains for each branch i the proportion c_i of the genome that has been found to be non-recombinant on that branch, as well as the recombination-corrected length x_i of each branch. The branch length estimate in $\mathcal{P}$ is related to s_i, the number of substitutions observed in the *non-recombinant* portions of the genome, and c_i by the formula $x_i = s_i/c_i$. Such a recombination-corrected phylogeny could be dated as if it were the output of standard phylogenetic software but that may underrepresent uncertainty in the dating because only partial sequence was used to estimate x_i, especially when the fractions $1 - c_i$ of recombinant material are large. Instead, we implemented dating of such trees based on a modified likelihood function that accounts for the fact that only the non-recombinant regions are informative about the branch lengths. This is achieved by considering the distribution of the number $s_i = x_i c_i$ of substitutions in the non-recombinant regions and scaling down the substitution rates by a factor c_i. For example, in the case of a relaxed clock model, both μ and σ are scaled down by a factor c_i so that the likelihood in (Equation 6) is modified to give:

$$\begin{aligned}p(\mathcal{P}|\mathcal{T}, \Theta) &= p(x_{1..b}|l_{1..b}, \mu, \sigma) \\ &= \prod_{i=1}^{b} f_{\text{Gamma}}\left(c_i x_i \left| \frac{c_i l_i\mu^2}{\mu + c_i l_i\sigma^2}, 1 + \frac{c_i l_i\sigma^2}{\mu}\right.\right) \quad (7)\end{aligned}$$

As before, the case of a strict clock is obtained by setting $\sigma = 0$ in (Equation 7), so that the shape and scale parameters of the Gamma distribution become simply $c_i l_i \mu$ and 1, respectively. We have implemented functions that can read directly the output files of ClonalFrameML (29) and Gubbins (30) in order to date recombination-corrected phylogenies

using this approach. The approach described above offers more statistical power than removing all recombinant sites prior to reconstructing a dated phylogeny (e.g. (33,40)), for the same reason that reconstructing a standard phylogeny based on such an alignment would not result in a phylogeny as accurate as estimated by ClonalFrameML or Gubbins. As an extreme example of this, if all sites are recombinant in at least one phylogenetic branch then there would be no site left in the recombination-filtered alignment, whereas even in this situation the phylogenetic relationships between the genomes can be derived from the number m_i of mutations observed in the non-recombinant part c_i of each branch i.

Markov Chain Monte Carlo methodology

We sample from the posterior distribution in (Equation 1) using a Markov Chain Monte Carlo (MCMC). Most parameters, such as the age of each node in the dated genealogy $\mathcal{T}$ are updated using Metropolis-Hastings moves with normal proposals centred on the current value. One exception is the coalescent time unit α for which a Gibbs move is available, by noticing that in (Equation 2) the rate $1/\alpha$ admits a Gamma conjugate prior. Specifically, we consider a Gamma(k, θ) prior on $1/\alpha$, so that the posterior distribution of α is distributed as:

$$\alpha \sim \text{InvGamma}\left(n+k-1, \frac{2\theta}{2+\theta\sum_{i=2}^{2n-1} k_i(k_i-1)(t_i-t_{i+1})}\right) \quad (8)$$

The priors on the parameters μ, σ and $1/\alpha$ are Gamma(0.001,1000) by default and in all applications below.

We have so far been assuming that the root of the phylogeny $\mathcal{P}$ was predetermined for example using one or several outgroup sequences, and also that all sampling dates of the genomes in $\mathcal{P}$ were known exactly. However, both of these assumptions can easily be relaxed via data augmentation in which the location of the root in $\mathcal{P}$ and the unknown sampling dates are treated as additional parameters co-estimated using additional MCMC moves (41). For the location of the root, we consider as prior that all points on the phylogeny are equally likely to be the root and use two Metropolis-Hastings moves, one proposing to move the root from its current location to one of the branches directly underneath, and another proposing to move the root while staying on the same branch. For the sampling dates, the user can specify the bounds of the uniform prior considered as possible dates, or by default the range of all known sampling dates is used, and a Metropolis-Hastings move proposes to update the unknown sampling dates within their allowed range.

Options are available to perform inference under the strict clock model (Equation 3) or under the relaxed clock model (Equation 6), but by default we consider a mixture of the two models, in which half of the prior weight is given to each model. Mixing between the two models is implemented using reversible jumps to propose moves between the strict ($\sigma = 0$) and relaxed ($\sigma > 0$) models (42). This allows us to perform model comparison between the two models, and in particular to estimate the Bayes Factor as the ratio of MCMC iterations spent in each model (43). In summary, each MCMC iteration consists of the following MCMC moves, all of which are used by default but can be deactivated by the user:

- A Metropolis-Hastings move proposing to update the value of the mean substitution rate μ
- A Gibbs move updating the coalescent unit α
- When using the relaxed clock model, a Metropolis-Hastings move proposing to update the standard deviation σ of the per-branch substitution rates
- A reversible-jump move proposing to move from the strict clock model to the relaxed clock model or vice-versa
- For each internal node of the tree, a Metropolis-Hastings move proposing to update its date
- For each leaf of the tree with unknown sampling date, a Metropolis-Hastings move proposing to update its date
- Two Metropolis-Hastings moves proposing to update the root location

By default, the MCMC is run for a total of 10^5 iterations, with the first half discarded as MCMC burnin and the remainder sampled every 100 iterations. For all results presented below, the convergence and mixing of the chains was assessed using the R package coda (44). The effective sample size of the inferred parameters α, μ and σ were computed to make sure that they were >200. Furthermore, multiple chains were run separately and compared to ensure that the multivariate version of the Gelman-Rubin diagnostic (45,46) was lower than 1.1.

Implementation

The methodology described above was implemented in a new R package called BactDating and freely available at `https://github.com/xavierdidelot/BactDating`. For maximum computational efficiency, the likelihood and prior functions described in (Equations 2-7) were written in C++ and integrated into the R package using Rcpp (47). BactDating also includes functions to simulate dated coalescent trees from (Equation 2), and phylogenetic trees from (Equations 3 and 6), which we used to simulate datasets and assess the performance of our inference methodology.

BactDating also includes a function to perform root-to-tip linear regression analysis, including optimisation of the root to maximize the coefficient of determination R^2, and implementation of a previously described test to assess the significance of the temporal signal based on random permutations of sampling dates (48). This linear regression procedure is used to provide a good default starting point for the MCMC algorithm. Finally, several studies have proposed that the significance of the temporal signal can be tested by comparison with a run where all sampling dates are set equal (1,49–51), and we implemented this approach by computing the deviance information criterion DIC (52) for the two runs with and without sampling dates set equal.

RESULTS

Application to a single simulated dataset

To demonstrate the use of our Bayesian methodology, we first simulated a single dataset, consisting of 100 individu-

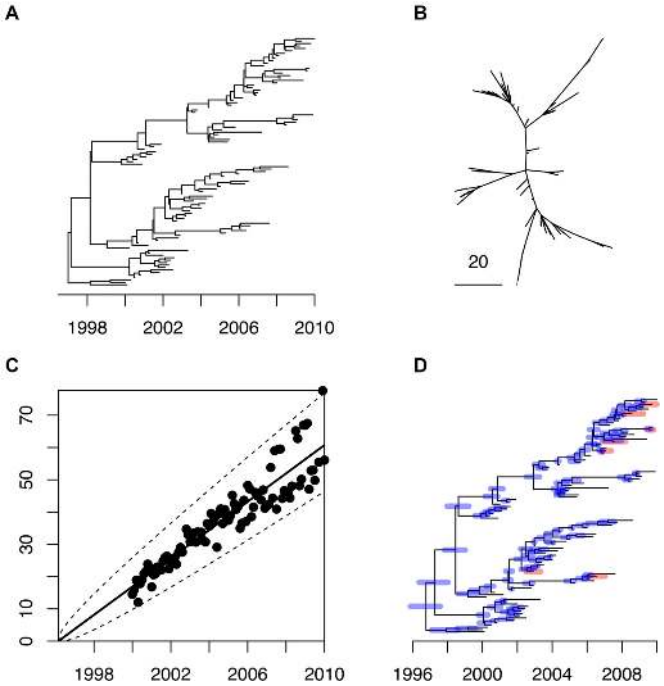

Figure 1. Application to a single simulated dataset. (**A**) the correct dated genealogy. (**B**) the unrooted phylogeny used as input. (**C**) Linear regression of root-to-tip (y-axis) versus sampling dates (x-axis). (**D**) Estimated dated genealogy, with blue bars indicating 95%CI for ancestral dates and red bars representing the 95%CI for the unknown sampling dates.

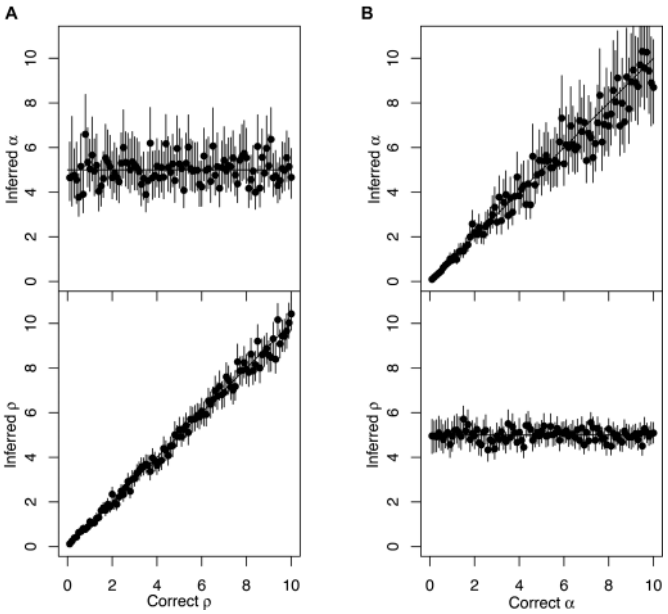

Figure 2. Application to multiple datasets simulated with a strict clock. (**A**) One hundred simulated datasets were analysed, each of which used parameters $\alpha = 5$ and $0.1 < \mu < 10$ (x-axis), and for both parameters the inferred mean (y-axis, dot) and 95%CI intervals (y-axis, line) are shown. (**B**) Same as panel A, but using a different set of 100 simulations for which the true parameters were $0.1 < \alpha < 10$ (x-axis) and a fixed $\mu = 5$.

als, sampled at regular intervals between the year 2000 and 2010. The genealogy was drawn from the heterochronous coalescent model (Equation 2) with coalescent time unit equal to $\alpha = N_e g = 5$ years (Figure 1A). The strict molecular clock model (Equation 3) was applied to this genealogy with mean rate of $\mu = 5$ substitutions per year to obtain an unrooted phylogenetic input tree (Figure 1B). We also consider the sampling dates as part of the input, except that each individual had a 10% probability of having an unknown sampling date. We first performed a linear regression analysis of root-to-tip distance versus sampling dates (when known), with the root position selected to optimize temporal signal. This resulted in a slightly underestimated clock rate of $\mu = 4.38$ substitutions per year, and a root located on the correct branch as in Figure 1A, but with an estimated date of 21 February 1996, underestimated compared to the correct root date 28 December 1996. This linear regression had a high fraction of variance explained by the model, $R^2 = 0.86$, with all points falling within or very close to the 95% confidence intervals (Figure 1C), and a highly significant p-value of $P < 10^{-4}$ based on a permutation test (48).

The clock rate and tree root estimated by the linear regression were both used as starting point for our MCMC procedure. The run time for the default 10^5 iterations was ~10 min on a standard desktop computer. Values in square brackets below represent the 95% credible intervals (95%CI) of estimated parameters. The posterior distribution of the coalescent time unit α had mean 4.69 years [3.66-5.98], which includes the correct value of 5 years used in the simulation. The substitution rate μ had mean 4.96 per year [4.46–5.47], which also includes the correct value of 5 per year. The posterior probability of the root location was highest for the correct branch, but only equal to 0.56 with the remaining probability being shared between the two branches directly below the short branch stemming from the real root (Figure 1A). Because of the shortness of this branch it is not surprising that there is uncertainty about the exact location of the root. Posterior mean and 95%CI were also estimated for the dates of all ancestral nodes and leaves for which the sampling dates were unknown (Figure 1D). In particular, the root of the tree had a mean date 24 September 1996 [30 October 1995–4 August 1997] which covers the correct date 28 December 1996.

Application to multiple simulated datasets

We repeated the procedure described above for 100 simulated datasets, each of which was generated with the same coalescent time unit $\alpha = 5$ years but with the substitution rate μ varying between 0.1 and 10 per year. For each dataset, we estimated the mean and 95%CI of the two parameters α and μ (Figure 2A). We found that estimated values for α remained around the correct value of 5, with most 95%CI covering 5, whereas the estimates of μ increased with the correct value of μ, with once again most 95%CI covering the correct values. We then repeated the procedure again for another 100 simulated datasets, but this time keeping $\mu = 5$ fixed and varying α between 0.1 and 10 year. As expected, we found that in these conditions the estimated values of μ remained constant and that the estimated values of α followed the correct values used in the simulations (Figure 2B).

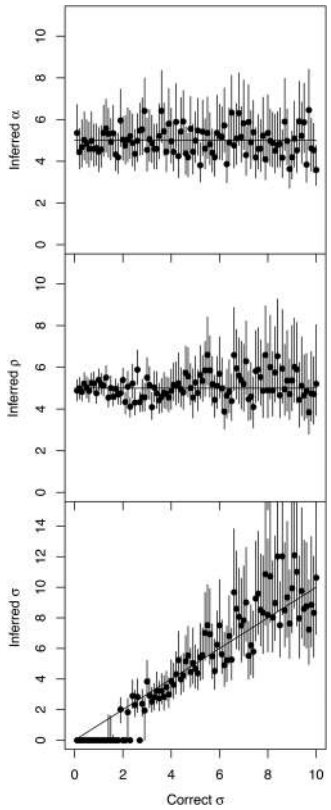

Figure 3. Application to multiple datasets simulated with a relaxed clock. One hundred simulated datasets were analysed, each of which used parameters $\alpha = 5$, $\mu = 5$ and $0.1 < \sigma < 10$ (x-axis), and for these parameters the inferred mean (y-axis, dot) and 95%CI intervals (y-axis, line) are shown.

The simulations considered so far were generated using a strict molecular clock (Equation 3) and inferred using a 50–50 mixture of the strict and relaxed clock models (cf Methods). The inferred Bayes Factors were always overwhelmingly in favour of the correct strict model, with the exception of only the first two simulations in Figure 2A, for which $\mu = 0.1$ and $\mu = 0.2$ substitutions per year, respectively. The strict clock rate used in these simulation was too low to rule out a relaxed clock model, and doing so would require a sampling interval of more than 10 years. We now consider a new set of 100 simulations performed under the relaxed clock model (Equation 6), in which the coalescent time unit is $\alpha = 5$, the average rate is $\mu = 5$ per year, and the standard deviation of the clock rate σ varies between 0.1 and 10. Inference was performed once again under the mixed model, in exactly the same conditions are previously. The estimates of the coalescent unit α and the average clock rate μ remained around the correct value of 5, with most 95%CI covering this value, but we note that as the standard deviation σ increased, so did the uncertainty on μ (Figure 3). The inferred values of σ followed the correct values, except when σ was <2, in which case σ was often inferred to be zero (Figure 3). This corresponds to datasets in which the model was incorrectly inferred to be the strict clock model ($\sigma = 0$) instead of the relaxed clock model ($\sigma > 0$). This behaviour is expected, since when the standard deviation σ of the per-branch clock rates is small (relative to its mean μ) the relaxation of the clock has little effect and therefore the data is hard to differentiate from data generated under the strict clock model. This incorrect model selection is therefore not an issue, and other parameter estimates such as the coalescent time unit α and evolutionary rate μ are unaffected (Figure 3). However, this behaviour demonstrates that our algorithm is relatively conservative in calling the clock relaxed, as a result of our choice of a highly uninformative prior on σ in the relaxed clock model which has a direct impact on model selection (53).

Taken altogether, these results on simulated data indicate that our MCMC procedure is correct, and that there is significant statistical power to estimate the key parameters of the models, and therefore to accurately perform Bayesian inference on the ancestral dates of a phylogeny, at least in the conditions used for simulating these datasets. The range of parameters used in the simulations above were selected to be representative of typical situations that arise in the genomic epidemiology of bacterial populations. In particular, the genome-wide substitution rate varies between species in the same order of magnitude considered above between 0.1 and 10 substitutions per year (2,24,54). Sequencing a sample of 100 genomes is also frequently achievable nowadays thanks to the recent reduction in cost and time required to sequence whole bacterial genomes (55). The assumption of a uniform unbiased sampling frame over 10 years represents a good case scenario, which is not always achievable. When it is not, the statistical power to accurately date a phylogenetic tree is likely to be reduced, and therefore the uncertainty in reconstructions is increased, which our Bayesian method is well suited to capture.

Comparison with other methods on benchmark data

The simulated datasets described above were designed to emulate real bacterial genomic datasets, and the same model was used for both simulation and inference. In order to test the robustness of our method to deviations in the underlying model, and to benchmark it against other methods, we also applied BactDating to previously described simulated datasets (12). These simulations were intended to emulate the evolution of HIV between hosts, based on a birth-death process with periodic sampling times, a mean clock rate of 0.006 substitutions per site per year and sequences of length 1000 bp (12). Two sampling schemes were considered: 25 individuals sampled at 3 times separated by 10 years and 10 individuals sampled at 11 times separated by 2 years. Two molecular clock models were used to generate the datasets: a strict clock model and an uncorrelated lognormal relaxed clock. For each of the four resulting combinations of sampling schemes and clock models, 100 datasets were simulated, and the results are shown in Supplementary Figure S1 (first sampling scheme, strict clock), Supplementary Figure S2 (first sampling scheme, relaxed clock), Supplementary Figure S3 (second sampling scheme, strict clock) and Supplementary Figure S4 (second sampling scheme, relaxed clock). We used BactDating to estimate the clock rate and date of the root based on the previously published unrooted phylogenies estimated using PhyML (19). We compared our results to the previously published results from a root-to-tip analysis similar to Tem-

pEst (13), LSD (12) and BEAST using a strict or relaxed clock model (6). The same previously published simulated data has also been used to benchmark treedater (16) and TreeTime (17). On the strict clock datasets, we found that BactDating performed at least as well as all other methods (Supplementary Figures S1 and S3). On the relaxed clock datasets, BactDating performed similarly to other methods, although there was a slight tendency to underestimate the mean clock rate (Supplementary Figures S2 and S4). This is probably due to saturation on branches with a high mutation rate, which would explain why LSD has a similar bias. However, such saturation would not happen in bacterial genomics due to much lower per site evolutionary rates (2). Overall, these results show that BactDating performs well even in conditions that it was not designed for, with minimal negative impact of the approximation in the phylogenetic reconstruction step, and high robustness to misspecification of the tree and evolutionary models.

Application to an ancient bacterial pathogen using aDNA

Mycobacterium leprae is the causative agent of leprosy, a debilitating disease that was endemic throughout Europe in the Middle Ages, and still remains a critical health threat in some parts of the developing world (56). Here we reanalyse previously published data from (31) including ten recent genomes (sampled between 1982 and 2012) and five ancient genomes (sampled between 990 and 1369). An unrooted phylogenetic tree was reconstructed using PhyML (19) (Supplementary Figure S5). After selecting the root that maximizes the coefficient of determination $R^2 = 0.9$, we find a strong correlation between sampling dates and root-to-tip distances (Figure 4A), with an estimated rate of 0.0353 substitution per genome per year and estimated root date of 928 BCE. All root-to-tip distances fall within the interval expected under a strict molecular clock (Figure 4A) and despite the low number of tree leaves, a date randomization test (48) found that the temporal signal is significant ($P < 10^{-4}$).

We performed the default 10^5 MCMC iterations, which took less than a minute to run. The dated phylogeny produced (Figure 4B) has the same root as for the root-to-tip analysis above, with mean dating of 1396 BCE and a broad 95%CI of [2735-490] BCE (Figure 4B). A strict clock model was inferred, with a Bayes Factor of 141.85. The clock rate had a posterior mean of 0.0314 substitutions per genome per year [0.0219–0.0419] (Supplementary Figure S6). These estimates are in excellent agreement with the original analysis of this data using BEAST (31). The substitution rate is low compared to values reported in similar bacterial phylogenomic studies as was previously reported (24,31), which is probably a result of both a low mutation rate in *M. leprae* and the negative dependency between substitution rate estimates and sampling time (2,24,57). To test further the significance of the temporal signal in this dataset, the MCMC was rerun under the assumption that all genomes were sampled on the same date. The deviance information criterion DIC (52) in this run was 243.28 compared to 170.57 when the correct dates were used, which indicates conclusively that the temporal signal is significant (49).

Application to a locally emerging clonal bacterial lineage

The four *Shigella* species are Enterobacteriaceae that have adapted to a human-restricted pathogenic lifestyle and become some of the most prevalent causes of human dysentery (58). The recent spread of antibiotic resistant lineages of *S. sonnei* to several developing countries where *S. sonnei* is traditionally rate is a major global health concern (59,60). We reanalysed previously published genomic data on the spread of the VN clade in Vietnam (32). *S. sonnei* is a clonal species, with only a single recombination event reported in a species-wide genomic study (59). No recombination event was reported in the VN dataset (32) and a ClonalFrameML (29) analysis found no recombination event either. A phylogenetic tree was constructed using PhyML (19) using 161 whole genomes sampled from Ho Chi Minh City (Vietnam) between 1995 and 2010 (Supplementary Figure S7). This phylogeny contained six outgroup genomes which were used to establish the location of the root for the remaining genomes (Supplementary Figure S7). As previously reported (32), the correlation between root-to-tip distances and isolation dates is very strong with a coefficient of determination $R^2 = 0.91$, and this result was found to be statistically significant according to a randomisation test ($P < 10^{-4}$, Supplementary Figure S8). This linear regression suggests a clock rate of 3.74 and a root date of 1982.68.

Running our algorithm for the default 10^5 MCMC iterations on this dataset took about ten minutes on a standard computer. Since only the year of the isolation dates were known, we allowed them to vary using a uniform prior within that year. The resulting dated phylogeny (Figure 5A) has mean dating 14 June 1983 [28 December 1977–18 November 1986], which is in excellent agreement with the previous report based on BEAST of 1982 [1978–1986] (32). A relaxed clock model was selected with a Bayes Factor greater than 1000 against the strict clock model. The inferred substitution rates had mean $\mu = 4.22$ substitutions per year [3.66-4.85]. This is equivalent to 8.34×10^{-7} [7.24×10^{-7}–9.59×10^{-7}] substitutions per site per year, which is in excellent agreement with the previous estimate from BEAST of 8.5×10^{-7} [7.6×10^{-7}–9.5×10^{-7}] (32).

The per-branch standard deviation of the relaxed clock model rate was estimated to be $\sigma = 2.24$ [1.57–3.09]. This is relatively high especially given that in the root-to-tip analysis almost all the genomes were within the 95% intervals expected under a strict clock model (Supplementary Figure S8). However, such a root-to-tip analysis is not a statistically powerful way of ensuring the validity of a strict clock model, because the root-to-tip distances are not independent of each other. To illustrate the inadequacy of a strict clock model, the number of substitutions on each branch was considered as a function its duration, along with the 95% ranges expected under both the strict clock and relaxed clock model (Figure 5B). Several branches have numbers of substitutions that fall outside of the strict clock range but within the relaxed clock range, illustrating the better fit of the relaxed clock model compared to the strict clock model.

Application to a recombining bacterial lineage

Streptococcus pneumoniae is a nasopharyngeal commensal and respiratory pathogen of humans, causing a high burden

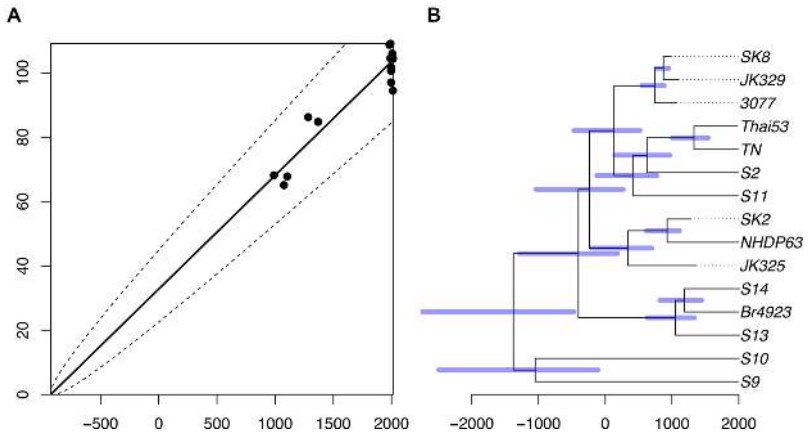

Figure 4. Analysis of *Mycobacterium leprae* dataset. (**A**) Linear regression of root-to-tip (y-axis) versus sampling dates (x-axis). (**B**) Estimated dated genealogy, with blue bars indicating 95%CI for ancestral dates.

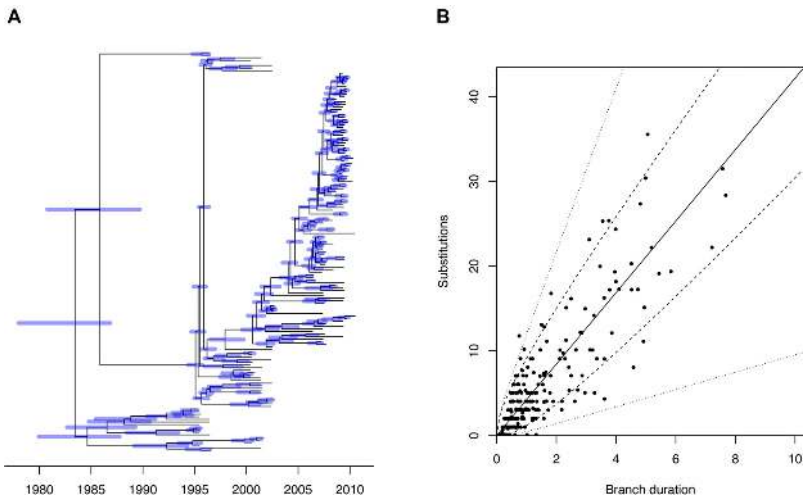

Figure 5. Analysis of *Shigella sonnei* VN dataset. (**A**) Estimated dated genealogy, with blue bars indicating 95%CI for ancestral dates. (**B**) Branch-by-branch comparison of duration in years (x-axis) and number of observed substitutions (y-axis). The expectation of the clock model is represented by the solid line, the 95% interval for the strict clock model is represented by the dashed lines and the 95% interval of the relaxed clock model is represented by the dotted lines.

of bacterial pneumonia, sepsis and meningitis worldwide. Originally detected in Spain, the PMEN1 lineage was one of the first multidrug-resistant *S. pneumoniae* found to have spread to multiple continents, and by the late 1990s was responsible for ∼40% of infant penicillin-resistant pneumococcal disease in the USA (61). Here we reanalyse previously published genomic data from 238 isolates (33), sampled between 1984 and 2008, although the sampling date was missing for 20 genomes. A phylogenetic tree uncorrected for recombination was constructed using RAxML (18) (Supplementary Figure S9) and a tree corrected for recombination was built using Gubbins (30) (Supplementary Figure S10). It was previously reported that correcting for recombination improved the temporal signal, and applying BEAST to the non-recombinant regions resulted in a PMEN1 root date estimate of 1969 [1958–1977] (33). Indeed, we find a coefficient of determination $R^2 = 0.22$ for a linear regression of root-to-tip distances against isolation dates based on the uncorrected tree (Supplementary Figure S11), compared with $R^2 = 0.59$ for the corrected tree (Supplementary Figure S12). Performing such a linear regression analysis on the uncorrected tree suggests a clock rate of 9.98 substitutions per year and a root date of 1981, whereas on the corrected tree the clock rate is estimated to be 3.21 substitutions per year and the root date 1971.

To illustrate the importance of accounting for recombination when dating lineages, we applied our MCMC algorithm to both the corrected and the uncorrected trees in exactly the same conditions. Each run took approximately 10 minutes using the default settings. Based on the uncorrected tree, a relaxed clock model was inferred with a mean rate μ of 3.72 [2.60–4.91] substitutions per year, and per branch standard deviation σ of 5.68 [3.91–7.66]. The higher value of σ compared to μ indicates that the clock is very relaxed, so that estimated dates are highly uncertain (Figure 6A). The root date for example is estimated to be 1523 with a 95% credible interval covering more than six centuries, from 1219 to 1885. The deviance information criterion DIC (52) was 3226.98 which is comparable to 3286.34 when all sampling dates were assumed identical, which suggests that the

Figure 6. Dating of *Streptococcus pneumoniae* PMEN1 before and after correcting for recombination. (**A**) Application of dating based on the RAxML tree uncorrected for recombination. (**B**) Application of dating based on the Gubbins tree corrected for recombination.

temporal signal is not strongly statistically significant in this uncorrected tree (49), even though a permutation test on the root-to-tip analysis (48) suggests it is ($P < 10^{-4}$).

When dating was applied to the recombination corrected tree, a relaxed clock model was also selected but this time the mean rate μ was 3.09 substitutions per genome per year [2.68–3.53] and the standard deviation σ was 1.04 [0.77–1.40]. Thus the clock is less relaxed than for the uncorrected tree, and the dates are more accurate (Figure 6B), for example the date of the root was estimated to be 1972 [1966–1977] which is in excellent agreement with the previous estimate of 1970 based on both root-to-tip analysis and BEAST (33). The deviance information criterion DIC was 3631.94 compared to 6725.77 when all sampling dates were set equal, which suggests that the temporal signal is definitely significant in the recombination corrected tree.

DISCUSSION

We have presented a new Bayesian approach called BactDating to produce dated phylogenies from a set of bacterial genomes. A key aim was to make sure that our method was fast and scalable to the large numbers of bacterial genomes that can be sequenced thanks to recent improvements in sequencing technologies (55). Several other fast scalable methods have been recently developed (12,16,17) but unlike these tools BactDating is based on the Bayesian statistical framework. Bayesian dating provides many advantages (22), such as the ability to naturally quantify uncertainties in parameter estimates, to consider different evolutionary models and to compare them. BactDating is slower than some of these non-Bayesian approaches, but remains fast enough to be applied to datasets of hundreds of genomes in a matter of minutes.

BactDating shares many similitudes with BEAST (3–6), including the use of a Markov Chain Monte Carlo to perform Bayesian inference, and the applications we presented on three real datasets showed that BactDating and BEAST produce highly consistent results. BactDating is several orders of magnitude faster and more scalable than BEAST, and this is achieved by assuming that the phylogenetic relationships between the genomes have been previously reconstructed using standard phylogenetic software. A first drawback of this approach lies in the computational cost of having to perform this previous analytical step, however this is not a significant issue in practice thanks to the recent development of fast maximum-likelihood phylogenetic software (18–21) which in most studies are already applied in parallel to dating. A more fundamental drawback concerns the fact that uncertainties associated with phylogenetic reconstruction are not accounted for in the dating. This could be addressed by running BactDating on multiple phylogenetic trees as was proposed in other applications where accounting for phylogenetic uncertainty was a concern (62–64). The high overall computation cost of this strategy could be avoided through the use of parallel computing, with each node computing for example a bootstrap replicate of the phylogenetic tree and performing dating using BactDating. BEAST explores the full space of unconstrained dated phylogenies, but it should be noted that this creates other issues such as difficulty in MCMC convergence and mixing (65,66), particularly in the presence of recombination (67), the need to build a consensus tree (68) and the occasional occurrence of non-sensical branches of negative lengths in such trees (69). On the other hand, the use by BactDating of previously assessed phylogenetic relationships can be a significant advantage if the phylogenetic software accounted for the disruptive effect of bacterial recombination, as do ClonalFrameML (29) and Gubbins (30).

Dating phylogenetic events without a prior idea of clock rate is only possible if the temporal signal in the dataset is significant and strong enough (13). This signal is typically assessed using a linear regression of root-to-tip distances versus isolation dates, but this is well known to be problematic since the root-to-tip distances are not independent of one another. Instead, we implemented a previously proposed approach which consists of comparing the results of dating with correct sampling dates and with all sampling dates set equal to one another (1,49–51). However, BactDating is also well suited to exploring other options, for example the idea of comparing the results of dating using the correct sampling dates to multiple runs where the sampling dates are randomized (14,24,70,71). So far, this approach has been used rarely in practice because it requires the anal-

ysis to be run many times, but our computationally efficient Bayesian framework makes this approach much more applicable than before.

Different substitution models can be used within BactDating as we illustrated by comparing strict and relaxed molecular clock models (Equations 3 and 6) on both simulated and real data. Another extension of the substitution model would be to account for the time dependency of substitution rates. The fact that observed substitution rates are lower on longer time scales compared to recent time scales has been well documented in viral phylogenetics (57,72,73) and more recently also in bacteria (2,24). A model for this dependency, for example an exponential decay equation (24,57), could be integrated into the distribution of number of substitutions for a given branch in order to test the validity of such a model and to account for this dependency in the dating. A different type of extension would be to consider alternative prior models for the dated phylogeny. Here we assumed a coalescent model with constant population size (Equation 2) which is relatively standard and uninformative, and therefore well suited to perform phylodynamic analysis in a subsequent step using tools that require a dated phylogeny as input (74,75). Alternatively, other models could easily be implemented within BactDating, either parametric such as an exponential growth model (35) or non-parametric such as a skyline model (7). Because it is both Bayesian and computationally efficient, BactDating is well suited to explore and compare such models extensions in future work.

DATA AVAILABILITY

BactDating is freely available for download from `https://github.com/xavierdidelot/BactDating`.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was funded by the Medical Research Council (grant MR/N010760/1) and the UK National Institute for Health Research (NIHR) Health Protection Research Unit (HPRU) in Modelling Methodology at Imperial College London in partnership with Public Health England (PHE) (grant HPRU-2012-10080). DJW is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (grant 101237/Z/13/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Conflict of interest statement. None declared.

REFERENCES

1. Drummond,A.J., Pybus,O.G., Rambaut,A., Forsberg,R. and Rodrigo,A.G. (2003) Measurably evolving populations. *Trends Ecol. Evol.*, **18**, 481–488.
2. Biek,R., Pybus,O.G., Lloyd-Smith,J.O. and Didelot,X. (2015) Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.*, **30**, 306–313.
3. Drummond,A.J. and Rambaut,A. (2007) BEAST : Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, **7**, 214.
4. Drummond,A.J., Suchard,M.A., Xie,D. and Rambaut,A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.
5. Bouckaert,R., Heled,J., Kühnert,D., Vaughan,T., Wu,C.-H., Xie,D., Suchard,M.A., Rambaut,A. and Drummond,A.J. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **10**, e1003537.
6. Suchard,M.A., Lemey,P., Baele,G., Ayres,D.L., Drummond,A.J. and Rambaut,A. (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.*, **4**, vey016.
7. Ho,S.Y.W. and Shapiro,B. (2011) Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.*, **11**, 423–434.
8. Didelot,X., Fraser,C., Gardy,J. and Colijn,C. (2017) Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.*, **34**, 997–1007.
9. Volz,E.M., Koelle,K. and Bedford,T. (2013) Viral Phylodynamics. *PLoS Comput. Biol.*, **9**, e1002947.
10. Lemey,P., Rambaut,A., Drummond,A.J. and Suchard,M. (2009) Bayesian phylogeography finds its roots. *PLoS Comput. Biol.*, **5**, e1000520.
11. Ward,M.J., Gibbons,C.L., McAdam,P.R., van Bunnik,B.A.D., Girvan,E.K., Edwards,G.F., Fitzgerald,J.R. and Woolhouse,M.E.J. (2014) Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Appl. Environ. Microbiol.*, **80**, 7275–7282.
12. To,T.-H., Jung,M., Lycett,S. and Gascuel,O. (2016) Fast dating using least-squares criteria and algorithms. *Syst. Biol.*, **65**, 82–97.
13. Rambaut,A., Lam,T.T., Max Carvalho,L. and Pybus,O.G. (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.*, **2**, vew007.
14. Duchêne,S., Geoghegan,J.L., Holmes,E.C. and Ho,S.Y. (2016) Estimating evolutionary rates using time-structured data: A general comparison of phylogenetic methods. *Bioinformatics*, **32**, 3375–3379.
15. Jones,B.R. and Poon,A.F. (2017) Node.dating: Dating ancestors in phylogenetic trees in R. *Bioinformatics*, **33**, 932–934.
16. Volz,E.M. and Frost,S.D.W. (2017) Scalable relaxed clock phylogenetic dating. *Virus Evol.*, **3**, vex025.
17. Sagulenko,P., Puller,V. and Neher,R.A. (2018) TreeTime: Maximum likelihood phylodynamic analysis. *Virus Evol.*, **4**, vex042.
18. Stamatakis,A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
19. Guindon,S., Dufayard,J.-F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
20. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
21. Nguyen,L.T., Schmidt,H.A., Von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
22. Bromham,L., Duchene,S., Hua,X., Ritchie,A., Duchene,D. and Ho,S. (2018) Bayesian molecular dating: opening up the black box. *Biol. Rev.*, **93**, 1165–1191.
23. Hedge,J. and Wilson,J. (2014) Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *MBio*, **5**, e02158-14.
24. Duchêne,S., Holt,K.E., Weill,F.-X., Le Hello,S., Hawkey,J., Edwards,D.J., Fourment,M. and Holmes,E.C. (2016) Genome-scale rates of evolutionary change in bacteria. *Microb. Genomics*, **2**, e000094.
25. Schierup,M.H. and Hein,J. (2000) Consequences of recombination on traditional phylogenetic analysis. *Genetics*, **156**, 879–891.
26. Didelot,X. and Falush,D. (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics*, **175**, 1251–1266.
27. Vaughan,T.G., Welch,D., Drummond,A.J., Biggs,P.J., George,T. and French,N.P. (2017) Inferring ancestral recombination graphs from bacterial genomic data. *Genetics*, **205**, 857–870.
28. Didelot,X., Lawson,D.J., Darling,A.E. and Falush,D. (2010) Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*, **186**, 1435–1449.

29. Didelot,X. and Wilson,D.J. (2015) ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.*, **11**, e1004041.
30. Croucher,N.J., Page,A.J., Connor,T.R., Delaney,A.J., Keane,J.A., Bentley,S.D., Parkhill,J. and Harris,S.R. (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.*, **43**, e15.
31. Schuenemann,V.J., Singh,P., Mendum,T.A., Krause-Kyora,B., Jäger,G., Bos,K.I., Herbig,A., Economou,C., Benjak,A., Busso,P. *et al.* (2013) Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science*, **341**, 179–183.
32. Holt,K.E., Vu,T., Nga,T., Pham,D., Vinh,H., Wook,D., Phan,M. and Tra,V. (2013) Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 17522–17527.
33. Croucher,N.J., Harris,S.R., Fraser,C., Quail,M.A., Burton,J., van der Linden,M., McGee,L., von Gottberg,A., Song,J.H., Ko,K.S. *et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science*, **331**, 430–434.
34. Kingman,J. (1982) The coalescent. *Stoch. Process. Appl.*, **13**, 235–248.
35. Drummond,A.J., Nicholls,G.K., Rodrigo,A.G. and Solomon,W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.
36. Zuckerkandl,E. and Pauling,L. (1962) *Molecular Disease Evolution, and Genic Heterogeneity*. Academic Press.
37. Liò,P. and Goldman,N. (1998) Models of molecular evolution and phylogeny. *Genome Res.*, **8**, 1233–1244.
38. Yang,Z. and Rannala,B. (2012) Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, **13**, 303–314.
39. Drummond,A.J., Ho,S.Y.W., Phillips,M.J. and Rambaut,A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.
40. Didelot,X., Dordel,J., Whittles,L.K., Collins,C., Bilek,N., Bishop,C.J., White,P.J., Aanensen,D.M., Parkhill,J., Bentley,S.D. *et al.* (2016) Genomic analysis and comparison of two gonorrhea outbreaks. *MBio*, **7**, e00525-16.
41. van Dyk,D.A. and Meng,X.-L. (2001) The art of data augmentation. *J. Comput. Graph. Stat.*, **10**, 1–50.
42. Green,P.J. (1995) Reversible Jump Markov Chain Monte Carlo Computation and bayesian model determination. *Biometrika*, **82**, 711–732.
43. Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **18**, 773–795.
44. Plummer,M., Best,N., Cowles,K. and Vines,K. (2006) CODA: convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.
45. Gelman,A. and Rubin,D.B. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–511.
46. Brooks,S.P.B. and Gelman,A.G. (1998) General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.*, **7**, 434–455.
47. Eddelbuettel,D. (2011) Seamless R and C++ Intgration with Rcpp. *J. Stat. Softw.*, **40**, 1–18.
48. Ramsden,C., Holmes,E.C. and Charleston,M.A. (2009) Hantavirus evolution in relation to its rodent and insectivore hosts: No evidence for codivergence. *Mol. Biol. Evol.*, **26**, 143–153.
49. Rambaut,A. (2000) Incorporating Non-Contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, **16**, 395–399.
50. Baele,G., Lemey,P., Bedford,T.B.C., Rambaut,A., Suchard,M.a. and Alekseyenko,A.V. (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.*, **29**, 2157–2167.
51. Murray,G.G.R., Wang,F., Harrison,E.M., Paterson,G.K., Mather,A.E., Harris,S.R., Holmes,M.A., Rambaut,A. and Welch,J.J. (2016) The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.*, **7**, 80–89.
52. Spiegelhalter,D., Best,N., Carlin,B. and Van der Linde,A. (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)*, **64**, 583–639.
53. Sinharay,S. and Stern,H.S. (2002) On the sensitivity of Bayes factors to the prior distributions. *Am. Stat.*, **56**, 196–201.
54. Didelot,X., Bowden,R., Wilson,D.J., Peto,T. E.A. and Crook,D.W. (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.*, **13**, 601–612.
55. Loman,N.J. and Pallen,M.J. (2015) Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.*, **13**, 787–794.
56. Scollard,D.M., Adams,L.B., Gillis,T.P., Krahenbuhl,J.L., Truman,W. and Williams,D.L. (2006) The continuing challenges of leprosy the continuing challenges of leprosy. *Clin. Microbiol. Rev.*, **19**, 338–381.
57. Ho,S.Y.W., Lanfear,R., Bromham,L., Phillips,M.J., Soubrier,J., Rodrigo,A.G. and Cooper,A. (2011) Time-dependent rates of molecular evolution. *Mol. Ecol.*, **20**, 3087–3101.
58. The,H.C., Thanh,D.P., Holt,K.E., Thomson,N.R. and Baker,S. (2016) The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nat. Rev. Microbiol.*, **14**, 235–250.
59. Holt,K.E., Baker,S., Weill,F.-X., Holmes,E.C., Kitchen,A., Yu,J., Sangal,V., Brown,D.J., Coia,J.E., Kim,D.W. *et al.* (2012) Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.*, **44**, 1056–1059.
60. Thompson,C.N., Duy,P.T. and Baker,S. (2015) The rising dominance of Shigella sonnei: An intercontinental shift in the etiology of bacillary dysentery. *PLoS Negl. Trop. Dis.*, **9**, 1–13.
61. Corso,A., Severina,E.P., Petruk,V.F., Maurlz,Y.R. and Tomasz,A. (1998) Molecular characterization of penicillin-resistant Streptococcus pneumoniae isolates causing respiratory disease in the United States. *Microb. Drug Resist.*, **4**, 325–337.
62. Parker,J., Rambaut,A. and Pybus,O.G. (2008) Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. *Infect. Genet. Evol.*, **8**, 239–246.
63. Nylander,J. A.A., Olsson,U., Alström,P. and Sanmartín,I. (2008) Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: Turdus). *Syst. Biol.*, **57**, 257–268.
64. Didelot,X., Gardy,J. and Colijn,C. (2014) Bayesian inference of infectious disease transmission from whole genome sequence data. *Mol. Biol. Evol.*, **31**, 1869–1879.
65. Kulkarni,M.A., Walimbe,A.M., Cherian,S. and Arankalle,V.A. (2009) Full length genomes of genotype IIIA Hepatitis A Virus strains (1995-2008) from India and estimates of the evolutionary rates and ages. *Infect. Genet. Evol.*, **9**, 1287–1294.
66. Eldholm,V., Pettersson,J.H., Brynildsrud,O.B., Kitchen,A. and Michael,E. (2016) Armed conflict and population displacement as drivers of the evolution and dispersal of Mycobacterium tuberculosis. *PNAS*, **113**, 13881–13886.
67. Dearlove,B.L., Cody,A.J., Pascoe,B., Méric,G., Wilson,D.J., Sheppard,S.K., Daniel,J. and Sheppard,S.K. (2016) Rapid host switching in generalist Campylobacter strains erodes the signal for tracing human infections. *ISME J.*, **10**, 721–729.
68. Holder,M.T., Sukumaran,J. and Lewis,P.O. (2008) A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Syst. Biol.*, **57**, 814–821.
69. Heled,J. and Bouckaert,R.R. (2013) Looking for trees in the forest: Summary tree from posterior samples. *BMC Evol. Biol.*, **13**, 221.
70. Firth,C., Kitchen,A., Shapiro,B., Suchard,M.A., Holmes,E.C. and Rambaut,A. (2010) Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Mol. Biol. Evol.*, **27**, 2038–2051.
71. Duchêne,S., Duchêne,D., Holmes,E.C. and Ho,S.Y. (2015) The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.*, **32**, 1895–1906.
72. Ho,S.Y.W. and Larson,G. (2006) Molecular clocks : when times are a-changin'. *Trends Genet.*, **22**, 79–83.
73. Ho,S.Y.W., Shapiro,B., Phillips,M.J., Cooper,A. and Drummond,A.J. (2007) Evidence for time dependency of molecular rate estimates. *Syst. Biol.*, **56**, 515–522.
74. Karcher,M.D., Palacios,J.A., Lan,S. and Minin,V.N. (2017) phylodyn: an R package for phylodynamic simulation and inference. *Mol. Ecol. Resour.*, **17**, 96–100.
75. Volz,E.M. and Didelot,X. (2018) Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst. Biol.*, **67**, 719–728.