VIRUS EVOLUTION

Virus Evolution, 2023, **9(1)**, 1–7
DOI: https://doi.org/10.1093/ve/veac120
Advance access publication date: 21 December 2022
Resources

bayroot: Bayesian sampling of HIV-1 integration dates by root-to-tip regression

Roux-Cil Ferreira,[1,†] Emmanuel Wong,[1] and Art F. Y. Poon[1,2,3,4,‡,*]

[1]Department of Pathology and Laboratory Medicine, Western University, London, ON N6A 5C1, Canada, [2]Department of Microbiology and Immunology, Western University, London, ON N6A 3K7, Canada, [3]Department of Computer Science, Western University, London, ON N6A 5B7, Canada and [4]Health Sciences Addition, Western University, H422, London, Ontario N6A 5C1, Canada
[†]https://orcid.org/0000-0002-8242-7862
[‡]https://orcid.org/0000-0003-3779-154X
*Corresponding author: E-mail: apoon42@uwo.ca

Abstract

The composition of the latent human immunodeficiency virus 1 (HIV-1) reservoir is shaped by when proviruses integrated into host genomes. These integration dates can be estimated by phylogenetic methods like root-to-tip (RTT) regression. However, RTT does not accommodate variation in the number of mutations over time, uncertainty in estimating the molecular clock, or the position of the root in the tree. To address these limitations, we implemented a Bayesian extension of RTT as an R package (*bayroot*), which enables the user to incorporate prior information about the time of infection and start of antiretroviral therapy. Taking an unrooted maximum likelihood tree as input, we use a Metropolis–Hastings algorithm to sample from the joint posterior distribution of three parameters (the rate of sequence evolution, i.e., molecular clock; the location of the root; and the time associated with the root). Next, we apply rejection sampling to this posterior sample of model parameters to simulate integration dates for HIV proviral sequences. To validate this method, we use the R package *treeswithintrees* (*twt*) to simulate time-scaled trees relating samples of actively and latently infected T cells from a single host. We find that *bayroot* yields significantly more accurate estimates of integration dates than conventional RTT under a range of model settings.

Key words: HIV-1; latent viral reservoir; molecular clock; root-to-tip regression; Bayesian inference.

1. Introduction

Root-to-tip (RTT) regression is a simple method to locate the earliest point in time in a phylogenetic tree (i.e., rooting the tree; Huelsenbeck, Bollback, and Levine 2002), to measure the rate of evolution (Drummond et al. 2003), or to reconstruct the divergence times of common ancestors. This method assumes the existence of a strict molecular clock, i.e., that the rate of molecular evolution is roughly constant (Bromham and Penny 2003). Accordingly, the number of nucleotide substitutions accumulating in a sequence should increase linearly over time. Hence, this method is a linear regression of the evolutionary divergence of sequences from their common ancestor against the times when those sequences were observed. The primary input of RTT regression is an unrooted phylogenetic tree with branch lengths measured in units of evolutionary time (i.e., the expected number of substitutions per site; Tajima and Nei 1984), which is the standard output of maximum likelihood methods for reconstructing phylogenies. The tips of the tree representing observed sequences are labelled with sampling times. Thus, RTT becomes an optimization over three parameters: the location of the root in the tree, the time associated with the root (x-intercept), and the molecular clock (slope of regression).

RTT has a broad range of applications. Since many viruses have a very rapid rate of evolution, RTT can be applied to sequences collected over a number of months or years. For instance, RTT has recently been used to estimate the origin date and clock rate of severe acute respiratory syndrome coronavirus 2 within the first few months of the pandemic (Duchene et al. 2020). We are particularly interested in the use of RTT to estimate the integration dates of HIV-1 proviruses within hosts (Jones et al. 2018). HIV-1 converts its RNA genome into double-stranded DNA that becomes integrated into the host genome as part of the virus replication cycle. In some cases, this integrated provirus becomes reversibly dormant in a transcriptionally inactive host cell (Siliciano and Siliciano 2004). This long-lived reservoir of latently infected cells is the primary obstacle to an effective cure for HIV-1. Consequently, characterizing the composition and dynamics of the latent reservoir has significant implications for HIV-1 cure research (e.g., Gondim et al. 2021).

For instance, we can estimate the molecular clock (the slope of the regression) from longitudinal samples of plasma HIV-1 RNA sequences before the start of antiretroviral therapy (ART). If we reconstruct a tree relating both these RNA sequences and proviral sequences from the latent reservoir, we can then use our clock estimate to extrapolate integration dates for the latter (Jones et al. 2018). This relies on the assumption that the integrated HIV-1 genome ceases to accumulate mutations upon integrating into the

host genome. Since we are reconstructing a tree relating individual HIV-1 sequences from a single host, the resulting tree is technically a 'genealogy' instead of a phylogeny, and we are counting mutations in an individual lineage instead of the accumulation of substitutions in a population. Consequently, we will avoid using the terms 'phylogeny' and 'substitution' from this point onward.

Due to its simplicity, RTT has a number of significant limitations. It implicitly assumes that the input tree is known without error. In practice, each proviral sequence is mapped to the regression line for a given tree to yield one and only one estimate of its integration date. Although one could generate interval estimates for integration dates, this is not trivial because we need to consider the joint confidence for the regression slope and intercept and to invert the model to predict dates. Furthermore, variation in the number of mutations after a given amount of time is expected, even under a strict molecular clock (Langley and Fitch 1974). In other words, a proviral sequence may by chance carry more mutations than expected given its actual date of integration. This random outcome can cause RTT to project a sequence's integration date estimate into the future, past its time of sampling or even past the start of ART, when the infection of new cells should be nearly completely suppressed.

Here we describe a Bayesian extension of the RTT method to estimate HIV-1 integration dates. Adopting a Bayesian approach provides a means of quantifying our uncertainty in estimating integration dates, as well as incorporating prior information about the time of infection and the start of ART. We detail our implementation of this method as an R package called *bayroot* and use a simulation model of within-host population dynamics to validate *bayroot* in comparison to conventional RTT methods.

2. Methods

2.1 Regression model

We start with an unrooted tree T relating n observed sequences. A strict molecular clock assumes that mutations accumulate at a constant rate μ over time, such that the number of mutations per unit time follows a Poisson distribution. Let Y_i be the number of mutations in the i^{th} observed sequence, which is determined by the location of the root in T. Since Y_i is an integer-valued outcome, we must rescale the input tree T by multiplying its branch lengths by the sequence length, such that lengths are in units of the expected number of mutations per genome. Note that because these measures of evolutionary time are derived from a continuous-time Markov model of sequence evolution, multiple hits and reversions are accounted for. Let t_0 be the origin time associated with the root. Let Δt_i be the time that has elapsed between the i^{th} sample and the root. The log-likelihood for a set of RNA sequences $\{Y_i, \Delta t_i\}$ is:

$$\log L(Y_i, \Delta t_i) = \sum_i Y_i \log(\mu \Delta t_i) - \mu \Delta t_i - \log\Gamma(Y_i + 1), \quad (1)$$

where $\Gamma(x)$ is the gamma function. Equation (1) is sometimes referred to as the Langley–Fitch model (Langley and Fitch, 1974).

Following prior work (Huelsenbeck, Bollback, and Levine 2002; Didelot et al. 2018), we assume a uniform prior distribution for possible locations of the root over the entire length of the tree. We also assume a uniform prior distribution for t_0, as standard practice in applications of Bayesian inference to HIV-1 infections (e.g., Sweeting et al. 2010; Stirrup and Dunn 2018). If a seroconversion window, i.e., the time interval between the last HIV seronegative visit and the first seropositive visit, is available for the host individual, these visit dates can be used to set lower and upper bounds for the uniform prior on t_0. If this information is not available, then these bounds may be based on other data such as viral load and CD4 cell count measurements (Pantazis et al. 2019). Otherwise, one may set the upper bound to the first sample collection date, and the lower bound may be derived from regional estimates of the time to HIV-1 diagnosis (e.g., Van Sighem et al. 2015). Finally, we assume a lognormal prior distribution on the clock rate μ, which can be informed by previous measurements of HIV-1 mutation rates within hosts e.g., Alizon and Fraser (2013).

With these prior distributions and the model likelihood, we implemented a Metropolis–Hastings sampling algorithm in R. A proposal function shifts the root along a branch by some distance $d \sim \text{Unif}(0, \delta_r)$, selecting a branch at random if it encounters an internal node (i.e., split) as it traverses the length of the tree. If, however, a terminal node is encountered before the root has been shifted by distance d, then the remaining distance is travelled by reflecting back from this terminus. This results in a symmetric proposal distribution. We also used a uniform proposal $\mu' \sim \text{Unif}(\mu - \delta_c, \mu + \delta_c)$ for the clock rate and a truncated normal proposal $t_0' \sim N(t_0, \sigma)$ for the origin time. The sampling algorithm returns an S3 object storing a data frame of sampled parameter values and a character vector of sampled trees serialized into Newick strings.

2.2 Sampling integration dates

Given a posterior sample of parameters Y, μ, and t_0, we need to propagate this information to the distribution of integration times associated with proviral DNA sequences sampled post-ART initiation. Using Bayes rule, the probability of integration time t_j for the j^{th} DNA sequence given divergence Y_j is:

$$P(t_j | Y_j) = \frac{P(Y_j | t_j) P(t_j)}{P(Y_j)}, \quad (2)$$

where we index by j instead of i to emphasize a shift from HIV-1 RNA to DNA sequences. We assume a uniform prior for integration times, $P(t_j) = (T - t_0)^{-1}$ for $t_0 \le t_j \le T$ and $P(t_j) = 0$ otherwise, where t_0 is the origin date and T is the time of ART initiation. Substituting Equation (1) as $P(Y_j | t_j)$ into the denominator $P(Y_j) = \int_{t_0}^{T} P(Y_j | t_j) P(t_j) dt$ and setting $s = t - t_0$, we solve the definite integral:

$$P(Y_j) = \frac{\int_0^{T-t_0} (\mu s)^{Y_j} \exp(-\mu s) \mathrm{d}s}{(T - t_0)\Gamma(Y_j + 1)} = \frac{\gamma(Y_j + 1, \mu(T - t_0))}{\mu(T - t_0)\Gamma(Y_j + 1)}, \quad (3)$$

where $\gamma(a, x)$ is the lower incomplete gamma function, $\int_0^x t^{a-1} \exp(-t) dt$. Finally, substituting Equations (1) and (3) into (2), we can write:

$$P(t_j | Y_j) = \frac{\mu M^{Y_j} \exp(-M)}{\gamma(Y_j + 1, M)}, \quad (4)$$

where we use a shorthand $M = \mu(T - t_0)$ to simplify the equation. To generate a sample of integration dates from this distribution, we use a simple rejection sampling method. For a given posterior sample of Y_j, μ, and t_0, we use Brent's algorithm to find the maximum of Equation (4), initialized at the midpoint $t = t_0 + (T - t_0)/2$. This maximum was used as an upper bound for rejection sampling for times drawn from the prior distribution, $t \sim \text{Unif}(t_0, T)$.

The Bayesian regression and integration date sampling methods described above were implemented in R as a package called *bayroot*. All source code is publicly available under the MIT license at https://github.com/PoonLab/bayroot.

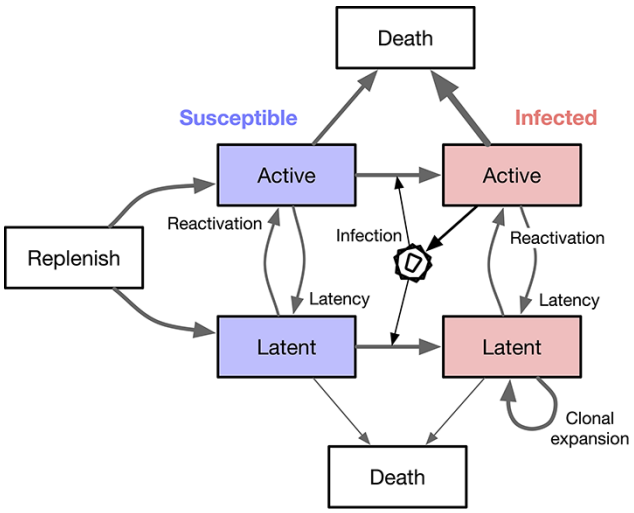

Figure 1. A schematic diagram of the compartmental model used to simulate cell population dynamics. Each box represents a well-mixed population of cells sharing the same rate parameters. We assume that only actively infected cells release virus particles that go on to infect other susceptible cells.

2.3 Simulating data

To validate the above method, we used the R package *twt* ('treeswithintrees', https://github.com/PoonLab/twt) to simulate cell population dynamics forward in time and then to simulate trees by sampling lineages backwards in time to their common ancestors. This package uses exact stochastic simulation of discrete events (Gillespie, 1977). In brief, it calculates the total rate of all events (Λ), draws an exponentially distributed waiting time to the first event $\tau \sim \Lambda \exp(-\Lambda)$, and then draws a uniform random number to determine which event occurs. We implemented a compartmental model of cell population dynamics (Fig. 1) that can be represented by the following set of differential equations:

$$\begin{aligned}
\frac{dT}{dt} &= -\rho T, \\
\frac{dA_S}{dt} &= \rho k T + m_{LA} L_S - \lambda_{AA}(t) A_I A_S - m_{AL} A_S - \mu_{A_S} A_S, \\
\frac{dA_I}{dt} &= \lambda_{AA}(t) A_I A_S + m_{LA} L_I - m_{AL} A_I - \mu_{A_I} A_I, \\
\frac{dL_S}{dt} &= r(1-k)T + m_{AL} A_S - \lambda_{AL}(t) A_I L_S - \lambda_{LL} L_I L_S - m_{LA} L_S - \mu_L L_S, \\
\frac{dL_I}{dt} &= \lambda_{AL}(t) A_I L_S + \lambda_{LL} L_I L_S + m_{AL} A_I - m_{LA} L_I - \mu_L L_I. \qquad (5)
\end{aligned}$$

This model is a simplified version of the system described by Rong and Perelson (2009). Most notably, our version does not model changes in the viral load. T represents a finite population of naive CD4+ T cells from which the populations of active (A) and resting (latent, L) cells are replenished at rates $k\rho$ and $(1-k)\rho$, respectively, for $0 \leq k \leq 1$. The S and I subscripts denote susceptible and infected subpopulations of active and latent cells. A branching event (λ_{xy}) requires a source cell to induce a target cell to undergo a change of state (switch compartments from x to y). For example, λ_{AA} represents the infection rate of a susceptible active T cell by a virus released from an actively infected cell. We assume that virus replication is completely blocked by the initiation of ART at time t^* (Kearney et al. 2014; Brodin et al. 2016), such that $\lambda_{A\bullet}(t \geq t^*) = 0$. A transition event occurs when a cell spontaneously migrates from compartments x to y at rate m_{xy}. For example, m_{LA} represents the reactivation rate of a latent cell. Finally, we assume constant cell death rates μ_x for each compartment x.

The simulation is initialized at time zero with user-specified population sizes of susceptible cells in each compartment, and a single actively infected cell, $A_I(0) = 1$. We simulated the integer-valued population size trajectories $\{T, A_S, A_I, L_S, L_I\}(t)$ forward in time until a stopping time of $t = 20$ simulation time units. We generated 50 replicate sets of trajectories under two different scenarios by exact stochastic simulation. The rate parameters were set to the following values: $r = 0.02$, $k = 0.5$, $\lambda_{AA}(t < t^*) = 0.002$, $\lambda_{AL}(t < t^*) = 10^{-4}$, $m_{AL} = m_{LA} = 0.001$, $\mu_{A_S} = 0.005$, $\mu_{A_I} = 0.1$, and $\mu_L = 0.001$. ART was initiated at $t^* = 10$ time units post-infection in Scenario 1 and at $t^* = 15$ in Scenario 2. For each iteration of the simulation, we calculated the rates for every type of event, adjusted by the respective compartment size at the current time t. For example, the rate of transmissions from A_I to A_S was set to $\lambda_{AA}(t) A_I(t) A_S(t)$. We drew an exponential waiting time given the total rate of all event types:

$$\Lambda(t) = \sum_{x,y} \lambda_{xy}(t) N_x(t) N_y(t) + \sum_{x,y} m_{xy}(t) N_x(t),$$

and then determined which event type occurred with probability $\lambda_{xy}(t) N_x(t) N_y(t) / \Lambda(t)$ or $m_{xy}(t) N_x(t) / \Lambda(t)$. Next, we incremented or decremented the respective population sizes for compartments affected by the event type. The time, type, and compartments of this event is recorded in a log that is later used to simulate trees. An example set of population size trajectories simulated using this algorithm under Scenario 1 is illustrated in Fig. 2.

To generate a tree relating the sampled lineages in *twt*, we applied another exact stochastic simulation algorithm in reverse time. For the 50 replicate sets of trajectories generated under Scenario 1, we sampled 10 HIV-1 RNA lineages at times $t = 3$, 6, and 9 post-infection. For trajectories generated under Scenario 2, we sampled 10 HIV-1 RNA lineages at $t = 11$, 13, and 15 post-infection. In both scenarios, we sampled 10 latently infected cells at $t = 20$ post-infection, for a total of 40 sampled lineages per replicate tree. These lineage sampling times defined the initial conditions for the reverse-time simulation of trees. Next, the algorithm samples events from the log generated in the forward-time simulation to build up a tree relating the sampled lineages. The stopping condition of the tree sampling algorithm is that the sampled lineages converge to a single common ancestor, which becomes the root.

We modified *twt* to output a Newick serialization of this 'transmission tree' among cells, labelling tips with sampling times. This tree included internal nodes with only one descendant branch, representing lineage state transitions, or transmissions to/from an unsampled lineage. Internal nodes were labelled with strings encoding the event type, node states (compartments), and unique identifiers for the individual cells involved. These annotations enabled us to 'colour' the branches of the tree by lineage state. The true integration dates for sampled latently infected cells were recorded to a separate file. An example of a tree generated by this process is shown in Fig. 2.

To simulate molecular evolution, we collapsed all branches corresponding to latently infected cells and used the resulting tree as input for INDELible (version 1.03; Fletcher and Yang 2009). We assigned an HIV-1 *env* sequence at the root (GenBank accession number AY772699). This sequence is one of the HIV-1 subtype C references curated by the Los Alamos National Laboratory HIV Sequence Database (http://www.hiv.lanl.gov). We configured INDELible to use the Tamura–Nei (TrN) model of nucleotide evolution with transition rates $\kappa_1 = 4$ and $\kappa_2 = 8$ and stationary base frequencies $f_A = 0.4$ and $f_C = f_G = f_T =$

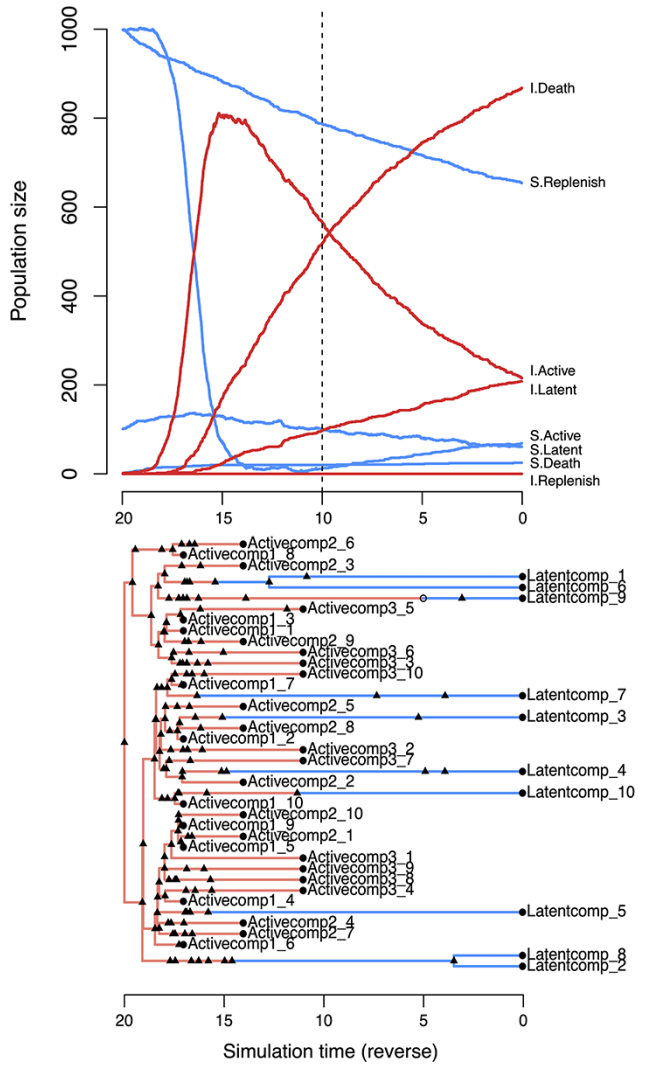

Figure 2. Examples of *twt* simulation outputs for a model of cell dynamics in the latent reservoir. (top) Population dynamics simulated forward in time. Each line represents the population size of a different compartment. S = susceptible, I = infected. The dashed vertical line indicates the time of ART initiation. This plot was produced by calling the generic plot method on the S3 object from the *twt* function *sim.outer.tree*. (bottom) A tree simulated in reverse-time, relating 10 cells sampled from the latently infected compartment at $\tau = 0$ and 30 from the actively infected compartment at $\tau = 11, 14, 17$ (Scenario 1), where $\tau = 20 - t$. Triangles represent transmission events, open circles represent transitions, and closed circles represent sampling times. Branches representing cell lineages in a latent state (Latentcomp_N) are collapsed prior to simulating virus evolution. This tree visualization was generated from the same S3 object using the R package *ggfree* (https://github.com/ArtPoon/ggfree).

0.2. In addition, we rescaled the tree such that the expected number of mutations per nucleotide site over its entire length was 1. Finally, we used FastTree (version 2.1.11, compiled for double precision; Price, Dehal and Arkin 2010) to reconstruct unrooted maximum likelihood trees from these simulated alignments.

2.4 Model validation

We ran our Bayesian sampling method on each of the 100 simulated trees for 2×10^4 steps, discarding a burn-in of 2,000 steps and thinning the remaining chain down to 1,000 steps. We set the lognormal prior distribution on clock rates to $\mu = -5$ and $\sigma = 2$, and the uniform prior distribution on root dates to a minimum of one simulation time unit before the true origin and a maximum of the first HIV RNA sampling time. In addition, we set the proposal parameters to $\delta_r = 0.01$ for the root location, $\sigma = 0.33$ for the time of infection, and $\delta_c = 0.01$ for the clock rate. In preliminary runs, we found that these settings were sufficient for replicate chain samples to converge to the same posterior distribution. To sample integration dates for each DNA sequence, we further thinned the chain down to a total of 200 samples from the posterior distribution to reduce auto-correlation.

To compare our results against conventional RTT regression, we censored the sampling times associated with tips that represented DNA sequences and then rooted the tree using the *rtt* function in the R package *ape* (implementation by R. M. McCloskey; Paradis and Schliep 2019). We extracted the RTT distances from the resulting tree and fit a simple linear regression of these distances against sampling times. Finally, we used the *inverse.predict* function from R package *chemCal* (Massart et al. 1997) to extract the predicted integration dates for the 200 samples from the posterior distribution.

To quantify the discordance between estimated ($\hat{t}$) and actual (t) integration dates, we calculated the root mean square error, $\text{RMSE} = \sqrt{\sum_{i=1}^{n} (\hat{t}_i - t_i)^2 / n}$, where n is the number of DNA sequences. We also calculated the mean absolute percentage error, $\text{MAPE} = 100\% \times \sum_{i=1}^{n} \left(|\hat{t}_i - t_i| / t_i \right) / n$, as an alternative measure of estimation error that is less sensitive to extreme values. We used a paired Wilcoxon rank-sum test to evaluate the significance of differences between the RMSE (or MAPE) values obtained from *bayroot* and conventional RTT.

3. Results

To compare conventional RTT regression to our Bayesian approach (*bayroot*), we simulated the proliferation of HIV-1 among active and latent CD4+ T cells with an exact stochastic method. Our simulation workflow yielded a total of 100 trees reconstructed from HIV-1 RNA and integrated proviral DNA sequences. We assumed that HIV-1 RNA was sampled before the start of ART and that HIV-1 proviral DNA was sampled from the latent reservoir post-ART initiation (Fig. 2). Fifty of the trees were simulated such that HIV-1 RNA was sampled at three time points starting at 3 simulation time units post-infection, at intervals of 3 time units (Scenario 1). For the remaining 50 trees, HIV-1 RNA sampling was delayed to 11 time units post-infection and taken at narrower intervals of 2 time units (Scenario 2).

Figure 3 compares the estimates of HIV-1 DNA integration dates produced by RTT and *bayroot*. Under Scenario 1, both methods tended to produce similar estimates because the sampling conditions were favourable for fitting the molecular clock (Fig. 3A). The median RMSE was 0.947 for RTT and 0.889 time units for *bayroot*. On a case-by-case basis, *bayroot* produced significantly more accurate estimates than RTT (paired Wilcoxon test, $P = 3.55 \times 10^{-4}$, Fig. 3B). The overall difference between estimates was numerically small. For instance, the median difference in RMSE between RTT and *bayroot* was 0.059 (interquartile range, IQR = 0.004 − 0.201) time units. In some cases, integration dates were mapped by RTT to the time period after ART initiation, leading to higher RMSE values (Fig. 3B). Since *bayroot* incorporates the prior information that HIV-1 integration should not occur during effective ART, its estimates are constrained to times preceding ART initiation. We found no significant difference between methods (paired Wilcoxon test, $P = 0.66$; Fig. S1A) when error was measured by MAPE, which is less

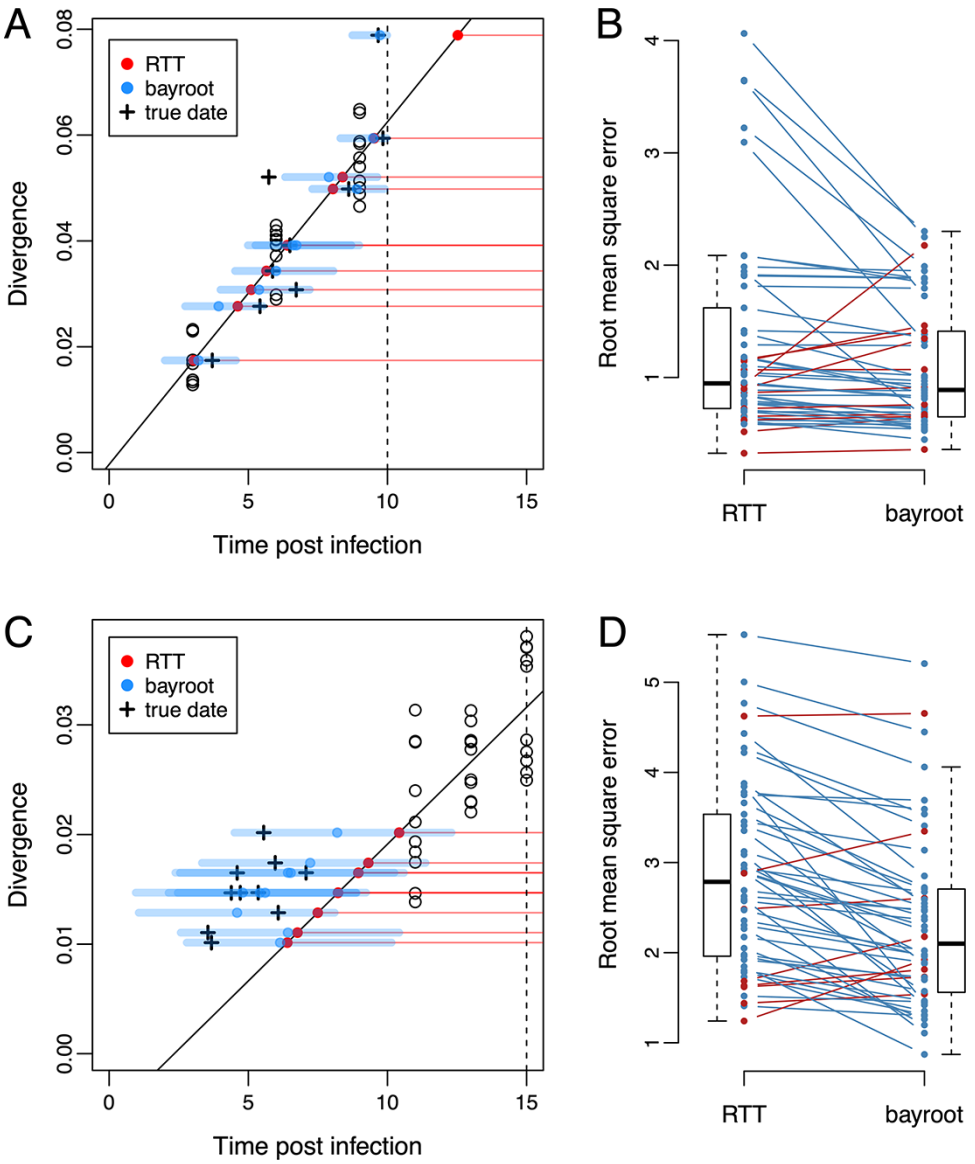

Figure 3. Comparison of results from *bayroot* and conventional RTT regression. (A) A scatterplot of RTT distance (divergence) against sampling times post-infection, for a representative example generated under Scenario 1. A solid line represents the RTT regression fitted to the RNA sequence data (open circles), which we expect to intercept the horizontal axis at $t = 0$. A vertical dashed line marks the start of ART. Red points represent estimates of integration dates from the RTT model for DNA sequences sampled at time $t = 20$, as indicated by horizontal red, thin lines. Blue points and thick line segments represent the median and 95 per cent credible interval for integration date estimates from *bayroot*. Cross marks indicate the actual integration dates. (B) A slopegraph comparing the RMSE of integration date estimates from RTT and *bayroot* for all 50 simulations generated under Scenario 1. Line segments are coloured red if the RMSE for a given simulation was greater for *bayroot*, and blue otherwise. (C) and (D) A scatterplot and slopegraph for simulations generated under Scenario 2. Slopegraphs were generated using R package *ggfree* (https://github.com/ArtPoon/ggfree).

influenced by the largest errors than the RMSE. Furthermore, 89.8 per cent of the actual integration dates fell within the 95 per cent credible intervals generated by *bayroot*.

For Scenario 2, both methods became less accurate with median RMSEs of 2.79 and 2.10 time units for RTT and *bayroot*, respectively (Fig. 3D). Because the sampling times of the RNA sequences used to calibrate the molecular clock were closer together and more distant from the actual time of infection in this scenario (Fig. 3C), we are less certain about all three parameters of the regression, i.e., the location of the root in the tree, the time associated with the root (*x*-intercept), and the clock rate (slope). Under these conditions, *bayroot* benefits from having prior information about the time of infection. For our simulations where $t = 0$ is the actual time, we constrained the time of infection variable to the interval from −1 to 3 simulation time units. (In practice, one could use a uniform prior bounded by the last seronegative and first seropositive dates for that individual.) In other words, prior information about the time of infection 'anchors' the RTT regression when there are insufficient data to accurately estimate the *x*-intercept (Fig. 3C). As a result, *bayroot* was significantly more accurate than RTT under this second scenario (paired Wilcoxon test, $P = 3.82 \times 10^{-7}$, Fig. 3D). The median difference in RMSE between RTT and *bayroot* was 0.405 (IQR $0.190 - 0.807$) time units—about seven times greater than scenario 1. In addition, this difference between methods remained significant when error was measured as MAPE (paired Wilcoxon test, 3.4×10^{-7}; Fig. S1B). Decomposition of the mean squared error into bias and variance components indicated that the difference in RMSE was driven

more by a reduction in bias in either scenario (Fig. S2). Finally, 89.4 per cent of actual integration dates fell within the 95 per cent credible intervals from *bayroot*. There was no significant association in this outcome between scenarios (Fisher's exact test, odds ratio = 0.5, $P = 0.34$).

Running a chain sample for 2×10^4 steps in *bayroot* required a median of 47.3 (IQR $45.0 - 48.8$) seconds in R version 4.2.0 for Linux on a single core of an AMD Ryzen ThreadRipper 1950X processor. If the user is not processing a large number of samples, as we have done here for replicate simulations, we suggest running chain samples for at least 10^6 steps with a thinning interval of 500 steps.

4. Discussion

The reconstruction of HIV-1 integration dates is a challenging problem. Cells carrying replication-competent provirus in the latent reservoir comprise a small fraction of resting CD4+ T cells (approximately 0.01–10 per million cells; Prodger et al. 2020; Crooks et al. 2015). Sequences of plasma HIV-1 RNA or integrated DNA often cover only a portion of the virus genome (Laskey et al. 2016), making it difficult to resolve their evolutionary relationships. In addition, the development of phylogenetic and statistical methods for analysing these sequence data (Ferreira et al. 2021) has lagged behind ongoing improvements in molecular techniques (Cho et al. 2022; Sun et al. 2022). Here we have described a Bayesian extension of a widely used regression method for estimating HIV-1 integration dates from sequence variation in the latent reservoir (Jones et al. 2018; Brodin et al. 2016; Brooks et al. 2020). Our method provides a means of incorporating additional data about the infection—e.g., the estimated date of infection, time of ART initiation, and previous measures of the rate of HIV-1 evolution within hosts—as prior information. Furthermore, adopting a Bayesian approach enables us to quantify our uncertainty about parameter estimates by sampling from the posterior distribution. We expect this will be important for studies where there is limited access to longitudinal plasma samples for retrospective sequencing, for instance.

Of course, our method also retains some significant limitations of conventional approaches to RTT regression. First, we are assuming that the unrooted tree relating HIV-1 RNA and DNA sequences is known without error. It is possible to relax this assumption by adopting a hierarchical approach and replicating our regression analysis on a posterior sample of unrooted trees that may be generated by a Bayesian phylogenetic program such as MrBayes (Ronquist and Huelsenbeck 2003) or BEAST (Drummond and Rambaut 2007). This is less efficient than sampling from the joint posterior distribution of unrooted trees, mutation model, and the RTT regression parameters. Additionally, we are assuming that the divergence of each sequence is an independent outcome. This convenient approximation is clearly untrue because of identity by descent: sequences that share a more recent common ancestor will have a similar RTT distance because they have inherited the same set of mutations. It is possible to overcome this limitation by adapting the covariance matrix of the regression model to the phylogenetic structure of the data (Neher 2018).

Not all studies use RTT regression to estimate HIV-1 integration dates. For example, one of the methods described by Abrahams et al. (2019) uses approximate maximum likelihood to reconstruct a host-specific tree relating HIV-1 RNA and DNA sequences and then locates the closest tip representing an RNA sequence for every tip representing a DNA sequence, which is assigned the sampling time of the RNA tip. Hence, the DNA sequences can only be associated with a finite number of integration dates. This approach benefits from extensive sampling of HIV-1 plasma RNA over the time period spanning the start of infection to ART initiation. If the ancestral HIV-1 RNA sequence most closely related to an HIV-1 provirus is not represented in the tree, then the latter would be mapped to another branch that may be associated with a sampling time that does not accurately estimate the integration date. In contrast, RTT methods directly use the number of mutations carried by an individual DNA sequence to estimate its integration date. The other sequences are used to calibrate the linear model mapping this divergence to the timeline.

Data availability

The R package *bayroot* is publicly available under the MIT license at https://github.com/PoonLab/bayroot. We have also provided the simulated data and R scripts used to perform the method validation and generate figures in this repository. The R package *twt* is publicly available under the GNU Affero General Public License v3.0 (AGPL-3.0) at https://github.com/PoonLab/twt.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

We thank Dr Jessica Prodger, with whom our discussion about the use of RTT regression to estimate the integration dates of HIV provirus from the latent reservoir motivated the development of this Bayesian method. This work was presented at the 29th International Dynamics and Evolution of Human Viruses meeting in La Jolla, California, USA.

References

Abrahams, M. -R. et al. (2019) 'The Replication-Competent HIV-1 Latent Reservoir is Primarily Established near the Time of Therapy Initiation', *Science Translational Medicine*, 11: eaaw5589.

Alizon, S. and C., Fraser (2013) 'Within-Host and Between-Host Evolutionary Rates Across the HIV-1 Genome', *Retrovirology*, 10: 1–10.

Brodin, J. et al. (2016) 'Establishment and Stability of the Latent HIV-1 DNA Reservoir', *eLife*, 5: e18889.

Bromham, L. and D., Penny (2003) 'The Modern Molecular Clock', *Nature Reviews Genetics*, 4: 216–224.

Brooks, K. et al. (2020) 'HIV-1 Variants are Archived Throughout Infection and Persist in the Reservoir', *PLoS Pathogens*, 16: e1008378.

Cho, A. et al. (2022) 'Longitudinal Clonal Dynamics of HIV-1 Latent Reservoirs Measured by Combination Quadruplex Polymerase Chain Reaction and Sequencing', *Proceedings of the National Academy of Sciences*, 119: e2117630119.

Crooks, A. M. et al. (2015) 'Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies', *Journal of Infectious Diseases*, 212: 1361–1365.

Didelot, X. et al. (2018) 'Bayesian Inference of Ancestral dates on Bacterial Phylogenetic Trees', *Nucleic Acids Research*, 46: e134–e134.

Drummond, A., O. G., Pybus and A., Rambaut (2003) 'Inference of viral Evolutionary Rates from Molecular Sequences', *Adv Parasitol*, 54: 331–358.

Drummond, A. J. and A., Rambaut (2007) 'BEAST: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology*, 7: 1–8.

Duchene, S. et al. (2020) 'Temporal Signal and the Phylodynamic Threshold of SARS-CoV-2', *Virus Evolution*, 6: veaa061.

Ferreira, R. -C. et al. (2021) 'Quantifying the Clonality and Dynamics of the Within-Host HIV-1 Latent Reservoir', *Virus Evolution*, 7: veaa104.

Fletcher, W. and Z., Yang (2009) 'INDELible: a Flexible Simulator of Biological Sequence Evolution', *Molecular Biology and Evolution*, 26: 1879–1888.

Gillespie, D. T. (1977) 'Exact Stochastic Simulation of Coupled Chemical Reactions', *Journal of Physical Chemistry*, 81: 2340–2361.

Gondim, M. V. P. et al. (2021) 'Heightened Resistance to Host Type 1 Interferons Characterizes HIV-1 at Transmission and after Antiretroviral Therapy Interruption', *Science Translational Medicine*, 13: eabd8179.

Huelsenbeck, J. P., J. P., Bollback and A. M., Levine (2002) 'Inferring the Root of a Phylogenetic Tree', *Systematic Biology*, 51: 32–43.

Jones, B. R. et al. (2018) 'Phylogenetic Approach to Recover Integration Dates of Latent HIV Sequences Within-Host', *Proceedings of the National Academy of Sciences*, 115: E8958–E8967.

Kearney, M. F. et al. (2014) 'Lack of Detectable HIV-1 Molecular Evolution During Suppressive Antiretroviral Therapy', *PLoS pathogens*, 10: e1004010.

Langley, C. H. and W. M., Fitch (1974) 'An Examination of the Constancy of the Rate of Molecular Evolution', *Journal of Molecular Evolution*, 3: 161–177.

Laskey, S. B. et al. (2016) 'Evaluating Clonal Expansion of HIV-Infected Cells: Optimization of PCR Strategies to Predict Clonality', *PLoS Pathogens*, 12: e1005689.

Desiré Luc Massart, Bernard G M Vandeginste, Lutgarde M C Buydens, Sijbrand de Jong, Paul J Lewi, and Johanna Smeyers-Verbeke. *Handbook of Chemometrics and Qualimetrics: Part A*. Data Handling in Science and Technology: 20A. Elsevier Science, 1997.

Neher, R. A. (2018) 'Efficient Estimation of Evolutionary Rates by Covariance Aware Regression', *bioRxiv*, 408005.

Pantazis, N. et al. (2019) 'Determining the Likely Place of HIV Acquisition for Migrants in Europe Combining Subject-Specific Information and Biomarkers Data', *Statistical Methods in Medical Research*, 28: 1979–1997.

Paradis, E. and K., Schliep (2019) 'ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R', *Bioinformatics*, 35: 526–528.

Price, M. N., P. S., Dehal and A. P., Arkin (2010) 'FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments', *PLoS One*, 5: e9490.

Prodger, J. L. et al. Reduced HIV-1 Latent Reservoir Outgrowth and Distinct Immune Correlates Among Women in Rakai, Uganda', *JCI Insight*, 5: 2020.

Rong, L. and A. S., Perelson (2009) 'Modeling Latently Infected Cell Activation: Viral and Latent Reservoir Persistence, and Viral Blips in HIV-Infected Patients on Potent Therapy', *PLoS Computational Biology*, 5: e1000533.

Ronquist, F. and J. P., Huelsenbeck (2003) 'MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models', *Bioinformatics*, 19: 1572–1574.

Siliciano, J. D. and R. F., Siliciano (2004) 'A Long-Term Latent Reservoir for HIV-1: Discovery and Clinical Implications', *Journal of Antimicrobial Chemotherapy*, 54: 6–9.

Stirrup, O. T. and D. T., Dunn (2018) 'Estimation of Delay to Diagnosis and Incidence in HIV using Indirect Evidence of Infection Dates', *BMC Medical Research methodology*, 18: 1–14.

Sun, C. et al. (2022) 'Droplet-Microfluidics-Assisted Sequencing of HIV Proviruses and Their Integration Sites in Cells from People on Antiretroviral Therapy', *Nature Biomedical Engineering* 6, 1004–1012.

Sweeting, M. J. et al. (2010) 'Estimating the Distribution of the Window Period for Recent HIV Infections: a Comparison of Statistical Methods', *Statistics in Medicine*, 29: 3194–3202.

Tajima, F. and M., Nei (1984) 'Estimation of Evolutionary Distance between Nucleotide Sequences', *Molecular Biology and Evolution*, 1: 269–285.

Van Sighem, A. et al. (2015) 'Estimating HIV Incidence, Time to Diagnosis, and the Undiagnosed HIV Epidemic using Routine Surveillance Data', *Epidemiology (Cambridge, Mass.)*, 26: 653.