

1 Validation of dated phylogenies in microbial population genetics

2 Xavier Didelot<sup>1,2,\*</sup>, Jake Carson<sup>1,3</sup>, Paolo Ribeca<sup>4,5</sup>, ...

3 <sup>1</sup> School of Life Sciences, University of Warwick, United Kingdom

4

5 <sup>2</sup> Department of Statistics, University of Warwick, United Kingdom

6

7 <sup>3</sup> Mathematics Institute, University of Warwick, United Kingdom

8

9 <sup>4</sup> UK Health Security Agency, London, United Kingdom

10

11 <sup>5</sup> Biomathematics and Statistics Scotland, The James Hutton Institute, Edinburgh, United Kingdom

12

13 \* Corresponding author. Tel: 0044 (0)2476 572827. Email: `xavier.didelot@gmail.com`

# INTRODUCTION

Dated phylogenies, also known as tip-calibrated, time-stamped or time-calibrated phylogenies, have become a ubiquitous tool in the study of microbial population genetics (Drummond et al. 2003; Biek et al. 2015; Rieux and Balloux 2016). In a dated phylogeny, the branch lengths are measured in a unit of time, for example years or days, rather than a unit of evolution as in a standard phylogeny. Consequently, the tips of a dated phylogeny are aligned with the (typically known) dates of sampled genomes and the internal nodes are aligned with the (typically inferred) dates of common ancestors between the genomes. Many tools exist to build dated phylogenies, either from a sequence alignment using for example BEAST (Suchard et al. 2018) or BEAST2 (Bouckaert et al. 2019), or by dating the nodes of a standard phylogeny, using for example LSD (To et al. 2016), node.dating (Jones and Poon 2017), treedater (Volz and Frost 2017), BactDating (Didelot et al. 2018) and TreeTime (Sagulenko et al. 2018). The dated phylogeny is interesting in itself since it depicts the ancestral relationships of sampled genomes over time, but it is also often used as the foundation for further analysis (Didelot and Parkhill 2022), such as inference of demographics (Baele et al. 2016), phylogeography (Lemey et al. 2009) or transmission between hosts (Didelot et al. 2017).

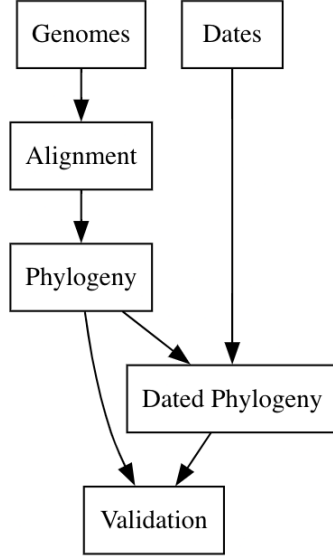
There are many factors that can invalidate the results of a dated phylogenetics analysis. This includes in particular the confounding effect that population structure can have on dating (Duchene et al. 2015; Murray et al. 2016). This is especially true when the substructures are imbalanced (Duchêne et al. 2015), are sampled at different dates (Tong et al. 2018), have different clock rates (Wertheim et al. 2012) and when the population structure is strong (Navascués and Emerson 2009). More generally, any incorrect assumptions made by the model under which the dating analysis is performed can invalidate the results.

One approach that has been used to ensure that there are no incorrect assumptions being made in the model is to perform inference under multiple models and perform model comparison, typically by computing a Bayes Factor (Baele et al. 2012; Li and Drummond 2012; Bouckaert and Drummond 2017). However, this requires multiple runs under different models, and only provides a relative measure of model appropriateness, with no indication of how good the best model actually is in absolute terms. Another related line of research involves testing the significance of the temporal signal (Duchene et al. 2015, 2020). This can be done for example by comparing results with and without dates (Rambaut 2000) or by randomizing the leaf dates (Duchene et al. 2015).

Here we present an alternative approach, in which we seek to evaluate the correctness of an inference and detect if there are any reasons to believe that the inference is not valid. This approach is sometimes referred to as model checking, model diagnostics or model validation, and it is complementary with the model comparison methodology mentioned above. We study the distribution of residuals after fitting a model, following methodology reminiscent of regression models (Cox and Snell 1968; Dunn and Smyth 1996), but also previously applied more generally for example to epidemic models (Lau et al. 2014) or Hidden Markov Models (Zucchini and MacDonald 2009; Buckby et al. 2020).

We use simulated datasets to demonstrate that this approach can detect a wide range of problems in the inference, including the aforementioned confounding effect of population structure (Murray et al. 2016). We also demonstrate the usefulness of this approach in practice on real datasets.

A



B

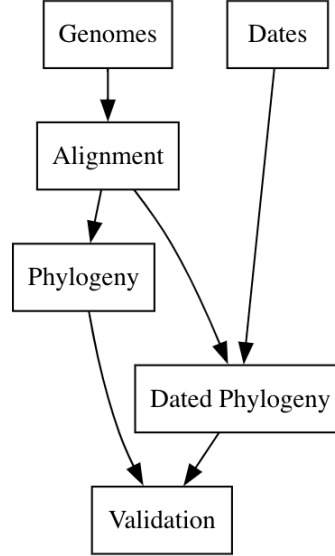


Figure 1: (A) General approach for validation of a dated phylogeny built by dating the nodes of a standard phylogeny. (B) General approach for validation of a dated phylogeny built directly from a sequence alignment.

## RESULTS

### General approach

We want to validate a dated phylogeny  $\mathcal{D}$ , previously constructed using any method. We propose to do so by comparing the dated phylogeny  $\mathcal{D}$  with an undated phylogeny  $\mathcal{L}$ . If the method used to construct  $\mathcal{D}$  involved dating the nodes of an undated phylogeny, for example TreeTime (Sagulenko et al. 2018) or treedater (Volz and Frost 2017), then this is readily available for validation and we therefore focus on this case in this article (Figure 1A). However, the validation methodology below can also be applied to methods that build a dated phylogeny directly from the alignment such as BEAST (Suchard et al. 2018), simply by constructing a separate undated phylogeny from the same alignment using for example PhyML (Guindon et al. 2010) or RAxML (Stamatakis 2015) (Figure 1B). For each branch in the dated phylogeny  $\mathcal{D}$  we can consider its inferred length, the number of substitutions happening on that branch in the standard phylogeny  $\mathcal{L}$ , and the model and parameters used when building the dated phylogeny  $\mathcal{D}$ , in order to compute a residual for that branch (see Methods). If the inference is valid, these residuals will follow their theoretical distribution (Cox and Snell 1968; Dunn and Smyth 1996). We use this property as a way to test the validity of the dated phylogeny  $\mathcal{D}$ .

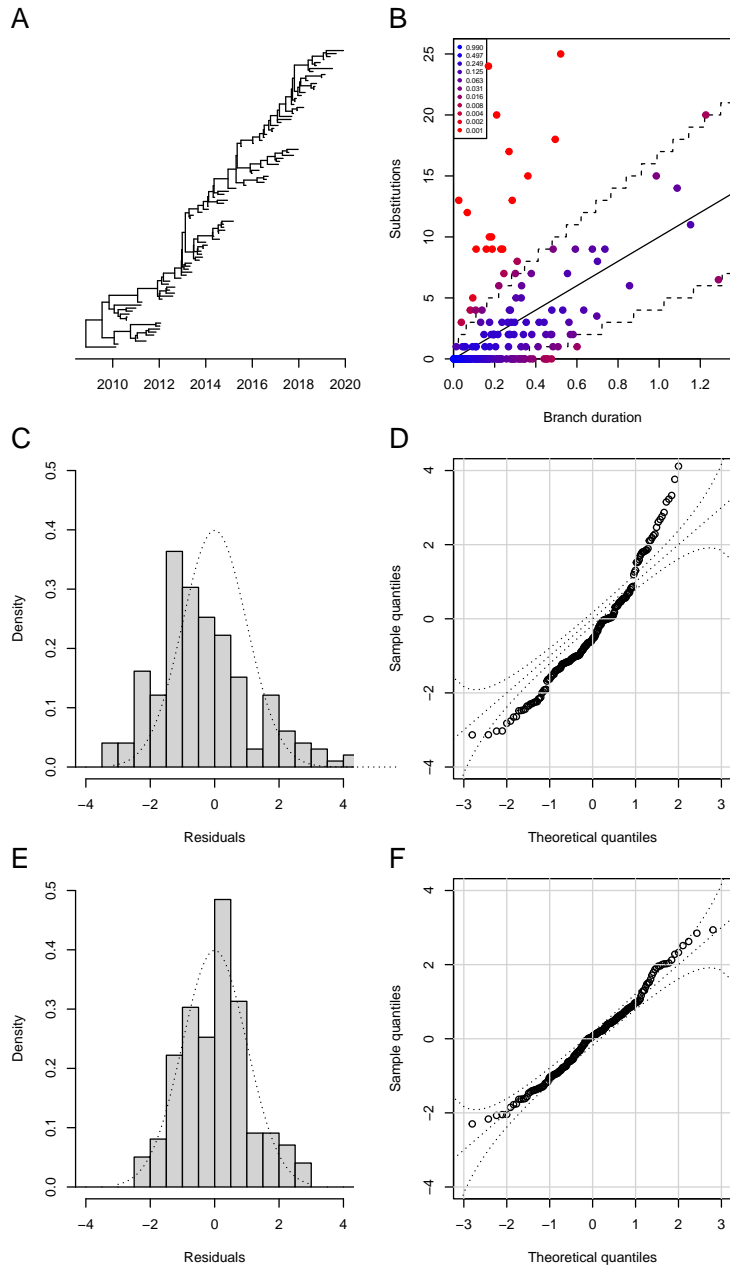


Figure 2: (A) Simulated dated phylogeny. (B) Distribution of substitutions generated by a relaxed clock model on the branches of the dated phylogeny, with their probability under a strict clock model. (C) Distribution of residuals after inference under a strict clock model. (D) QQ plot of residuals after inference under a strict clock model. (E) Distribution of residuals after inference under a relaxed clock model. (F) QQ plot of residuals after inference under a relaxed clock model.

## 69 Motivating example

70 A dated phylogeny was simulated including 100 leaves uniformly distributed between 2010 and 2020,  
71 under the heterochronous coalescent model (Drummond et al. 2002) with constant population size  
72  $N_e g = 1$  year (Figure 2A). We applied the additive relaxed clock model (Didelot et al. 2021) to this  
73 dated phylogeny, with mean clock rate  $\mu = 10$  substitutions per year and relaxation parameter  $\omega = 5$   
74 (Equation 3). Consequently, some branches had many more or less substitutions compared to what  
75 would be expected under a strict clock model with  $\mu = 10$ , and the probabilities of these branches under  
76 this model would be low (Figure 2B). Nevertheless, a root-to-tip regression seemed very satisfactory,  
77 with  $R^2 = 0.94$  and  $p < 10^{-4}$  for a date randomization test (Figure S1).

78 We applied BactDating (Didelot et al. 2018) to reconstruct the dated tree, incorrectly assuming a strict  
79 clock model (Equation 1). The clock rate was estimated to be  $\mu = 10.5$  [9.4;11.6] and the root date  
80 2008.6 [2008.1;2009.1] which is approximately correct. However, when looking at a single sample from  
81 the posterior, the residuals for the branches were not distributed as Normal(0,1) (Figure 2C) and a  
82 QQ plot revealed significant deviation (Figure 2D). The Anderson-Darling test rejects the hypothesis  
83 of standard normality of the residuals ( $p = 3.03 \times 10^{-6}$ ). The same residual analysis performed on  
84 multiple samples from the posterior showed that they all had p-values below 0.05. We repeated the  
85 same analysis incorrectly assuming a strict clock model using LSD (To et al. 2016), node.dating (Jones  
86 and Poon 2017), treedater (Volz and Frost 2017) and TreeTime (Sagulenko et al. 2018), all of which  
87 led to similar results (Figure S2).

88 We applied BactDating again, but this time used the correct additive relaxed clock model (Equation  
89 3). The clock rate estimated to be  $\mu = 11.3$  [8.8;14.1], the root date was 2008.9 [2007.7;2009.8] and  
90 the relaxation parameter was  $\omega = 6.4$  [4.2;8.9], all of which is approximately correct. The residuals  
91 for a single sample from the posterior looked approximately distributed as they should be both when  
92 plotting them against their theoretical distribution (Figure 2E) and when constructing a QQ plot  
93 (Figure 2F). The Anderson-Darling test did not reject the hypothesis of standard normality of the  
94 residuals ( $p = 0.465$ ). Repeating this residual analysis on multiple samples from the posterior showed  
95 that only 4.5% of them had p-values below 0.05.

## 96 Confounding effect of population structure

## 97 Benchmarking

## 98 Real data examples

## 99 DISCUSSION

100 TODO

# 101 MATERIALS AND METHODS

## 102 Molecular clock models

103 The molecular clock model determines the distribution of number of substitutions  $l_i$  on a branch of the  
104 dated tree with duration  $d_i$ . We consider four types of molecular clock models, for each combination  
105 of discrete vs continuous and strict vs relaxed. In the discrete strict clock model (Zuckerandl and  
106 Pauling 1962) with rate  $\mu$ , substitutions occur on the branches as a Poisson process with rate  $\mu$  and  
107 therefore:

$$l_i \sim \text{Poisson}(d_i\mu) \quad (1)$$

108 A continuous version of the strict clock model can be formed based on a Gamma process with the  
109 same mean and variance (Didelot et al. 2021):

$$l_i \sim \text{Gamma}(d_i\mu, 1) \quad (2)$$

110 Strict clock models are based on the assumptions that the substitution rate is constant throughout  
111 the branches of the tree, but this is not always true in which case a relaxed clock model can be used  
112 which allows the rate to vary (Drummond et al. 2006). In particular here we use the additive relaxed  
113 clock model (Didelot et al. 2021), in which  $\mu$  is the mean clock rate and  $\omega$  determines how much this  
114 rate varies on the branches. The discrete version of this model is given by:

$$l_i \sim \text{NegativeBinomial}\left(\frac{d_i\mu}{\omega}, \frac{1}{1+\omega}\right) \quad (3)$$

115 A continuous additive relaxed clock model can again be defined by considering a Gamma process with  
116 the same mean and variance:

$$l_i \sim \text{Gamma}\left(\frac{d_i\mu}{1+\omega}, 1+\omega\right) \quad (4)$$

117 Note that throughout this article Gamma distributions are parametrised by shape and scale and  
118 Negative Binomials by number of successes and probability of success. In the four models we have  
119 that the mean of  $l_i$  is equal to  $d_i\mu$ . The variance of  $l_i$  is equal to its mean in the two strict clock  
120 models, and equal to its mean times  $(1+\omega)$  in the two relaxed clock models.

## 121 Approximate posterior sampling given a point estimate

122 We want to validate a dated phylogeny  $\mathcal{D}$  by comparison with an undated phylogeny  $\mathcal{L}$ . If the dated  
123 phylogeny  $\mathcal{D}$  was sampled from its posterior distribution for example using BactDating (Didelot et al.  
124 2018), then residuals can be calculated directly as described in the next subsection. This analysis can

125 be performed for multiple posterior samples in order to generate a posterior distribution of p-values  
 126 (TODO cites).

127 If on the other hand  $\mathcal{D}$  is the result of maximum likelihood estimation, then first we need to generate  
 128 samples from the posterior before residuals can be computed.<sup>1</sup> To illustrate this, let us first consider  
 129 the discrete strict clock model (Equation 1) and that the true branch durations  $d_i$  are independent  
 130 and identically distribution as:

$$d_i \sim \text{Gamma}(k, \theta) \quad (5)$$

131 Let  $\hat{d}_i$  be a branch length in  $\mathcal{D}$ , on which there are  $l_i$  substitutions in the undated phylogeny  $\mathcal{L}$ . If  $\hat{d}_i$   
 132 is a maximum likelihood estimate of  $d_i$  then  $\hat{d}_i = l_i/\mu$ . By conjugacy of the Gamma prior and Poisson  
 133 likelihood, we can deduce that the posterior of  $d_i$  is:

$$d_i \sim \text{Gamma}\left(k + \hat{d}_i\mu, \frac{\theta}{1 + \theta\mu}\right) \quad (6)$$

134 We can simulate from this distribution to get a posterior sample  $d_i$ , from which we can then compute  
 135 the residuals.

136 TODO non-iid case based on coalescent simulation.

137 Maybe use improper  $\text{InvGamma}(0, \infty)$  prior on  $\alpha$  so that posterior is:

$$\alpha \sim \text{InvGamma}\left(n - 1, \frac{2}{\sum_{i=2}^{2n-1} k_i(k_i - 1)(t_i - t_{i+1})}\right) \quad (7)$$

138 2

## 139 Computation of residuals

140 3

141 Let  $d_i$  be the duration of a given branch in  $\mathcal{D}$  and  $l_i$  be the number of substitutions on the corresponding  
 142 branch of  $\mathcal{L}$ , that is the branch that separates the leaves in the same way. There is a unique  
 143 corresponding branch in  $\mathcal{L}$  for all branches in  $\mathcal{D}$  except for the two branches  $a$  and  $b$  connected to the  
 144 root of  $\mathcal{D}$  for which there is only a single corresponding branch  $x$ . We therefore split the substitutions  
 145 on  $x$  proportionally between the two branches  $a$  and  $b$  by defining:

$$l_a = \frac{l_x d_a}{d_a + d_b} \text{ and } l_b = \frac{l_x d_b}{d_a + d_b} \quad (8)$$

146 The distribution of  $l_i$  given  $d_i$  is given by the molecular clock model. Let us for now consider that the  
 147 distribution is continuous (as in Equations 2 and 4) and we will return later to the discrete case (as

<sup>1</sup>Could mention somewhere case of a summary tree from the posterior distribution (Heled and Bouckaert 2013)

<sup>2</sup>Need to do other models than strict clock model. Conjugacy of priors and posteriors is not as readily available as for Poisson case. May need a Monte-Carlo method (short run of MH algorithm?) to get sample.

<sup>3</sup>Notations are not consistent.  $\mathcal{D}$  sometimes refer to input dated phylogeny and sometimes to sampled dated phylogeny.

148 in Equations 1 and 3). Instead of a specific model, we consider the general case where  $F_i(l_i)$  is the  
 149 cumulative distribution function of  $l_i$  given  $d_i$ . Let  $u_i$  denote the uniform residual for the observation  
 150  $l_i$ , defined as:

$$u_i = F_i(l_i) = p(L_i \leq l_i | d_i) \quad (9)$$

151 If the inference is valid, then the uniform residual  $u_i$  should be distributed as Uniform(0,1), because  
 152 for any random variable  $X$  with cumulative distribution function  $F$  we have that  $U = F(X)$  is  
 153 Uniform(0,1). However, it is difficult to assess how close to zero or one a value needs to be in order to  
 154 be an outlier. We therefore define the normal residuals  $n_i$ , analogous to the residuals commonly used  
 155 in regression models (Cox and Snell 1968; Dunn and Smyth 1996). The normal residuals are obtained  
 156 by transforming the uniform residuals with the inverse of the cumulative distribution function  $\Phi$  of a  
 157 Normal(0,1) random variable:

$$n_i = \Phi^{-1}(u_i) \quad (10)$$

158 If the inference is valid, then the normal residual  $n_i$  should be distributed as Normal(0,1) which is  
 159 more convenient to work with than the Uniform(0,1) for uniform residuals. The uniform and normal  
 160 residuals above can be computed directly when the clock model is continuous (Equations 2 and 4) but  
 161 when the clock model is discrete (Equations 1 and 3) we need to make the following adjustment (Dunn  
 162 and Smyth 1996; Brockwell 2007; Lau et al. 2014):

$$u_i \sim \text{Unif}(F_i(l_i), F_i(l_i + 1)) \quad (11)$$

## 163 Analysis of residuals

164 After computation of the uniform residuals  $u_i$  and normal residuals  $n_i$ , we use several methods to assess  
 165 the validity of the dated phylogeny inference. The uniform residuals  $u_i$  can be plotted as a histogram  
 166 to compare their distribution with the theoretical Uniform(0,1) distribution, but as previously noted  
 167 this can be difficult to interpret. We therefore prefer to use the normal residuals  $n_i$  which can be  
 168 plotted as a histogram to compare their distribution with the theoretical Normal(0,1). A quantile-  
 169 quantile plot (QQ plot) can be used to compare the distribution of the residuals to their theoretical  
 170 distribution. Anderson-Darling test (Lewis 1961) in DescTools R package implements simple hypothesis  
 171 testing (unlike nortest package which is composite test). Use this to test that normal residuals are  
 172 standard normal. Note this is exactly equivalent to testing the uniform residuals against Uniform(0,1).  
 173 Anderson-Darling simple hypothesis testing was used in (Lau et al. 2014) via implementation in package  
 174 ADGofTest, returns same results as DescTools. Both use the same C code from (Marsaglia and  
 175 Marsaglia 2004).

## 176 Data simulation

177 Some simulations using DetectImports (Didelot et al. 2023b).



178 Some simulations using Master (Vaughan and Drummond 2013) to simulate under the structured  
179 coalescent model (Nordborg 1997).

180 Some simulations using mlesky (Didelot et al. 2023a) to simulate each population with non-constant  
181 population size. The size of the  $j$ -th population follows a previously studied model of clonal expansion  
182 (Helekal et al. 2021):

$$N_j(t) = \frac{M_j(t - s_j)^2}{h_j^2 + (t - s_j)^2} [t \geq s_j] \quad (12)$$

183 Note that the square brackets are Iverson brackets. Each population starts at time  $s_j$  with size  
184  $N(s_j) = 0$  and grows logistically up to its maximum  $N_j(\infty) = M_j$ , with  $h_j$  being the time taken to  
185 reach half of this since  $N_j(s_j + h_j) = M_j/2$ .

## 186 Real data

187 TODO

## 188 Implementation

189 We implemented the analytical methods described in this paper in a new R package entitled  
190 *ValidateDating* which is available at <https://github.com/xavierdidelot/ValidateDating> for R  
191 version 3.5 or later. All code and data needed to replicate the results are included in the “run”  
192 directory of the *ValidateDating* repository.

## 193 ACKNOWLEDGEMENTS

194 We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection  
195 Research Unit in Genomics and Enabling Data.

# References

- Baele G, Lemey P, Bedford TBC, Rambaut A, a Suchard M, Alekseyenko AV. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*. 29:2157–2167.
- Baele G, Suchard MA, Rambaut A, Lemey P. 2016. Emerging concepts of data integration in pathogen phylodynamics. *Systematic biology*. 00:1–24.
- Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015. Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*. 30:306–313.
- Bouckaert R, Vaughan TG, Fourment M, Gavryushkina A, Heled J, Denise K, Maio ND, Matschiner M, Ogilvie H, Plessis L, et al. (11 co-authors). 2019. BEAST 2.5 : An Advanced Software Platform for Bayesian Evolutionary Analysis. *PLoS computational biology*. 15:e1006650.
- Bouckaert RR, Drummond AJ. 2017. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evolutionary Biology*. 17:42.
- Brockwell A. 2007. Universal residuals: A multivariate transformation. *Statistics & Probability Letters*. 77:1473–1478.
- Buckby J, Wang T, Zhuang J, Obara K. 2020. Model Checking for Hidden Markov Models. *Journal of Computational and Graphical Statistics*. 29:859–874.
- Cox DR, Snell EJ. 1968. A General Definition of Residuals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*. 30:248–265.
- Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research*. 46:e134–e134.
- Didelot X, Franceschi V, Frost SDW, Dennis A, Volz EM. 2023a. Model design for non-parametric phylodynamic inference and applications to pathogen surveillance. *Virus Evolution*. 9:vead028.
- Didelot X, Fraser C, Gardy J, Colijn C. 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular Biology and Evolution*. 34:997–1007.
- Didelot X, Helekal D, Kendall M, Ribeca P. 2023b. Distinguishing imported cases from locally acquired cases within a geographically limited genomic sample of an infectious disease. *Bioinformatics*. 23:btac761.
- Didelot X, Parkhill J. 2022. A scalable analytical approach from bacterial genomes to epidemiology. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 377:20210246.
- Didelot X, Siveroni I, Volz EM. 2021. Additive uncorrelated relaxed clock models for the dating of genomic epidemiology phylogenies. *Molecular Biology and Evolution*. 38:307–317.
- Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology*. 4:e88.
- Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*. 161:1307–1320.
- Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends in Ecology and Evolution*. 18:481–488.

235 Duchêne D, Duchêne S, Ho SYW. 2015. Tree imbalance causes a bias in phylogenetic estimation of  
236 evolutionary timescales using heterochronous sequences. *Molecular Ecology Resources*. 15:785–794.

237 Duchene S, Duchêne D, Holmes EC, Ho SY. 2015. The performance of the date-randomization test in  
238 phylogenetic analyses of time-structured virus data. *Molecular Biology and Evolution*. 32:1895–1906.

239 Duchene S, Lemey P, Stadler T, Ho SY, Duchene DA, Dhanasekaran V, Baele G. 2020. Bayesian  
240 evaluation of temporal signal in measurably evolving populations. *Molecular Biology and Evolution*.  
241 37:3363–3379.

242 Dunn PK, Smyth GK. 1996. Randomized Quantile Residuals. *Journal of Computational and Graphical*  
243 *Statistics*. 5:236–244.

244 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and  
245 methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0.  
246 *Systematic biology*. 59:307–21.

247 Heled J, Bouckaert RR. 2013. Looking for trees in the forest: Summary tree from posterior samples.  
248 *BMC Evolutionary Biology*. 13:221.

249 Helekal D, Ledda A, Volz E, Wyllie D, Didelot X. 2021. Bayesian inference of clonal expansions in a  
250 dated phylogeny. *Systematic Biology*. p. syab095.

251 Jones BR, Poon AF. 2017. Node.dating: Dating ancestors in phylogenetic trees in R. *Bioinformatics*.  
252 33:932–934.

253 Lau MS, Marion G, Streftaris G, Gibson GJ. 2014. New model diagnostics for spatio-temporal systems  
254 in epidemiology and ecology. *Journal of the Royal Society Interface*. 11:1–10.

255 Lemey P, Rambaut A, Drummond AJ, Suchard MA. 2009. Bayesian phylogeography finds its roots.  
256 *PLoS computational biology*. 5:e1000520.

257 Lewis PA. 1961. Distribution of the Anderson-Darling statistic. *The Annals of Mathematical Statistics*.  
258 pp. 1118–1124.

259 Li WLS, Drummond AJ. 2012. Model averaging and Bayes factor calculation of relaxed molecular  
260 clocks in Bayesian phylogenetics. *Molecular biology and evolution*. 29:751–61.

261 Marsaglia G, Marsaglia J. 2004. Evaluating the Anderson-Darling Distribution. *Journal of Statistical*  
262 *Software*. 9.

263 Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, Holmes MA, Rambaut A,  
264 Welch JJ. 2016. The effect of genetic structure on molecular dating and tests for temporal signal.  
265 *Methods in Ecology and Evolution*. 7:80–89.

266 Navascués M, Emerson BC. 2009. Elevated substitution rate estimates from ancient DNA: Model  
267 violation and bias of Bayesian methods. *Molecular Ecology*. 18:4390–4397.

268 Nordborg M. 1997. Structured coalescent processes on different time scales. *Genetics*. 146:1501–1514.

269 Rambaut A. 2000. Incorporating Non-Contemporaneous Sequences Into Maximum Likelihood  
270 Phylogenies. *Bioinformatics*. 16:395–399.

271 Rieux A, Balloux F. 2016. Inferences from tip-calibrated phylogenies: A review and a practical guide.  
272 *Molecular Ecology*. 25:1911–1924.

273 Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum likelihood phylodynamic analysis. *Virus*  
274 *Evolution*. 4:vex042.

- 275 Stamatakis A. 2015. Using RAxML to Infer Phylogenies. *Current Protocols in Bioinformatics*.  
276 51:6.14.1–6.14.14.
- 277 Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic  
278 and phylodynamic data integration using BEAST 1.10. *Virus Evolution*. 4:vey016.
- 279 To TH, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms.  
280 *Systematic Biology*. 65:82–97.
- 281 Tong KJ, Duchêne DA, Duchêne S, Geoghegan JL, Ho SYW. 2018. A comparison of methods for  
282 estimating substitution rates from ancient DNA sequence data. *BMC Evolutionary Biology*. 18:70.
- 283 Vaughan TG, Drummond AJ. 2013. A stochastic simulator of birth-death master equations with  
284 application to phylodynamics. *Molecular biology and evolution*. 30:1480–93.
- 285 Volz EM, Frost SDW. 2017. Scalable relaxed clock phylogenetic dating. *Virus Evolution*. 3:vex025.
- 286 Wertheim JO, Fourment M, Kosakovsky Pond SL. 2012. Inconsistencies in Estimating the Age of  
287 HIV-1 Subtypes Due to Heterotachy. *Molecular Biology and Evolution*. 29:451–456.
- 288 Zucchini W, MacDonald IL. 2009. *Hidden Markov Models for Time Series: An Introduction Using R*.  
289 Chapman and Hall/CRC.
- 290 Zuckerkandl E, Pauling L. 1962. Molecular Disease, Evolution, and Genic Heterogeneity. In: Kasha  
291 M, Pullman B, editors, *Horizons in Biochemistry*, New York: Academic Press, pp. 189–222.

Rate=1.30e+01,MRCA=2009.35,R2=0.94,p<1.00e-04

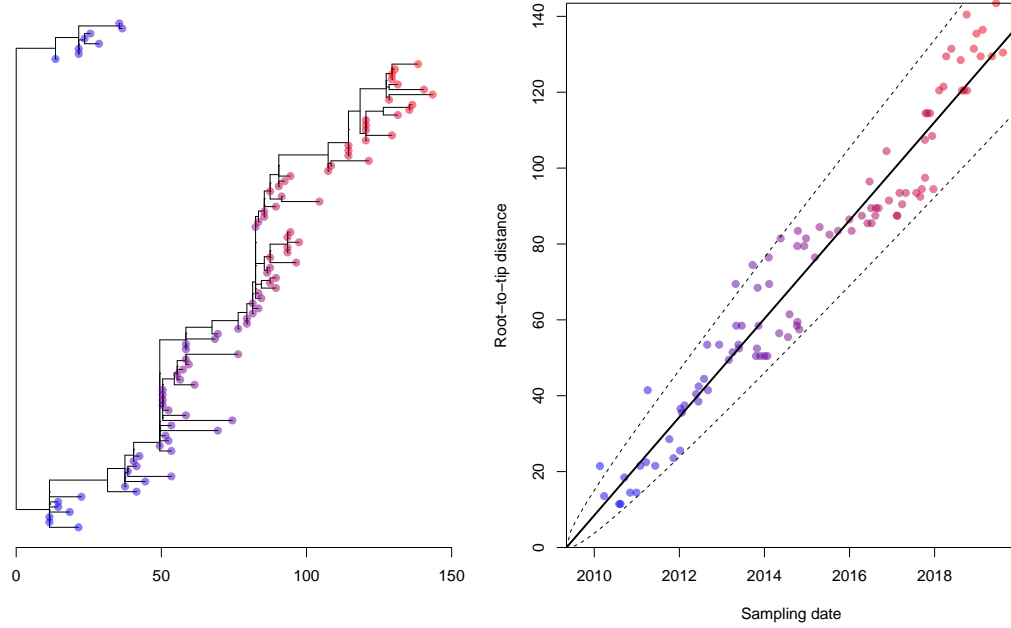


Figure S1: Root-to-tip regression analysis for the motivating example.

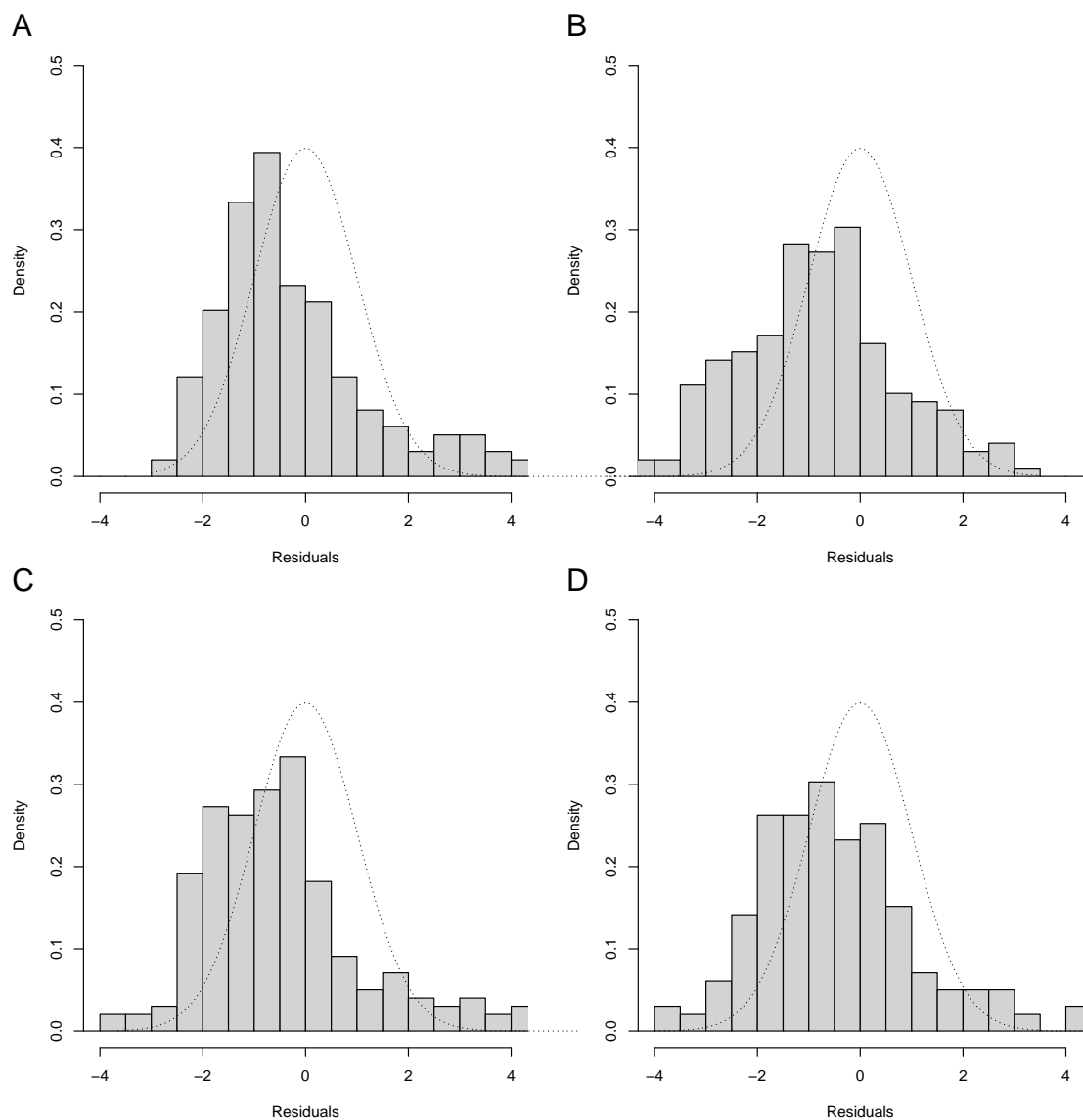


Figure S2: Residuals after application on the motivating example of a strict clock model using LSD (A), node.dater (B), treedater (C) and TreeTime (D).