Supplementary Information
node.dating: Timing the nodes of ancestral dates in phylogenetic trees

Bradley R. Jones[1,2,*] and Art F.Y. Poon[2,3]

[1]Faculty of Health Sciences, Simon Fraser University, Burnaby, V5A 1S6, Canada
[2]BC Centre for Excellence in HIV/AIDS, Vancouver, V6Z 1Y6, Canada
[3]Department of Medicine, University of British Columbia, V5Z 1M9, Canada
[*]Corresponding author (email: brj1@sfu.ca)

1 Simulation

To verify the accuracy of `node.dating`, we applied it to simulated data. we simulated 50 phylogenetic trees using a birth-death model with the the R package *TreeSim* (Stadler, 2015) using the parameters: $\lambda = 5.116 \times 10^{-2}$ day^{-1}, $\delta = 5.006 \times 10^{-2}$ day^{-1}, and $s = 5.237 \times 10^{-3}$. we then applied a strict molecular clock to trees wt the R package *NELSI* (Ho et al., 2015) using the parameters: $\mu = 1.964 \times 10^{-4}$ and $\sigma = 1.417 \times 10^{-5}$ substitutions per generation. we used these trees to generate simulated HIV sequences with *INDELible* 1.03 (Fletcher and Yang, 2009) using a HKY85 nucleotide substitution model (Hasegawa et al., 1985) with a stationary distribution of 0.42, 0.15, 0.15, 0.28 for A, C, G, T respectively and a transitional bias of 8.5. Finally we reconstructed phylogenetic trees from the sequences using *RAxML* 8.2.4 (Stamatakis, 2014) with the GTR model and rooted the trees using the `rtt` function of the *APE* package. This process is engineered to replicate phylogenetic trees derived from real data.

2 Weighted RMSE

The dates of the MRCA of each pair of tips of the original birth-death tree were saved and compared with the results of `node.dating` using a weighted root mean squared error (RMSE) as the error metric. Specifically:

$$\text{RMSE} = \sqrt{\frac{\sum_{1\leq i<j\leq N} w_{i,j}\left(d_{\text{MRCA}_{t_r}(i,j)} - \delta_{\text{MRCA}_{t_p}(i,j)}\right)^2}{\sum_{1\leq i<j\leq N} w_{i,j}}}$$

where t_r and t_p are the real (resp. predicted) phylogenies each with N tips; $\text{MRCA}_t(i,j)$ is the MRCA of tip i and j in the phylogeny, t; d_m and δ_m are the real (resp. predicted) dates of the MRCA, m; and $w_{i,j}$ is the weight of the pair of tips, i and j, and is given by:

$$w_{i,j} = \sqrt{1/\left(x_{\text{MRCA}_{t_r}(i,j)}\, y_{\text{MRCA}_{t_p}(i,j)}\right)}$$

with x_m and y_m as the number of pairs of tips in the real (resp. predicted) phylogeny whose MRCA is m. The MRCA was compared instead of the date of each internal node because the tree topologies may change after applying *RAxML*.

For our analysis we considered the mean of the RMSE for each phylogeny.

3 Acquisition of real data

Sequences were aligned using MUSCLE 3.8.31 (Edgar, 2004) and inspected and cleaned using AliView (Larsson, 2014). We trimmed the alignments so that each sequence had at least 50% coverage over each base. We reconstructed the phylogeny of the patient's sequences using *RAxML* 8.2.4 (Stamatakis, 2014) and rooted the tree using the `rtt` function of APE.

References

Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5), 1792–1797.

Fletcher, W. and Yang, Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8), 1879–1888.

Hasegawa, M. et al. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2), 160–174.

Ho, S. Y. W. et al. (2015) Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Molecular Ecology Resources*, 15, 688–696.

Larsson, A. (2014) AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22), 3276–3278.

Stadler, T. (2015) TreeSim: Simulating phylogenetic trees. *R package version 2.2*.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.