MOLECULAR ECOLOGY**Comment on Rieux and Balloux: calibration from tip-dating
can compromise topological accuracy and evolutionary
inference**

Journal:	<i>Molecular Ecology</i>
Manuscript ID	MEC-16-0344.R1
Manuscript Type:	Comment
Date Submitted by the Author:	n/a
Complete List of Authors:	Emerson, Brent; IPNA-CSIC, Ecology and Evolution; Alvarado-Serrano, Diego Hickerson, Michael; City College of New York, Biology
Keywords:	μ , θ , Molecular Evolution, substitution rate, Bayesian inference

Comment on Rieux and Balloux: calibration from tip-dating can compromise topological accuracy and evolutionary inference

Brent C. Emerson^{1,2*}, Diego F. Alvarado-Serrano³ and Michael J. Hickerson^{3,4,5}

1. Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), C/Astrofísico Francisco Sánchez 3, La Laguna, Tenerife, Canary Islands, 38206, Spain.

2. School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.

3. Biology Department, City College of New York, New York, NY, 10031, USA.

4. The Graduate Center, City University of New York, New York, NY, 10016, USA.

5. Division of Invertebrate Zoology, American Museum of Natural History, New York, NY 10024, USA.

* Contact author: bemerson@ipna.csic.es

Abstract

We contribute to the recent review of Rieux and Balloux (2016) on inferences from tip-dated phylogenies by developing their discussion on the influence of population size (N_e) under panmixia for the estimation of substitution rate (μ). We highlight how phylogenetic trees inferred with tip-dated sequences under large panmictic N_e tend to erroneously enforce an age-based coalescent pattern on the posterior distribution of trees, that in turn results in systematically inflated estimates of μ . We discuss the consequences of this, and suggest how to accommodate the issue in the short and long term.

For Review Only

In their recent review, Rieux and Balloux (2016) summarise the current state-of-the-art for tip dating, and put forward a “how to” guide to perform such analyses. While their review is both comprehensive and timely, we feel the need to elaborate on a consequential element of tip-dating that received only limited attention. Rieux and Balloux (2016) acknowledge that rate estimates from tip-dated phylogenies may be upwardly biased under less than simple demographic scenarios, such as structured populations (Navascués & Emerson 2009). One potential explanation put forward for this bias by Emerson and Navascués (2009) is that population structure may produce trees where ancient DNA (aDNA) and modern DNA (mDNA) segregate in different lineages. Under such a scenario tree shape may be incorrectly inferred due to Felsenstein’s Pulley Principle (Felsenstein 1981), where different roots obtain the same likelihood but can lead to higher inferred mutational events between sampling times, and thus higher estimated substitution rate (μ) (Navascués & Emerson 2009).

Rieux and Balloux (2016) also cite Emerson *et al.* (2015) as showing that larger effective population size (N_e) may also result in topologies where aDNA and mDNA segregate in different lineages. This is not correct, as Emerson *et al.* (2015) demonstrate with simulations that aDNA and mDNA tend to segregate randomly with respect to gene tree lineages under larger panmictic N_e values. These later authors further show that phylogenetic trees inferred by BEAST for tip-dated sequences that were generated under larger panmictic N_e tend to enforce an age-based coalescent pattern on the posterior distribution of trees, that in turn results in systematically inflated estimates of μ (Fig. 1, Emerson *et al.* 2015). An age-based coalescent pattern is where DNA sequences of older age diverge closer to the root, and thus cluster together more basally within a tree, whereas younger sequences tend to diverge and cluster together more terminally, as is expected under small N_e . However, as N_e increases topology will depart from an age-based coalescent because lineages of different ages can easily co-segregate in the same population (Box 1, Emerson *et al.* 2015). Here it is useful to point out that for a coalescent-based method, such as employed by BEAST, the gene genealogy can be viewed as a nuisance variable when the objective is to estimate mutation rate (Felsenstein 1988). In the definition of the likelihood of the parameters given the data, $L(\Theta|D)$ is proportional to $\Pr(D|\Theta)$, where $\Pr(D|\Theta) = \Pr(D|G,\Theta)\Pr(G|\Theta)$ (summed across all possible genealogies), D is the data, G is the tip-dated gene genealogy and Θ ; Θ is the model parameters (including $\theta = 2N_e\mu$).

While gene genealogy is a nuisance variable within the context of the joint estimation of model parameters N_e and μ by BEAST, it is also consequential. The tip-dated panmictic coalescent model employed in BEAST allows decoupling the posterior estimates of θ into N_e and μ by using the tip ages of the posterior distribution of trees. The results of Emerson *et al.* (2015) suggest that when true N_e is large, thereby resulting in gene trees deviating from an age-based coalescent pattern, the small prior for N_e leads to the sampling of suboptimal trees that are a better fit to an age-based coalescent pattern. When the suboptimal trees introduce homoplasy and associated higher numbers of mutations to accommodate the observed patterns of polymorphism, the prediction is for μ to be over-estimated (Emerson *et al.* 2015).

What we urge is not the abandonment of tip-dating, but a more thorough evaluation of parameter inference when tip-dating is undertaken. In the long-term, more is needed regarding the understanding of the relationships among the prior distributions as well as the posterior distributions of tip-dated trees, N_e and μ when using BEAST. Until then, it would seem wise to report and compare the posterior distribution of tip-date constrained tree topologies with that of unconstrained topologies, and this can be readily achieved by comparing BEAST-derived tip-date constrained topologies with those derived from MrBayes (e.g. Appendix S1, Emerson *et al.* 2015). Although this may not be an optimal solution, it is one approach to assess the magnitude of the problem. It is relevant to point out that as even only a small number of inferred mutations may influence the estimation of μ , tests for broad topological agreement between unconstrained and constrained topologies may not be sufficiently informative. Specific assessment of topological differences, and their consequences with regard to the amount of inferred mutational change within the tree will be required. The issue we highlight both here and in Emerson *et al.* (2015) is likely to be less consequential for viral analyses, where N_e is expected to be small, and thus an age-based pattern of coalescence is expected to be more common. However, it is still pertinent to demonstrate, rather than assume, that inferred trees are not confounding parameter inference.

References

Emerson BC, Serrano DFA, Hickerson MJ (2015) Model misspecification confounds the estimation of rates, and exaggerates their time dependency. *Molecular Ecology*, **24**, 6013-6020.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach.

Journal of Molecular Evolution, **17**, 368–376.

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annual*

Review of Genetics, **22**, 521-565.

Navascués M, Emerson BC (2009) Elevated substitution rate estimates from ancient DNA:

model violation and bias of Bayesian methods. *Molecular Ecology*, **18**, 4390-4397.

Rieux A, Balloux F (2016) Inferences from tip-calibrated phylogenies: a review and a practical

guide. *Molecular Ecology*, **25**, 1911-1924.

For Review Only