POTS VERSUS CLOCKS: INTEGRATING ARCHAEOLOGICAL EVIDENCE WITH BAYESIAN ESTIMATES OF DIVERGENCE TIMES IN OCEANIC LANGUAGES

BENEDICT KING[*1], MARY WALWORTH[1], AYMERIC HERMANN[1,2], SIMON J. GREENHILL[1,3] and RUSSELL D. GRAY[1]

[*]Corresponding Author: benedict_king@eva.mpg.de

[1]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]CNRS, Paris, France

[3]Faculty of Science, University of Auckland, New Zealand

Bayesian phylogenetics provides a quantitative method with which to estimate the timing of language divergences (Hoffmann *et al.* 2021). In such analyses, a "relaxed clock" of lexical change is used to date nodes in the language tree. A relaxed clock assumes an approximately predictable rate of lexical innovation, while allowing the rate to vary across the branches of the tree. In order to date a phylogenetic tree, the clock needs to be calibrated by integrating prior information to constrain the age of particular nodes in the tree (Maurits *et al.* 2020). The expansion of Austronesian speakers across Oceania is a topic of research in which Bayesian methods should provide valuable insight. Since the publication of the Austronesian phylogenetic tree of Gray *et al.* (2009), the Austronesian Basic Vocabulary Database (Greenhill *et al.* 2008) has undergone significant expansion. New data, combined with improvements in analysis methods and new archaeological dates should provide further insights into the timing and expansion of Austronesians into Oceania.

To obtain a time-scaled tree of Oceanic languages, we analysed lexical data in the form of cognate sets in the Bayesian phylogenetic software BEAST2 (Bouckaert *et al.* 2018). The tree was time-scaled using a relaxed clock and calibration distributions on the divergence of the Oceanic, Polynesian and Eastern Polynesian languages. Despite the relatively informative prior probability distributions on divergence time, our analysis produced an estimate for the divergence of Oceanic languages that strongly conflicted with the archaeological record (Fig. 1), a result that holds for analyses using a variety of different models of tree shape. This ancient date estimated for Oceanic languages evokes the "Rocks versus Clocks" debate in evolutionary biology, in which molecular clocks frequently produce ancient divergence dates that can predate the earliest fossil record of a group by 10s or 100s of millions of years (Benton 1999, Puttick *et al.* 2016). Surprisingly, these older dates for the Oceanic tree are found despite no apparent signal in the lexical data supporting such a divergence.

We investigated the ability of relaxed clock analyses to recover dates within an age range consistent with the archaeological data using a series of simulations. Data was simulated on a tree taken from an analysis in which the divergence time for Oceanic was enforced to be no more than 3300 years. Although the relaxed clock performed reasonably when simulations were performed according to randomised branch rates, we found that when data were simulated using the observed branch rates, the age of Oceanic was consistently overestimated, as were other nodes in the phylogeny. The clock model showed some degree of statistical inconsistency, in that larger simulated datasets returned more inaccurate age estimates. We also found that the relaxed clock was unable to correctly estimate the rates of lexical innovation on branches, consistently underestimating rates on branches simulated with a rapid rate. The relatively old date estimated for Oceanic may therefore be a statistical artefact resulting from a high rate of lexical innovation at the origin of the Central Pacific subgroup of Oceanic.

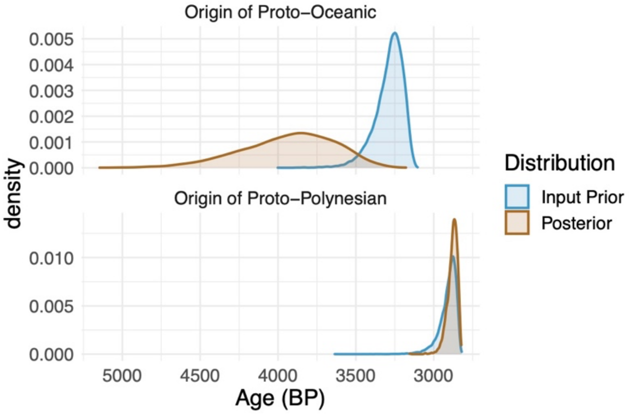

Figure 1: Comparison of prior age calibrations (based on archaeological data) with ages estimated from the data (posterior). For Oceanic as a whole, the estimated age strongly conflicts with the prior.

The inability of these analyses to recover the known age of Oceanic in a simulation scenario calls into question the accuracy of ancient dates estimated from relaxed clocks both in linguistics and evolutionary biology. In both fields, groups chosen for study might often be expected to show rapid rates of evolution (lexical or molecular) at the base of the tree (Beaulieu, 2016), for example due to migration to new territories or recovery from mass extinctions. Our study suggests that relaxed clocks overestimate the age of linguistic groups in at least some situations.

References

Beaulieu, J. M., O'Meara, B. C., Crane, P., & Donoghue, M. J. (2015). Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Systematic biology, 64*(5), 869-878.

Benton, M. J. (1999). Early origins of modern birds and mammals: molecules vs. morphology. *BioEssays*, *21*(12), 1043–1051.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ... & Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, *15*(4), e1006650.

Gray, R. D., Drummond, A. J., & Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, *323*(5913), 479–483.

Greenhill, S. J., Blust, R., & Gray, R. D. (2008). The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*, *4*, EBO-S893.

Hoffmann, K., Bouckaert, R., Greenhill, S. J., & Kühnert, D. (2021). Bayesian phylogenetic analysis of linguistic data using BEAST. *Journal of Language Evolution*, *6*(2), 119–135.

Maurits, L., de Heer, M., Honkola, T., Dunn, M., & Vesakoski, O. (2020). Best practices in justifying calibrations for dating language families. *Journal of Language Evolution*, *5*(1), 17–38.

Puttick, M. N., Thomas, G. H., & Benton, M. J. (2016). Dating placentalia: Morphological clocks fail to close the molecular fossil gap. *Evolution*, *70*(4), 873–886.