Running head: Simulated examination of priors in biogeographic dating

Title: Reciprocal influence of biological and geographic data in biogeographic dating methods

Authors: Luke Sparreo1, Ana Bedoya1,\*

1Center for Biodiversity and Evolution, New York Botanical Garden, 2900 Southern Boulevard, Bronx, NY, USA 10458

Correspondence: 2900 Southern Boulevard, Bronx, NY, USA 10458, E-mail: [abedoya@nybg.org](mailto:abedoya@nybg.org)

# Abstract

X

Keywords: biogeography, geogenomics, dating, molecular, simulated

Biological and geographic information have long reciprocally informed hypotheses on one another (Bueno-Hernández et al., 2023). Traditionally, pattern-based comparisons between geographic and biological distribution data informed most biogeographic hypotheses. Through the development of different strategies in statistical methods applying fossil calibrations to phylogenies (Parham et al., 2012), and an increase in high-quality DNA data (Rieux & Balloux, 2016), biogeographic research now not only involves searching for pattern congruence but also may explicitly test data-driven hypotheses. In fact, many large clades across the tree of life now are now broadly accepted as reliably dated (dos Reis et al., 2012; Harris & Davies, 2016). Research focused on dating phylogenies has recently trended towards applying biogeographic priors to time phylogenetic models. In both theoretical and empirical (Landis, 2017) systems, tip-based methods have been successful in using biogeographic information to calibrate dated phylogenies. Like fossil tip-dating methods, the observed biogeographic states of present organisms can successfully create a posterior distribution of speciation times to calibrate trees. When combined with known historic geographic priors, such as the advent of islands (Landis et al., 2018), calibration can be further refined.

While geographic history has now been repeatedly successful in analyzing biological data, research on the inverse of these reciprocal analyses has only just begun to inspire further investigation. The rapidly expanding resources of high-quality genetic data have led researchers to use molecular information to address questions about historic geographic processes. In “geogenomics”, genetic data is used to test or constrain geological hypotheses (Baker et al., 2014), allowing genetic information to fill in knowledge gaps where other paleogeographic methods are limited. Geogenomic methods do not just confirm or deny the validity of biogeographic patterns, but specifically aim to understand the mechanisms by which biological and geological forces act upon one another (Dolby et al., 2022). This geogenomic framework provides opportunities to model biogeographic history from molecular data, such as where topographic disruptions existed in presently continuous landscapes using phylogenetic divergence of extant species (Sanín et al., 2022), and the configuration of drainage basins over time (Bedoya et al., 2021). Further unexplored empirical hypotheses remain to be tested, such as using genetic data to predict timing of geological events like volcanic eruptions or landslides (Dolby et al., 2022). Yet, theoretical testing of the geogenomic method, especially the reciprocal influence of the input biological and geographic datasets, is lacking. We address this issue here with a simulated dataset.

In our model scenario, a river-dwelling angiosperm lineage occurs in a river basin that has experienced two historic biogeographic barriers. This model group of aquatic riparian species is suitable for analyses, because individuals do not migrate and species are restricted to highly specific and diagnosable ranges. We assume a barrier, such as mountain uplift, creates a branching in a river ecosystem at three and one million years before present. The ensuing three distinct river branches allow for speciation, with a phylogeny whose branches mirror the river ecosystem (Fig. 1). We use an existing methodological framework (Landis, 2017) to examine the influence of different types of biological and geographic priors on one another in biogeographic dating analyses. In using this simulated dataset with known evolutionary and biogeographic history, we test the potential for well-calibrated phylogenies to estimate biogeographic events. We also examine the impact of introducing gene flow to biological data, and the impact of incorrect assumptions in geographic history.

Methods

## *Demographic and molecular simulation*

A model phylogeny containing three taxa was generated in python using the packages msprime (Baumdicker et al., 2022), NumPy (Harris et al., 2020). The model phylogeny has taxon C splitting from taxa A and B three million years ago and taxon A splitting from taxon B one million years ago. Each taxon has a generation time of one year. Model sequences of 10,000 base pairs were generated for each taxa using the HKY model, which was used to limit the number of assumptions about our simulated dataset. We selected a mutation rate of 1e-8 and a recombination rate of 4.8e-8, realistic for angiosperms (Koch et al., 2000; Salomé et al., 2011). Divergence times were calculated for each node. All invariant sites were filled in with randomly selected bases. In one demographic model, gene flow between taxa was never allowed. In a second demographic model, gene flow occurred between ingroup taxon A and outgroup taxon C from 0.8 to 0.6 million years ago. Demographic models were visualized in python using the package DemesDraw, developed for the package Demes (Gower et al., 2022).

## *Biogeographic dating*

Biogeographic dating analyses for the model data was performed in RevBayes (Höhna et al., 2016) according to the Dispersal-Extinction-Cladogenesis (DEC) model (Landis, 2017; also see Landis et al., 2018). Scripts were adapted from the publicly available RevBayes analyses used in these studies. In this model, molecular sequence data, a phylogeny, present ranges and distances, and geological epoch times are defined as inputs to jointly estimate phylogeny and biogeography. Two Markov chains Monte Carlo were run with 10 million generations for each of the ten model scenarios. For all of the model scenarios, we manually set the root age of the tree to a uniform distribution between three and four million years ago and set the extirpation rate to zero. Results were visualized using Tracer v1.7.2 (Rembaut et al., 2018) and convergence assessment was examined with the R package Convenience (Fabreti & Höhna, 2022) to confirm adequate run times. The two generated log and tree files for each model were combined with the cat command after applying a 0.25 burn-in on each file, resulting in a concatenated file with 15 million generations. The final generated trees were visualized in FigTree v1.4.4.

First, we ran five biogeographic knowledge-level scenarios (Fig. 2); uniform-informed, normal-informed, normal-uninformed, normal-incorrect, and unknown. We then ran the same five biogeographic knowledge-level scenarios with the second demographic model that allowed for a period of gene flow. After finding comparable results with dispersal models of dispersal = 0.05 and 0.10, we report simulations with the higher dispersal model to best reflect biological reality.

# Results

# Discussion

Our results found that models with a flat, unknown geological prior performed nearly as well at estimating divergence times as models with a uniform or normal geological prior centered around correct times. Correct, well-informed simulated data successfully being able to predict the timing of the emergence of biogeographic barriers without prior historic geographic information suggests exciting opportunities to test new biogeographic hypotheses. These results confirm the critical idea in geogenomics that biology and geography may reciprocally inform each other. Furthering genomic data acquisition and empirical analyses in model systems such as montane or riparian species has been proposed (Bedoya, 2024) as the next steps in furthering the empirical testing of reciprocal geographic-biological dating methods.

With the introduction of incorrect prior information, however, models performed significantly worse in predicting geological history. This supports previous studies that have indicated incorrect assumptions within the DEC (Magalhaes et al., 2021) and fossilized birth-death (Barido-Sottani et al., 2020) models have large impacts on inference in biogeographic estimations. Further work should explore the impact of these historic geographic assumptions in other models like DEC+J (Matzke, 2014), which may not accurately parameterize the rates of speciation (Ree & Sanmartin, 2018). Researchers should be conservative and cautious when creating strict priors in geogenomic methods, especially considering the fact that with our accurate simulated molecular data, a flat geological prior performed nearly as well as correctly-informed priors.

When demographic models allowed for gene flow, the predicted cladogenesis times were found to be significantly younger, regardless of the historic geographic prior used. While gene flow may interfere with inferring the correct biogeographic history, detecting the influence of gene flow itself may be a tool to understand the history of past geographic connections. EXPAND.

**References**

**Figures**

**Figure 1:** Biogeographic model for simulated organisms. Dotted lines represent known biogeographic dates of uplift marking epoch transitions at three and one million years ago.

**Figure 2:** Five biogeographic knowledge-level scenarios. Red dotted line represents the true time for the node where taxon C splits from taxa A and B, and green dotted line represents the true time for the node where taxon A splits from taxon B.