Running title: Dating geological history is successful with simulated molecular data

# Introduction:

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 focus on 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 to 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 biogeographic 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 strictly 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 is lacking. We address this issue here with a simulated dataset.

We use an existing methodological framework (Landis, 2017) to ask if a well-supported, accurately-dated molecular phylogeny estimate the dates of biogeographic events without prior information. Answering this question will support the increasing interest in and need for applying biological data to answer biogeographic questions. In using this simulated dataset with known evolutionary and biogeographic history, we test the potential for well-calibrated phylogenies to estimate biogeographic events. In a 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 (Figure 1).

**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.

**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 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 files for each model were combined using LogCombiner in BEAST X v10.X (Suchard et al., 2018). The final generated trees were visualized in FigTree v1.4.4.

First, we ran five biogeographic knowledge-level scenarios (Figure 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.

**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.

**Results:**

**Figure X:** Consensus trees for 5 models without demographic history of gene flow

**Figure X:** Consensus trees for 5 models with demographic history of gene flow

**Discussion:**

Perfect simulated data successfully being able to predict the timing of the emergence of biogeographic barriers without prior information suggests exciting opportunities to test new biogeographic historical hypotheses. In fact, models with a flat, unknown prior performed just as well as models with a uniform or normal geological prior centered around correct times. 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 the idea that researchers should be conservative and cautious when creating strict priors in geogenomic methods.

Talk about gene flow results. Gene flow is not a problem to be solved, but another tool to understand past geographic connections! But also, talk about limitations.

**References:**

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretzschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., Quinto-Cortés, C. D., … Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3), iyab229. <https://doi.org/10.1093/genetics/iyab229>

Bedoya A. M. (2024). Botany and geogenomics: Constraining geological hypotheses in the neotropics with large-scale genetic data derived from plants. *American journal of botany*, *111*(4), e16306. <https://doi.org/10.1002/ajb2.16306>

Bedoya, A. M., Leaché, A. D., & Olmstead, R. G. (2021). Andean uplift, drainage basin formation, and the evolution of plants living in fast‐flowing aquatic ecosystems in northern South America. *New Phytologist*, *ss*(5), 2175-2190.

Bueno-Hernández, A., Barahona, A., J Morrone, J., Espinosa, D., & Juárez-Barrera, F. (2023). Historiographical approaches to biogeography: a critical review. History and philosophy of the life sciences, 45(3), 27. <https://doi.org/10.1007/s40656-023-00580-9>

Dolby, G. A., Bennett, S. E. K., Dorsey, R. J., Stokes, M. F., Riddle, B. R., Lira-Noriega, A., Munguia-Vega, A., & Wilder, B. T. (2022). Integrating Earth-life systems: a geogenomic approach. Trends in ecology & evolution, 37(4), 371–384. https://doi.org/10.1016/j.tree.2021.12.004

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C., & Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings*. *Biological sciences*, *279*(1742), 3491–3500. <https://doi.org/10.1098/rspb.2012.0683>

Fabreti, L. G., & Höhna, S. (2022). Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *Methods in Ecology and Evolution*, *13*(1), 77-90.

Gower, G., Ragsdale, A. P., Bisschop, G., Gutenkunst, R. N., Hartfield, M., Noskova, E., Schiffels, S., Struck, T. J., Kelleher, J., & Thornton, K. R. (2022). Demes: a standard format for demographic models. *Genetics*, *222*(3), iyac131. https://doi.org/10.1093/genetics/iyac131

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard- Marchant, P., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, L. W., & Davies, T. J. (2016). A Complete Fossil-Calibrated Phylogeny of Seed Plant Families as a Tool for Comparative Analyses: Testing the 'Time for Speciation' Hypothesis. *PloS one*, *11*(10), e0162907. <https://doi.org/10.1371/journal.pone.0162907>

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, *65*(4), 726–736. <https://doi.org/10.1093/sysbio/syw021>

Koch, M. A., Haubold, B., & Mitchell-Olds, T. (2000). Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in Arabidopsis, Arabis, and related genera (Brassicaceae). *Molecular biology and evolution*, *17*(10), 1483–1498. https://doi.org/10.1093/oxfordjournals.molbev.a026248

Landis M.J. 2017. Biogeographic dating of speciation times using paleogeographically informed processes. *Systematic Biology*, *66*(2), 128–144. 10.1093/sysbio/syw040

Landis, M. J., Freyman, W. A., & Baldwin, B. G. (2018). Retracing the Hawaiian silversword radiation despite phylogenetic, biogeographic, and paleogeographic uncertainty. *Evolution; international journal of organic evolution*, *72*(11), 2343–2359. <https://doi.org/10.1111/evo.13594>

Parham, J. F., Donoghue, P. C., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., ... & Benton, M. J. (2012). Best practices for justifying fossil calibrations. *Systematic Biology*, *61*(2), 346-359.

Rambaut A., Drummond A.J., Xie D., Baele G. and Suchard M.A. (2018). Posterior summarisation in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*. syy032. doi:10.1093/sysbio/syy032

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular ecology*, *25*(9), 1911–1924. <https://doi.org/10.1111/mec.13586>

Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*, *108*(4), 447–455. <https://doi.org/10.1038/hdy.2011.95>

Suchard M.A., Lemey P., Baele G., Ayres D.L., Drummond A.J. & Rambaut A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus* *Evolution*, *4*, vey016. DOI:10.1093/ve/vey016