Running head: Simulated examination of priors in biogeographic dating

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

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

Biological and geographic data have long reciprocally informed one another. In this study, we simulate 10 simple molecular phylogenies with known histories and date them with historic biogeographic priors of varying information levels and accuracy in order to test the influence of biogeographic knowledge and molecular data on phylogenetic dating. In our simulations, models prohibiting gene flow with unknown biogeographic histories performed just as well as models prohibiting gene flow with correctly informed biogeographic priors. The result that accurate molecular data performs similarly under correct and unknown geographic knowledge shows promise for the utilization of the geogenomic method, which has been used to date geographic events with biological data. However, the introduction of gene flow or the incorporation of incorrect historic biogeographic data shifted the accuracy of our phylogenetic dating. These findings show that systematists must remain cautious when conducting biogeographic dating.

Keywords: biogeography, geogenomics, dating, molecular, simulation

The reciprocal influence of biological and geographic information laid the foundation of the biogeographic field and has allowed for centuries of both observations and testable hypotheses (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.

**Figure 1**

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 (Fig 2.). 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.

**Figure 2**

Materials & 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 for each pair of the two model runs 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 then 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.

We ran five biogeographic knowledge-level scenarios (Fig. S1); 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

**Figure 3**

When using correctly informed geological priors (Figs. S2, S3, S7, S8), the age distribution for node two began at the true date of one million years ago (Table S1), regardless of the introduction of gene flow after the split of taxa A and B at node two. However, the age distribution node two became much younger when gene flow was incorporated into our demographic model. The absolute age of node two was predicted to be about 0.7 million years younger in models permitting gene flow. For correctly informed models, the choice between uniform and normal paleogeographic models had little influence on absolute node age regardless of gene flow, but the normal model did have a slightly older age distribution for node two when gene flow was prohibited.

In the geology unknown models (Figs. S4, S9), when both prohibiting and permitting gene flow, the models found an absolute age for both nodes similar (around 0.1 million years) to the correctly informed models. Like the correctly informed model, the introduction of a short period of gene flow in the demographic history of geology unknown models had a large influence on the predicted absolute age of node two. The absolute age of node two was found to be around 0.5 million years younger when gene flow was included when compared to the unknown model prohibiting gene flow. Additionally, the age distributions for node two were similar to the correctly informed models when gene flow was permitted. However, in the unknown model prohibiting gene flow, the age distribution for node two was dated more broadly than the correctly informed models that prohibited gene flow.

When incorrect geological priors were used in the model (Figs. S5, S6, S10, S11), results for absolute node age and the age distribution for node two shifted in models prohibiting gene flow. The absolute age for node two was about 0.5 years older in both incorrect models prohibiting gene flow, and the age distribution for node two was dated as older as well. Interestingly, in the models that permitted gene flow, the incorrect, correctly informed, and unknown models all produced similar (around 0.1 million years) results. Predictably, the estimated absolute age for node one remained roughly unchanged, as it was not manipulated in the epoch time estimation nor the demographic model. There were no notable differences in the age estimations between using the uniform or normal distributions in the incorrect models.

# Discussion

## *Geologically unknown prior data dates phylogenies comparably to accurate geological priors*

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.

While the strong performance of the molecular data is exciting, it should be noted that in our simulations, our molecular data was “perfect”: a model scenario in which genetic sequences had no missing data. Additionally, we assume our taxon sampling was complete; there were no unknown extant taxa nor extinct taxa in our group that lacked molecular data. Yet, has been thought that missing data at the molecular (Roure et al., 2013) and taxonomic levels (Zwickl & Hillis, 2002) may negatively influence the accuracy of phylogenies, which could in turn create issues downstream in dating methods. Additionally, our model dataset lacked common informatic errors like inaccurate molecular alignment (Ogden & Rosenberg, 2006) or erroneous taxonomic identification (de Almeida et al., 2023) both of which may lead to inaccurate phylogenies and biogeographic dating. In realistic datasets, imperfect molecular data may interfere with the ability of molecular data to accurately test geographic hypotheses.

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. Model systems such as these that lack known demographic histories of gene flow, have strongly-supported biogeographic histories, and already have accepted dated phylogenies through adequate fossil calibrations (such as Landis et al., 2021) will be necessary to explore the possibilities of dating geographic events through biological data.

## *Incorrect geographic historical data alters dating accuracy*

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. Considering debate continues over the timing of influential biogeographic events (White et al., 2013; Zheng, 2015), the selection of accurately resolved paleogeographic priors is critical. Further work should also explore the impact of these historic geographic assumptions in other biogeographic models like DEC+J (Matzke, 2014), which have been debated regarding ability to accurately parameterize the rates of speciation (Ree & Sanmartin, 2018), as the accuracy of model assumptions likely plays a large role in the accuracy of phylogenetic dating results.

Researchers should be conservative and cautious when creating strict priors in biogeographic dating methods, especially considering the fact that with our accurate simulated molecular data, a flat geological prior performed nearly as well as correctly-informed priors.

## *Demographic models that allow gene flow alter dating accuracy*

When demographic models allowed for gene flow, the predicted cladogenesis times were found to be significantly younger, regardless of the historic geographic prior used. The fact that all models allowing gene flow, regardless of the biogeographic information present, produced similar results suggests that the existence of gene flow in molecular data creates a significant barrier to the accurate implementation in using biogeographic data to date phylogenies. Current work continues to reveal histories of introgression (Stull et al., 2023), including from extinct lineages (Tricou et al., 2022), which did not exist in our model system. Researchers should take caution in using biogeographic priors to date phylogenies with known or suspected gene flow.

While gene flow may interfere with inferring the correct biogeographic history, further research may develop methods that allow systematists to properly date phylogenies of systems that include rampant gene flow. In fact, if properly diagnosed, the detection of gene flow itself may be a tool to understand the history of past biogeographic connections, as gene flow is more likely to occur between species that have geographic overlap (Liu et al., 2024).

# Acknowledgements

Code used to run our model simulations was adapted from existing code produced in Landis, 2017 and Landis et al., 2018.

# Literature Cited

de Almeida, R. F., Pellegrini, M. O. O., de Morais, I. L., Simão-Bianchini, R., Rattanakrajang, P., Cheek, M., & Simões, A. R. G. (2023). Barking up the wrong tree: the dangers of taxonomic misidentification in molecular phylogenetic studies. *Plant ecology and evolution*, *156*(2), 146–159.

Barido-Sottani, J., van Tiel, N. M. A., Hopkins, M. J., Wright, D. F., Stadler, T., & Warnock, R. C. M. (2020). Ignoring fossil age uncertainty leads to inaccurate topology and divergence time estimates in time calibrated tree inference. *Frontiers in ecology and evolution*, *8*.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., et al. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, *220*(3).

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.

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. *The New Phytologist*, *232*(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.

Dolby, G. A., Bennett, S. E. K., Dorsey, R. J., Stokes, M. F., Riddle, B. R., Lira-Noriega, A., Munguia-Vega, A., et al. (2022). Integrating Earth-life systems: a geogenomic approach. *Trends in Ecology & Evolution*, *37*(4), 371–384.

Fabreti, L. G., & Höhna, S. (2021). Convergence Assessment for Bayesian Phylogenetic Analysis using MCMC simulation. *Methods in Ecology and Evolution*.

Gower, G., Ragsdale, A. P., Bisschop, G., Gutenkunst, R. N., Hartfield, M., Noskova, E., Schiffels, S., et al. (2022). Demes: a standard format for demographic models. *Genetics*, *222*(3).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

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.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., et al. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, *65*(4), 726–736.

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.

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

Landis, M. J., Eaton, D. A. R., Clement, W. L., Park, B., Spriggs, E. L., Sweeney, P. W., Edwards, E. J., et al. (2021). Joint Phylogenetic Estimation of Geographic Movements and Biome Shifts during the Global Diversification of Viburnum. *Systematic Biology*, *70*(1), 67–85.

Landis, M. J., Freyman, W. A., & Baldwin, B. G. (2018). Retracing the Hawaiian silversword radiation despite phylogenetic, biogeographic, and paleogeographic uncertainty. *Evolution*, *72*(11), 2343–2359.

Liu, S.-Y., Yang, Y.-Y., Tian, Q., Yang, Z.-Y., Li, S.-F., Valdes, P. J., Farnsworth, A., et al. (2025). An integrative framework reveals widespread gene flow during the early radiation of oaks and relatives in Quercoideae (Fagaceae). *Journal of Integrative Plant Biology*, *67*(4), 1119–1141.

Magalhaes, I. L. F., Santos, A. J., & Ramírez, M. J. (2021). Incorporating Topological and Age Uncertainty into Event-Based Biogeography of Sand Spiders Supports Paleo-Islands in Galapagos and Ancient Connections among Neotropical Dry Forests. *Diversity*, *13*(9), 418.

Matzke, N. J. (2014). Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. *Systematic Biology*, *63*(6), 951–970.

Ogden, T. H., & Rosenberg, M. S. (2006). Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology*, *55*(2), 314–328.

Parham, J. F., Donoghue, P. C. J., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., Inoue, J. G., et al. (2012). Best practices for justifying fossil calibrations. *Systematic Biology*, *61*(2), 346–359.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904.

Ree, R. H., & Sanmartín, I. (2018). Conceptual and statistical problems with the DEC+J model of founder-event speciation and its comparison with DEC via model selection. *Journal of biogeography*, *45*(4), 741–749.

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J., & Yang, Z. (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings. Biological Sciences / the Royal Society*, *279*(1742), 3491–3500.

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a review and a practical guide. *Molecular Ecology*, *25*(9), 1911–1924.

Roure, B., Baurain, D., & Philippe, H. (2013). Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Molecular Biology and Evolution*, *30*(1), 197– 214.

Salomé, P. A., Bomblies, K., Fitz, J., Laitinen, R. A. E., Warthmann, N., Yant, L., & Weigel, D. (2012). The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity*, *108*(4), 447–455.

Stull, G. W., Pham, K. K., Soltis, P. S., & Soltis, D. E. (2023). Deep reticulation: The long legacy of hybridization in vascular plant evolution. *The Plant Journal*, *114*(4), 743–766.

Tricou, T., Tannier, E., & de Vienne, D. M. (2022). Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*, *71*(5), 1147–1158.

White, L. T., Gibson, G. M., & Lister, G. S. (2013). A reassessment of paleogeographic reconstructions of eastern Gondwana: Bringing geology back into the equation. *Gondwana Research*, *24*(3–4), 984–998.

Zheng, H. (2015). Birth of the Yangtze River: age and tectonic-geomorphic implications. *National science review*, *2*(4), 438–453.

Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, *51*(4), 588–598.

# Figures

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

**Figure 2:** The methodological approach of geogenomics that allows reciprocal influence of biological and geological data to test hypotheses.

**Figure 3:** Maximum clade credibility trees produced from ten different models with varying geographic historical knowledge, prior distributions, and demographic models.

# Supplemental Information

**Table S1**: The absolute ages predicted for node 1 (taxa C and stem group AB), node 2 (taxa A and B), and the distribution of age ranges for the demographically manipulated node 2 (taxa A and B)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Node 1 Absolute Age** | **Node 2 Absolute Age** | **Node 2 Age Distribution** |
| Informed uniform, no gene flow | 3.6051 | 2.1222 | 3.3801, 0.9284 |
| Informed normal, no gene flow | 3.6012 | 2.1295 | 3.2818, 1.0217 |
| Unknown, no gene flow | 3.5304 | 2.0042 | 3.5133, 0.5983 |
| Incorrect uniform, no gene flow | 3.633 | 2.6246 | 3.5475, 1.7175 |
| Incorrect normal, no gene flow | 3.6297 | 2.6844 | 3.5546, 1.9654 |
| Informed uniform, gene flow | 3.5067 | 1.3893 | 2.9633, 0.2038 |
| Informed normal, gene flow | 3.5074 | 1.403 | 2.9705, 0.2054 |
| Unknown, gene flow | 3.5198 | 1.4953 | 3.0648, 0.2794 |
| Incorrect uniform, gene flow | 3.5086 | 1.4076 | 2.9846, 0.2103 |
| Incorrect normal, gene flow | 3.5084 | 1.4034 | 2.9679, 0.2042 |

**Figure S1:** Prior distributions for 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.**Figures S2-S11:** Maximum clade credibility trees constructed for the ten biogeographic models tested in this study.

**Figures S2-S11:** Maximum clade credibility trees produced by each of the 10 models.