Solís-Lemus, C., & Tiley, G. (2025). Seeing the Network for the Trees: Methodological and Empirical Advances in Reticulate Evolution. *Bulletin of the Society of Systematic Biologists*. https://doi.org/10.18061/bssb.v4i1.10445

Reviews

Seeing the Network for the Trees: Methodological and Empirical Advances in Reticulate Evolution

Claudia Solís-Lemus[1,2], George Tiley[3]

[1] Department of Plant Pathology, University of Wisconsin–Madison, [2] Wisconsin Institute for Discovery, University of Wisconsin-Madison, [3] Southern Genetics, LLC

Keywords: Species networks

https://doi.org/10.18061/bssb.v4i1.10445

Bulletin of the Society of Systematic Biologists

Abstract

This special collection includes topics related to the development of novel methods for reconstructing phylogenetic networks from different mathematical, statistical, and computational approaches that highlight the challenges of network reconstruction and the needs of contemporary genomic data. In addition, the collection broadcasts diverse applications of phylogenetic networks on a wide variety of organisms across the Tree of Life.

1 Introduction

Biodiversity on Earth has been shaped by multiple biological processes. Some of these processes, like speciation, are appropriately modeled by a tree structure, while others like gene flow, hybridization or introgression are better depicted using a network. Our ability to detect reticulate evolutionary processes within the Tree of Life has been propelled by our improved technological power to collect data as well as by the development of new statistical methods and mathematical models that appropriately account for gene flow. However, our ability to infer phylogenetic networks lacks far behind our capabilities to infer trees. While we can easily infer trees of thousands of taxa, we can rarely estimate networks with more than a hundred taxa.

The inference of reticulation events in the Tree of Life is challenging. In particular, the complexity of the statistical models increases dramatically when we simultaneously model reticulate processes along with tree processes like incomplete lineage sorting. Furthermore, reticulation events themselves can be complex. For example, a hybrid speciation event that gives rise to a new species that further hybridizes with a sister lineage results in what is mathematically denoted a level-2 phylogenetic network (Huson et al., 2010). Level-2 (or the more general level-k network for $k > 1$) networks have not been sufficiently studied despite representing a biological scenario that is not uncommon. Last, data heterogeneity, sampling strategies, noise, and missingness patterns can diminish the signal of gene flow, or eliminate it entirely.

This special edition presents a collection of 1) empirical studies that illustrate the power of phylogenetic networks to address important evolutionary biological and systematic questions (Fauskee et al., 2024; Morales-Briones & Kadereit, 2023; Olave et al., 2023); 2) novel methodological advances to infer or use phylogenetic networks from different data datatypes (Barton et al., 2022; Teo et al., 2023), and 3) rigorous evaluations of existing tree and network methods under different levels of complexity of gene flow events (Bjorner et al., 2022; Cao et al., 2022; Hibbins & Hahn, 2022; Justison & Heath, 2024). We hope this special edition will advance network thinking among evolutionary biology research as the field reconciles the rampant evidence of gene flow within the Tree of Life with our methodological ability to detect those reticulate patterns.

2 Advancing Research in Reticulate Evolution

2.1 Evolutionary insights that we can learn by using network methods

Hybrid origins of photosynthesis mechanism transitions

The evolution of photosynthesis mechanisms such as C_3 and C_4 have long been of interest to the plant science community, especially considering the association between photosynthesis and climate or ecosystem change (Ehleringer et al., 1991). Morales-Briones and Kadereit (2023) used the model system *Flaveria* (Asteraceae), which included C_3, C_4, and intermediate species to propose a novel mechanism where an intermediary lineage arose through hybridization of C_3 ancestors and a C_4 lineage arose through hybridization of intermediary ancestors. Findings were bolstered by consistent results between hy-

bridization tests (Kubatko & Chifman, 2019) and phylogenetic networks (Yu & Nakhleh, 2015) from a rigorously processed transcirptome dataset that could serve as a blueprint for future studies built upon orthogroups. The results of Morales-Briones and Kadereit (2023) should prompt a fresh perspective and new analysis on other groups with photosynthesis transitions too.

Explaining contentious phylogenetic relationships

Madagascar's mouse lemurs (*Microcebus*, Cheirogaleidae) include at least 19 named species that have diversified across divergent ecological niches in about 1.5 million years (Van Elst et al., 2024), but the relationships among well-supported clades within the group has remained unlclear given conflict among previous studies. Fauskee et al. (2024) showed high gene tree discordance underlying the focal node, and that this discordance was likely due to introgression between ancestral lineages. The authors used hybridization tests and network inference, and included hypothesis testing via marginal likelihoods to show that introgression between the focal lineages offered the best explanation for their genomic data. The results of Fauskee et al. (2024) offer new perspectives into the phylogeographic history of mouse lemurs while demonstrating a battery of statistical approaches for differentiating signals of introgression from incomplete lineage sorting or other sources of gene tree discordance.

Understanding speciation in the context of historical gene flow

Characterizing the molecular, morphological, and ecological differences underlying species (i.e. integrative species delimitation) is crucial for reconstructing speciation mechanisms. Olave et al. (2023) demonstrated how detecting historical gene flow between the ancestors of extant species and placing reticulate evolution in the context of paleoclimatic niches can improve our understanding of how species arise. Here, the authors used *Barupis* (Carabidae) distributed across the Andes and RADseq data to combine molecular and morphometric species delimitation techniques. Notably, SNP-based methods were used to estimate a species network (Olave & Meyer, 2020) and a model of trait evolution accounting for historical gene flow (Bastide et al., 2018) supported a scenario of transgressive evolution of a potentially adaptive trait. Olave et al. (2023) provides a roadmap for integrative species delimitation studies that need to consider gene flow among ancestral species, regardless of choice in molecular data.

2.2 Current methodological limits to detect reticulate evolutionary histories.

Introgression misleads inference of the history of speciation

Genomes have been shaped through time on signals from a mosaic of evolutionary and biological processes, and thus, thorough investigation of the limits of existing methods to distinguish reticulate evolution from other biological processes is needed. Hibbins and Hahn (2022) show that even low amounts of introgression can mislead species tree estimation when the rate of incomplete lineage sorting is high. Their work studies species tree inference from gene trees or biallelic sites, both options equally misguided by the data supporting the history of introgression rather than that of the true species tree (history of speciation). However, the signal for the correct speciation history is indeed in the data, as evidenced by the accurate prediction by machine-learning methods. Namely, the authors trained a supervised learning method on simulated gene tree datasets to perform binary classification of the history of speciation on 21 features of the gene trees. They found that gene tree features such as variance in coalescence time, node distances and gene tree frequencies have strong prediction accuracy reaching 76.3% for naïve Bayes and 93.3% for random forest. Their work illustrates the applicability of machine-learning to classify speciation histories on real data. However, such classifiers need to be trained on large simulated datasets which could be prohibited for scenarios with many taxa.

Hybridization events become undetectable when ancient, complex or involving ghost lineages

Bjorner et al. (2022) perform a thorough simulation study under different types of hybridization scenarios and test the performance of widely used hybrid detection methods such as MSCquartets (Mitchell et al., 2019), TICR (Stenz et al., 2015), HyDe (Kubatko & Chifman, 2019), Patterson's D-Statistic (Patterson et al., 2012) (also known as ABBA-BABA test). By eliminating other sources of noise such as constant rates of incomplete lineage sorting, constant population sizes and low gene tree estimation error, the authors are able to identify the effect of number and depth of reticulations and the mixing parameter (inheritance probability) on the accuracy of hybrid detection. They found that all methods have similar good performance (high precision and low false positive/negative rates) on single shallow hybridizations involving few taxa. The ABBA-BABA test displayed the highest false positive rates among all methods, especially in cases involving more than one hybridization event. As more hybridizations are added, all methods have higher false negative rate which suggests that complex hybridization scenarios weaken the reticulate signal rather than create discordant signal (which would be evidenced by an increased false positive rate). Last, while HyDe is the only method that can identify which taxon is the hybrid taxon among the taxa involved in the hybridization event, it cannot perform well when the parents of hybridization are unsampled or extinct (ghost lineages) which confirms what other studies found (Pang & Zhang, 2022; Tricou et al., 2022).

Model misspecification weakens the performance of phylogenetic network inference methods

Model misspecification has many shapes. Cao et al. (2022) focus on gene tree estimation error (GTEE) and the assumption of a single substitution rate for all genomic loci. They show that GTEE negatively impacts the performance of test statistics of "treeness" such as large false positives when data was simulated under the true multispecies coalescent (MSC) model. These methods, however, were accurate to determine that the MSC was unfit when data was generated under the multispecies network coalescent model (MSNC). Last, network inference is worsened under per-locus rate heterogeneity as these methods interpret any signal that deviates from the MSC as evidence for reticulation. Full Bayesian inference methods such as PhyloNet (Wen & Nakhleh, 2018) and BEAST2 (Drummond & Rambaut, 2007) that account for substitution rate heterogeneity have improved inference performance.

Distribution of network classes under a birth-death-hybridization process

Distributions of phylogenies that arise under a model of species birth and death (Kendall, 1948) have been well studied and form the basis of contemporary analyses of diversification rates. Our understanding of networks expected under an analogous birth-death-hybridization process (e.g. (Zhang et al., 2018)) is still in early stages. Justison and Heath (2024) helps to bridge this gap with extensive simulations that provide some of the first expectations for different types of network classes across parameter space of the birth-death-hybridization process. The results of Justison and Heath (2024) is not only of pressing interest for method developers that have to make assumptions about the network class for searches, but also informs empiricists on how many types of networks and hybridization events may not be detectable with existing methods.

2.3 Novel theoretical contributions in phylogenetic networks inference

Statistical learning with phylogenetic networks invariants

It is known that the inference of phylogenetic networks from genetic sequences is computationally expensive. Barton et al. (2022) introduce a new model-based approach to infer 4-leaf level-1 phylogenetic networks (quarnets) from algebraic invariants that site pattern probability distributions from a Jukes-Cantor phylogenetic network model must satisfy. Their work can be extended to more than four taxa by puzzling the quarnets into larger networks. Their method (QNR-SVM) takes as input aligned DNA sequences for a set of four taxa and uses support vector classifiers to classify it as belonging to one of the 24 quarnet models. The method is highly accurate to identify hybridization cycles of 4 nodes, but fails to identify cycles with 3 nodes which was expected given the lack of identifiability of such cycles (Allman et al., 2024; Gross et al., 2021; Solis-Lemus & Ané, 2016). The method is implemented in R and publicly available.

Continuous trait evolution model simultaneously accounting for within-species variation and reticulation

Teo et al. (2023) introduce the first phylogenetic linear model in which the phylogeny can be a network and that accounts for within-species variation in the continuous response trait. Within-species trait variation has been commonly denoted "measurement error" which can be misleading as this variation can be due to genetic differences, plasticity or environmental variation within species. Teo's method has three main contributions compared to other trait models, in addition to using a network as backbone phylogeny for the covariance matrix. First, the model is the first to allow one or more species to have a single observation thanks to the assumption of equal within-species variance. Second, the method jointly estimates the within-species variance with other parameters as opposed assume it to be perfectly known. Third, their implementation uses restricted maximum likelihood (REML) instead of maximum likelihood (ML) which is known to correct the underestimation of variance components. Their implementation is publicly available in the PhyloNetworks Julia package (Solis-Lemus et al., 2017).

3 Challenges and opportunities ahead

The collection highlights advances in the development of network methods as well as contemporary challenges. While multiple hybrid detection or network estimation methods have been developed in a relatively short amount of time, the limitations of existing methods are still being characterized. This is increasingly important as more investigations move from detecting the presence of introgression among species to understanding how introgression can play a role in speciation and the evolution of interesting traits (Morales-Briones & Kadereit, 2023; Olave et al., 2023). While networks can go a long way to resolving uncertainty in phylogenetic relationships that have remained uncertain even with genomic data (Fauskee et al., 2024), it will become difficult to recover networks accurately for deeper relationships (Bjorner et al., 2022). It is increasingly evident that we need to address the effects of ancient introgression when reconstructing the evolution of species (Cao et al., 2022), but it is possible for some relationships to be beyond our current toolkit (Justison & Heath, 2024).

As species networks become as routine for evolutionary biologists as molecular phylogenies, new questions will certainly arise. What will a network mean to someone using phylogenetic systematics to resolve the taxonomy of a difficult group? How should lineages of potential hybrid origin be considered regarding large-scale biodiversity studies using species richness or phylogenetic diversity? Do the evolutionary origins of some traits as well as their genetic basis need to be revisited? These issues and others will certainly

arise as the methods experience continuous improvement with well-characterized statistical properties and user-friendly implementations. Species networks are becoming more accessible to empiricists and we suspect will present rewarding new avenues of statistical and computational research as needs for scalable and accurate methods under complex scenarios remain.

Acknowledgments

This work was partially funded by the National Science Foundation [DEB-2144367 to CSL]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 101026923, awarded to GPT.

Submitted: December 04, 2024 EDT. Accepted: January 13, 2025 EDT. Published: March 31, 2025 EDT.

References

Allman, E. S., Baños, H., Garrote-Lopez, M., & Rhodes, J. A. (2024). Identifiability of level-1 species networks from gene tree quartets. *arXiv Preprint arXiv:2401.06290*. https://doi.org/10.1007/s11538-024-01339-4

Barton, T., Gross, E., Long, C., & Rusinko, J. (2022). Statistical learning with phylogenetic network invariants. *arXiv Preprint arXiv:2211.11919*.

Bastide, P., Solis-Lemus, C., Kriebel, R., William Sparks, K., & Ané, C. (2018). Phylogenetic comparative methods on phylogenetic networks with reticulations. *Systematic Biology*, *67*(5), 800–820. https://doi.org/10.1093/sysbio/syy033

Bjorner, M., Molloy, E. K., Dewey, C. N., & Solis-Lemus, C. (2022). Detectability of varied hybridization scenarios using genome-scale hybrid detection methods. *arXiv Preprint arXiv:2211.00712*.

Cao, Z., Li, M., Ogilvie, H. A., & Nakhleh, L. (2022). The impact of model misspecification on phylogenetic network inference. *bioRxiv*, 2022–10. https://doi.org/10.1101/2022.10.24.513600

Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, *7*, 1–8. https://doi.org/10.1186/1471-2148-7-214

Ehleringer, J. R., Sage, R. F., Flanagan, L. B., & Pearcy, R. W. (1991). Climate change and the evolution of C4 photosynthesis. *Trends in Ecology & Evolution*, *6*(3), 95–99. https://doi.org/10.1016/0169-5347(91)90183-X

Fauskee, B., Crowl, A., Piatkowski, B., Yoder, A., & Tiley, G. (2024). Ancient introgression in mouse lemurs (microcebus: Cheirogaleidae) explains 20 years of phylogenetic uncertainty. *Bulletin of the Society of Systematic Biologists*, *3*(1). https://doi.org/10.18061/bssb.v3i1.9319

Gross, E., Iersel, L. van, Janssen, R., Jones, M., Long, C., & Murakami, Y. (2021). Distinguishing level-1 phylogenetic networks on the basis of data generated by markov processes. *Journal of Mathematical Biology*, *83*, 1–24. https://doi.org/10.1007/s00285-021-01653-8

Hibbins, M. S., & Hahn, M. W. (2022). Distinguishing between histories of speciation and introgression using genomic data. *BioRxiv*, 2022–09. https://doi.org/10.1101/2022.09.07.506990

Huson, D. H., Rupp, R., & Scornavacca, C. (2010). *Phylogenetic networks: Concepts, algorithms and applications* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511974076

Justison, J. A., & Heath, T. A. (2024). Exploring the distribution of phylogenetic networks generated under a birth-death-hybridization process. *Bulletin of the Society of Systematic Biologists*, *2*(3), 1–22. https://doi.org/10.18061/bssb.v2i3.9285

Kendall, D. G. (1948). On the generalized "birth-and-death" process. *The Annals of Mathematical Statistics*, *19*(1), 1–15. https://doi.org/10.1214/aoms/1177730285

Kubatko, L. S., & Chifman, J. (2019). An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evolutionary Biology*, *19*(1), 112. https://doi.org/10.1186/s12862-019-1439-7

Mitchell, J. D., Allman, E. S., & Rhodes, J. A. (2019). Hypothesis testing near singularities and boundaries. *Electronic Journal of Statistics*, *13*(1), 2150–2193. https://doi.org/10.1214/19-EJS1576

Morales-Briones, D. F., & Kadereit, G. (2023). Exploring the possible role of hybridization in the evolution of photosynthetic pathways in flaveria (asteraceae), the prime model of C4 photosynthesis evolution. *Bulletin of the Society of Systematic Biologists*, *2*, 1–16. https://doi.org/10.18061/bssb.v2i3.8992

Olave, M., Griotti, M., Carrara, R., Franchini, P., Meyer, A., & Roig-Juñent, S. A. (2023). Historical climate change dynamics facilitated speciation and hybridization between highland and lowland species of baripus ground beetles from patagonia. *Bulletin of the Society of Systematic Biologists*, *2*(3), 1–16. https://doi.org/10.18061/bssb.v2i3.9263

Olave, M., & Meyer, A. (2020). Implementing large genomic single nucleotide polymorphism data sets in phylogenetic network reconstructions: A case study of particularly rapid radiations of cichlid fish. *Systematic Biology*, *69*(5), 848–862. https://doi.org/10.1093/sysbio/syaa005

Pang, X.-X., & Zhang, D.-Y. (2022). Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. *Systematic Biology*. https://doi.org/10.1093/sysbio/syac047

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Solis-Lemus, C., & Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.*, *12*(3), e1005896. https://doi.org/10.1371/journal.pgen.1005896

Solis-Lemus, C., Bastide, P., & Ané, C. (2017). PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.*, *34*(12), 3292–3298. https://doi.org/10.1093/molbev/msx235

Stenz, N. W. M., Larget, B., Baum, D. A., & Ané, C. (2015). Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species *Arabidopsis thaliana* (L.) Heynh. *Systematic Biology*, *64*(5), 809–823. https://doi.org/10.1093/sysbio/syv039

Teo, B., Rose, J., Bastide, P., & Ané, C. (2023). Accounting for within-species variation in continuous trait evolution on a phylogenetic network. *Bulletin of the Society of Systematic Biologists*, *2*(3), 1–29. https://doi.org/10.18061/bssb.v2i3.8977

Tricou, T., Tannier, E., & Vienne, D. M. de. (2022). Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*. https://doi.org/10.1093/sysbio/syac011

Van Elst, T., Sgarlata, G. M., Schüßler, D., Tiley, G. P., Poelstra, J. W., Scheumann, M., Blanco, M. B., Aleixo-Pais, I. G., Rina Evasoa, M., Ganzhorn, J. U., & others. (2024). Integrative taxonomy clarifies the evolution of a cryptic primate clade. *Nature Ecology & Evolution*, 1–16. https://doi.org/10.1038/s41559-024-02547-w

Wen, D., & Nakhleh, L. (2018). Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Systematic Biology*, *67*(3), 439–457. https://doi.org/10.1093/sysbio/syx085

Yu, Y., & Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, *16*, 1–10. https://doi.org/10.1186/1471-2164-16-S10-S10

Zhang, C., Ogilvie, H. A., Drummond, A. J., & Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, *35*(2), 504–517. https://doi.org/10.1093/molbev/msx307