Species Delimitation Annotated Bibliography

Phyloseminar 2024

David Adelhelm

Wayne P. Maddison, Gene Trees in Species Trees, Systematic Biology, Volume 46, Issue 3, September 1997, Pages 523–536, https://doi.org/10.1093/sysbio/46.3.523

One of the earlier papers discussing the topic of gene tree/species tree discordance. Maddison explores several of the common causes of this discordance, including horizontal transfer, hybridization, lineage sorting, and gene duplication, as well as how to diagnose some of these causes. Next, the author discusses how these processes could be evaluated under both parsiomony and maximum likelihood optimality criteria to yield a species tree from many gene trees. Lastly, Maddison addresses the more philosophical question of "what is a species phylogeny?". Is it accurate to describe different gene trees as "right" or "wrong" given their agreement with an assumed species tree? What can our phylogeny tell us about interbreeding amongst and within different tips?

Kevin De Queiroz, Species Concepts and Species Delimitation, Systematic Biology, Volume 56, Issue 6, December 2007, Pages 879–886, https://doi.org/10.1080/10635150701701083

A somewhat famous paper in the topic of species delimitation. In this Article, De Queiroz highlights the many different ways in which a researcher may conceptualize a species and how that can conflict with species delimitation. The author starts by outlining several popular species concepts in biology and what properties they use to assess a species' status. Following this, an attempt at reconciling these many concepts is made by identifying commonalities amongst them. Namely, all the them categorize a species as a group evolving separately from it's relatives. De Queiroz proposes that different species concepts may be addressing different stages along a continuum that is speciation, with some claiming 'species' earlier in the continuum than others. In the remainder of the paper, De Queiroz addresses the impacts of a unified species concept on the study of species delimitation. First, it helps to clarify differences between conceptualization (what is a species) versus delimitation (how do we infer species boundaries and number). Under a unified species concept, the different properties of a proposed species may instead be used as different pieces of evidence in the methodology of species delimitation. Furthermore, none of these properties guarantees species identity and instead may be one possible piece of evidence for a larger study, and the absence of any property also does not completely exclude species identity.

James H. Degnan, Noah A. Rosenberg, Discordance of Species Trees with Their Most Likely Gene Trees, PLOS Genetics, Volume 2, Issue 5, May 2006, https://doi.org/10.1371/journal.pgen.0020068

A methods paper further exploring the topic of gene tree/species tree discordance, primarily the existence of an "anomaly zone" in gene tree inference where data addition can reliably lead to incorrect estimates of species trees. When inferring a species tree from gene trees, it has been common to estimate a species tree as the most commonly occurring topology amongst many gene trees. However, for certain topologies and regions in branch length space, deep coalescence can lead to discordant gene trees being inferred more often than those in agreement with the species tree. In such a scenario, we experience the phenomenon of statistical inconsistency, a situation where the addition of data (more loci) actually increases the odds of inferring a incorrect species tree. As a proposed solution, the authors explain that while the addition of more loci may exacerbate the problem when anomalous gene trees (AGT's) are present, increasing sample sizes may help to lessen the occurrence of AGT's. Additionally, since AGT's only begin to occur in a trees with 4 or more tips, it is possible to estimate a species tree by inferring all possible 3-tip species trees and then create a consensus species tree from these.

Joseph Heled, Alexei J. Drummond, Bayesian Inference of Species Trees from Multilocus Data, Molecular Biology and Evolution, Volume 27, Issue 3, March 2010, Pages 570–580, https://doi.org/10.1093/molbev/msp274

Another early methods paper demonstrating the application of Bayesian Inference to Coalescent Theory to coestimate several gene trees within a species tree, this workflow is now forms the basis for the popular divergence time estimation platform *BEAST (star-beast). By creating several simulated datasets, the author demonstrated that the *BEAST workflow performs moderately well even at low levels of sampling (4 individuals per species), and rapidly increases in accuracy with the addition of loci to the study. Additionally, error may be lessoned by increasing sequence length.

Ziheng Yang, Bruce Rannala, Bayesian Species Delimitation Using Multilocus Sequence Data, PNAS, Volume 107, Issue 20, May 2010, Pages 9264-9269, https://doi-org.dist.lib.usu.edu/10.1073/pnas.0913022107

The BPP publication, another bayesian method in species delimitation where the authors demonstrate a workflow capable of assessing posterior probability of species assignments given a guide topology and gene trees as input. A guide topology is required in this analysis because it allows the search algorithm to drastically reduce the space of parameters that must be integrated over, thus creating a more efficient computation.

Devon A DeRaad, John E McCormack, Nancy Chen, A Townsend Peterson, Robert G Moyle, Combining Species Delimitation, Species Trees, and Tests for Gene Flow Clarifies Complex

Speciation in Scrub-Jays, Systematic Biology, Volume 71, Issue 6, November 2022, Pages 1453–1470, https://doi.org/10.1093/sysbio/syac034

A more recent exploration of species delimitation given advancements in methodologies of the late 2010s/early 2020s. While advancements in genomic resources have allowed researchers to generate massive datasets for several nonmodel systems, it has increasingly become apparent that speciation is rarely 'neat'. "Complex Speciation", or the situation in which species continue to interbreed following speciation, presents a difficult challenge for those interested in species delimitation as it undermines many of the common assumptions of our early models. However, this challenge has inspired a great diversity of new methods in the past decade. In this study, the authors demonstrate the efficacy of several machine learning methodologies in species discovery of the American Scrub-Jays. By combining all of the methods, in addition to several analyses for population demography and gene flow, the authors are able to find strong signal for novel species identity within the scrub-jays despite their history of complex speciation

José-Rubén Montes, Pablo Peláez, Alejandra Moreno-Letelier, David S. Gernandt, Coalescent-based Species Delimitation in North American Pinyon Pines Using Low-copy Nuclear Genes and Plastomes, American Journal of Botany, Volume 109, Issue 5, May 2022, Pages 706-726, https://doi-org.dist.lib.usu.edu/10.1002/ajb2.1847

Similar to the previous paper on scrub-jays, this paper serves to demonstrate recent advancements in species delimitation methodology by applying 3 different methods (Generalized Mixed Yule Coalescent, Poisson Tree Process, and Trinomial Distribution of Triplets) to the North American Pinyon Pines, a group characterized by rampant ILS. Although the three methods varied significantly in their species identification, the authors found that two of the methods, GMYC and PTP, tended to oversplit species and often failed to identify species with high morphological dissimilarity. The authors believe these results to be unreliable and likely a result of the high variation in divergence times across the group. However, the Triplets method seemed to generate much more reliable results that were in agreement with existing delimitations built from different sources of data.

Shahan Derkarabetian, Stephanie Castillo, Peter K. Koo, Sergey Ovchinnikov, Marshal Hedin, A Demonstration of Unsupervided Machine Learning in Species Delimitation, Molecular Phylogenetics and Evolution, Volume 139, 2019, https://doi.org/10.1016/j.ympev.2019.106562

A really cool methods paper to come out of recent advancements in machine learning. A common first step in species delimitation research is the Species Discovery phase, where samples with no prior species assignment are given to an algorithm which must then attempt to cluster or group said samples into probable species groups. Fortunately, this is very conducive to unsupervised machine learning, where a machine is fed unlabeled data and must use the inherent data structure to generate groupings. Using three different forms on

unsupervised machine learning, the authors demonstrate that UML can correctly infer species boundaries without oversplitting taxa due to population structuring, as several model-based methods are liable to do.

Sonja Bamberger, Jie Xu, Bernhard Hausdorf, Evaluating Species Delimitation Methods in Radiations: The Land Snail Albinaria cretensis Complex on Crete, Systematic Biology, Volume 71, Issue 2, March 2022, Pages 439–460, https://doi.org/10.1093/sysbio/syab050

Another methods comparison test highlighting the difficulties of species delimitation (and coalescent methods) in systems of rapid radiations. Rapid radiations often yield gene trees with short and wide branches, creating a high probability of deep coalescence and gene tree discordance. These many discordant gene trees make species tree estimation exceedingly difficult by limiting the consensus available for the model to draw from. In this study, the authors sought to test recent advancements in species delimitation against a land snail radiation. For species discovery, the data were run through three different pipelines; the MSC (implemented through BPP), Gaussian clustering, and population structure through ADMIXTURE. These three analyses creating a somewhat goldilocks-like scenario: the MSC tended to oversplit species compared to existing delimitations (based on morphology), the gaussian clustering tended to overlump species, and ADMIXTURE seemed to agree most with existing delimitation (with moderate overlumping). Following this, the data were tested against three different species validation methods: BFD*, delimitR, and isolation by distance (IBD) testing. Both BFD* and delimitR were found to agree completely with species delimitations made by ADMIXTURE, and would also agree with moderate overlumping hypotheses (splitting species into subgroups). This seems to indicate that BFD* and delimitR cannot distinguish well between population and species boundaries. However, the IBD tests inferred that differentiation between several clusters inferred by ADMIXTURE can be explained by IBD and not speciation, cutting the total number of hypothesized species from 14 to 8. Given how much the IBD analyses seemed to agree with morphological classifications of the group, the authors recommend IBD for future validation studies.

Arley Camargo, Mariana Morando, Luciano J. Avila, Jack W. Sites, SPECIES DELIMITATION WITH ABC AND OTHER COALESCENT-BASED METHODS: A TEST OF ACCURACY WITH SIMULATIONS AND AN EMPIRICAL EXAMPLE WITH LIZARDS OF THE LIOLAEMUS DARWINII COMPLEX (SQUAMATA: LIOLAEMIDAE), Evolution, Volume 66, Issue 9, 1 September 2012, Pages 2834–2849, https://doi.org/10.1111/j.1558-5646.2012.01640.x

Referenced frequently in the How to Fail at Species Delimitation paper. This paper applies three different species delimitation methods (BPP, SpeDeSTEM, and ABC) to simulated data to determine their individual accuracies before testing them on a nonmonophyletic group of lizards. From their simulations, the authors found that in terms of accuracy (given their dataset) the ranking was BPP>ABC>SpeDeSTEM. Additionally, all three methods performed worse in the presence of gene flow, but ABC suffered the least. This highlights ABC as a good method to use when significant gene flow exists in your dataset.

Thomas C. Giarla, Jacob A. Esselstyn, The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews, Systematic Biology, Volume 64, Issue 5, September 2015, Pages 727–740, https://doi.org/10.1093/sysbio/syv029

Once again, a paper highlighting the difficulties of rapid radiation (and thus ILS) for species delimitation in empirical datasets. Using the genus *Crocidura* from the philippines as a case study for this methodological problem, the authors compile a genome-scale dataset including UCE's and mtDNA for use in three comparative methods; concatenation (via MrBayes), summary coalescent (MulRF and ASTRAL), and hierarchical coalescent (*BEAST). While the concatenation and summary coalescent methods seemed to agree well enough on topology compared to the very low support values of the *BEAST results, the simulation study implied that these may be false topologies (anomalous gene trees?). Part of this could be a result of model mis-specification or poor sampling (several species only had a single representative). Altogether, this casts doubt into any of their phylogenies and demonstrates the importance of including a large and representative sampling when tackling species delimitation.

Arong Luo, Cheng Ling, Simon Y W Ho, Chao-Dong Zhu, Comparison of Methods for Molecular Species Delimitation Across a Range of Speciation Scenarios, Systematic Biology, Volume 67, Issue 5, September 2018, Pages 830–846, https://doi.org/10.1093/sysbio/syy011

As the title suggests, a methods paper detailing the comparative efficacy of different models/methods in species delimitation. The authors use simulated data to test the accuracy of generalized mixed Yule-coalescent (GMYC), poisson tree process (PTP), and BPP. These simulated data were created under 5 unique scenarios with varying degrees of population size, crown age, sample size, gene flow, and speciation rate to get an idea of which factors have the greatest impact on model performance. The following tests and performance assessment found that BPP tends to perform very well under most scenarios, with a drop in performance under high rates of gene flow (BPP assumes no gene flow so this is somewhat expected). Additionally, for singe-locus datasets PTP and GMYC tended to perform similarly, but with with PTP being preferable for its ability to handle smaller species numbers and more specifically their scenario II (large populations, long divergence times). Further, the authors caution against GMYC and PTP more generally as they tend to assume equivalency between species and gene trees, which becomes problematic in the presence of discordance. The two singe-locus methods also struggled when faced with high N/t ratios, to the point where data addition failed to improve performance. Given these findings, the authors recommend BPP out of the three methods but warn that BPP's implicit assumptions regarding gene flow should be considering before proceeding.

Jeet Sukumaran, L. Lacey Knowles, Multispecies Coalescent Delimits Structure, Not Species, PNAS Biological Sciences, Volume 114, Issue 7, February 2017, Pages 1607-1612, https://doi-org.dist.lib.usu.edu/10.1073/pnas.1607921114

This paper warns users of the MSC against using a single line of evidence (genetic) as the only basis for species delimitation. As the title suggests, this study demonstrates using simulated

data that the MSC is often insufficient in distinguishing between population structuring and species boundaries, leading to overestimation of species in discovery-based methods. Some of the biggest takeaways from the results were that BPP tends to overestimation species number across most speciation rate paradigms, and that when speciation is exceedingly low, the delimitations made are “essentially random with respect to the actual number and specific boundaries of species, averaging 5–13 times more estimated species”. Additionally, BPP never underestimated species boundaries, only over. All of this is to say that population structuring and dynamics create genetic structuring that may not necessarily equate to species boundaries, but nevertheless that structuring is the main signal identified in MSC analyses, biasing our work towards overestimation.

Wayne P. Maddison, Jeannette Whitton, The Species as a Reproductive Community Emerging From the Past, Bulletin of the Society of Systematic Biologists, Volume 2, Issue 1, 2023, https://doi.org/10.18061/bssb.v2i1.9358

In a similar vein to De Queiroz 2007, this review seeks to address important considerations to the field of species delimitation. At the core of their thesis is the argument that a species should reflect the reproductive history of the lineage. By narrowing the argument to reproductive history, we can understand the community in retrospect. How do the traits we see presently explain mating behavior in the past? The authors address de Queiroz’s concept of a species as a “separately evolving metapopulation” by specifying that our concept needs temporal relevance, are we looking forwards, at the present, or backwards, when we assess this separate evolution? From this we gain the new “Retrospective Reproductive Community Concept”, the idea that species are cohesion-molded lineages of the past and that the traits seen today are shaped by processes of the past. The authors argue that this is a necessary revision to species concepts as it better addresses process as a key factor to the delineation of a species, and that process must be cohesive across the individuals. Furthermore, it connects our unit (the species) better to our method. We believe our individual gene histories (although occasionally discordant) share a linkage through the reproductive community, and we are implicitly assuming them when we concatenate our genes or apply the multispecies coalescent.