Effects of phylogenetic variation on prioritization of species for conservation

Kristina Wicke[1], Md Rejuan Haque[2], and Laura Kubatko[3,4,*]

[1] *Department of Mathematical Sciences, New Jersey Institute of Technology, Newark NJ, USA*
[2] *Division of Biostatistics, College of Public Health, The Ohio State University, Columbus OH, USA*
[3] *Department of Statistics, The Ohio State University, Columbus OH, USA*
[4] *Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus OH, USA*

Abstract

Conservation planning often involves difficult choices about which species to protect. One way to prioritize species for conservation is to consider their relative evolutionary isolation on a phylogeny. Several evolutionary isolation metrics have been discussed in the literature, all of which require a phylogenetic tree as input. The recent widespread recognition of the ubiquity of incongruence between gene trees and species trees thus leads to the natural question of which should be used as the phylogenetic basis for assessing conservation priorities. To address this, we explored the variability in species rankings based on gene trees as compared to the species tree for several empirical data sets. We used the Fair Proportion (FP) index, also known as the evolutionary distinctiveness score, as the measure of evolutionary isolation. The FP index apportions the total diversity of a phylogenetic tree among all leaves in such a way that each species receives a "fair proportion" of its ancestry. Our results indicate that prioritization rankings among species vary greatly depending on the underlying phylogeny, suggesting that the choice of phylogeny is crucial in assessing phylogenetic diversity in a conservation setting. Our aim is to raise awareness of this issue and inspire new research on which evolutionary information (species trees, gene trees, or a combination of both) should form the basis for conservation decisions.

Keywords: Fair proportion index, Gene tree-species tree discordance, multispecies coalescent, phylogenetic diversity

1 Introduction

The world is currently facing a major extinction crisis that will inevitably result in loss of biodiversity. Limited financial resources coupled with varying levels of recognition of the importance of conserving biodiversity necessitate the development of methods for prioritizing species for conservation. Quantitative approaches to assessing the value of individual species to overall biodiversity are thus a crucial component in developing a robust conservation strategy.

Various methods for quantifying biodiversity have been proposed, including those that consider the evolutionary heritage of a set of species as represented by the phylogenetic tree relating them. For example, the *phylogenetic diversity* (PD) measure introduced by [3] sums the branch lengths connecting a set of species, thus linking evolutionary history to feature diversity [3]. Preserving PD has become an integral part of conservation considerations (see, e.g., the "Phylogenetic Diversity Task Force" initiated by the IUCN [5]).

While PD captures the biodiversity of sets of species, evolutionary isolation metrics, also known as phylogenetic diversity indices, apportion the total diversity of a tree among its leaves and quantify the relative

*Corresponding author
Email address: kubatko.2@osu.edu

importance of species for overall biodiversity based on their placement in the tree. Various methods can be devised to distribute the total diversity of a tree across present-day species, and a variety of phylogenetic diversity indices for phylogenetic trees has been introduced in the literature (for an overview, see, e.g., [24, 17]). One of the most popular indices is the *Fair Proportion* (FP) index (also known as *evolutionary distinctiveness* (ED) score) introduced by [16, 4]. Its underlying idea is to assign each species a "fair proportion" of the total PD of a tree. More precisely, each species descended from a given branch in the phylogeny receives an equal proportion of that branch's length.

The FP index has been employed in the "EDGE of Existence project", a global conservation initiative focusing specifically on threatened species that represent a significant amount of unique evolutionary history ([4]; see also `https://www.edgeofexistence.org/`). This project aims at identifying species that are both **e**volutionary **d**istinct and **g**lobally **e**ndangered, where evolutionary distinctiveness is measured in terms of the FP index and the globally endangered score is based on the IUCN Red List Category. However, because the FP index heavily depends on the underlying phylogenetic tree and its branch lengths, different phylogenies for the same set of species may result in different prioritization orders.

Phylogenetic trees have traditionally been inferred from single genes and the resulting gene trees were assumed to be a valid estimate for the species tree, i.e., the "true" evolutionary history of the species under consideration. However, the advent of whole genome sequencing has resulted in increased appreciation of the differences between these gene trees and the species tree. Discordance among gene trees and between gene trees and the species tree is known to arise from numerous biological processes, including incomplete lineage sorting, lateral gene transfer, and gene duplication and loss (e.g., [8, 11, 12]). For the FP index, this discordance implies that different gene trees may result in different prioritization orders that conflict with one another as well as with the species tree.

Substantial variability in the ranking positions of species induced by the FP index for different gene tree and species tree estimates has recently been observed for a multilocus mammal data set [25]. In this note, we show that this is indeed a widespread phenomenon occurring across various groups of organisms by analyzing nine multilocus data sets from the literature and comparing the FP prioritization orders obtained on individual gene trees, on species trees, and by averaging over gene trees. We first describe the computation of the FP index and then discuss the results for these nine data sets. We conclude with a discussion of the issues that arise from these results and argue that this phenomenon deserves increased attention as genome-scale data become routinely available.

2 Materials and Methods

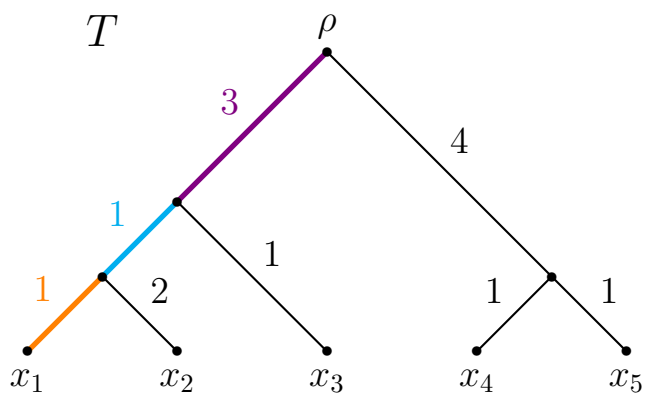
The FP index

Let T be a rooted phylogenetic tree with leaf set $X = \{x_1, \ldots, x_n\}$ and root ρ, where each edge e is assigned a non-negative length $l(e) \in \mathbb{R}_{\geq 0}$. Then, the FP index [16, 4] for leaf $x_i \in X$ is defined as $FP_T(x_i) = \sum_{e \in P(T;\rho,x_i)} \frac{l(e)}{n(e)}$, where $P(T;\rho,x_i)$ denotes the path in T from the root ρ to leaf x_i and $n(e)$ is the number of leaves descended from e. As an example, for tree T depicted in Fig. 1 and leaf x_1, we have $FP_T(x_1) = \frac{3}{3} + \frac{1}{2} + \frac{1}{1} = 2.5$. Analogously, $FP_T(x_2) = 3.5$, $FP_T(x_3) = 2$, $FP_T(x_4) = 3$, and $FP_T(x_5) = 3$. Note that the FP index induces a natural ordering of species, where species with the larger values of the FP index are given higher conservation priority.

Data collection

Nine multilocus data sets were curated from the literature. In some cases, we reduced the published data sets to a subset of species and/or genes due to large amounts of missing data or missing outgroups. Detailed information on both the data reduction process as well as the species and loci included in our analyses is available from the SI. The full data set was used when not otherwise noted below:

- Dolphin data set [9]: DNA sequence data from 24 genes for 47 aquatic mammals. We reduced the data set to 28 species and 22 genes.



$$FP_T(x_1) = \frac{3}{3} + \frac{1}{2} + \frac{1}{1} = 2.5$$

Figure 1: Rooted phylogenetic tree T with leaf set $X = \{x_1, \ldots, x_5\}$. As an example, for the FP index of x_1, we have $FP_T(x_1) = \frac{3}{3} + \frac{1}{2} + \frac{1}{1} = 2.5$.

- Fungi data set [19]: Amino acid sequence data, gene tree, and species tree estimates based on 683 genes for 25 individuals of bipolar budding yeasts and four outgroups.
- Mammal data set [20, 21]: DNA sequence data and gene tree estimates based on 447 genes for 33 species of mammals and four outgroups.
- Plant data set [19]: DNA sequence data, gene tree, and species tree estimates based on 363 genes for 48 Lamiaceae and four outgroup species. We reduced the data set to 318 genes for the 52 species.
- Primate data set [1]: DNA sequence data from 52 genes for four primate species.
- Rattlesnake data set [6]: DNA sequence data from 19 genes for 24 individuals of six subspecies of *Sistrurus* rattlesnakes and two outgroup species. We picked one sequence per subspecies and one outgroup sequence (i.e., seven sequences in total), and used 16 of the 19 loci.
- Rodent data set [19]: DNA sequence data, gene tree, and species tree estimates based on 794 genes for 37 rodent species. We reduced the data set to 761 genes for the 37 species.
- Snake data set [14]: DNA sequence data, gene tree, and species tree estimated based on 333 genes for 31 caenophidians and two outgroup species.
- Yeast data set [18]: DNA sequence data from 106 genes for eight yeast species.

Tree estimation and data analysis

Five of the nine data sets (fungi, mammal, plant, rodent, and snake) contained gene tree estimates, which we subsequently used in the analysis. For the remaining four data sets (dolphin, primate, rattlesnake, and yeast), we estimated gene trees under the GTR+Gamma model using RAxML version 8.2.12 [22].

Except for the fungi and snake data sets, a species tree estimate was obtained using SVDquartets [2] as implemented in the PAUP* package [23] (note that we re-estimated species trees for the rodent and plant data sets instead of using the published species trees since we reduced the corresponding data sets in the gene tree analysis). Finally, maximum likelihood branch lengths were computed for all species trees under

the GTR+Gamma model (for all but the fungi data set, for which we used the LG model for amino acid sequence evolution) with the molecular clock enforced using PAUP*.

FP indices on individual gene and species trees as well as average FP indices across gene trees were calculated and the resulting values were transformed into rankings (using the "1224" standard competition ranking for ties). Based on this, boxplots of the distribution of ranks across gene trees were generated for each species (an example is depicted in Fig. 2) and their interquartile ranges (IQRs) were computed. The mean, minimum, and maximum IQR across taxa, standardized by dividing by the number of taxa, was computed. Larger values indicate more variability in the distribution of ranks, and thus ranks that differ more across loci. Kendall's τ rank correlation was calculated between the rankings obtained from pairs of gene trees, and from individual gene trees and the species tree (Fig. 3), as well as from the average FP index and the species tree (Table 1). All statistical analyses were performed using the R Statistical Software [15].

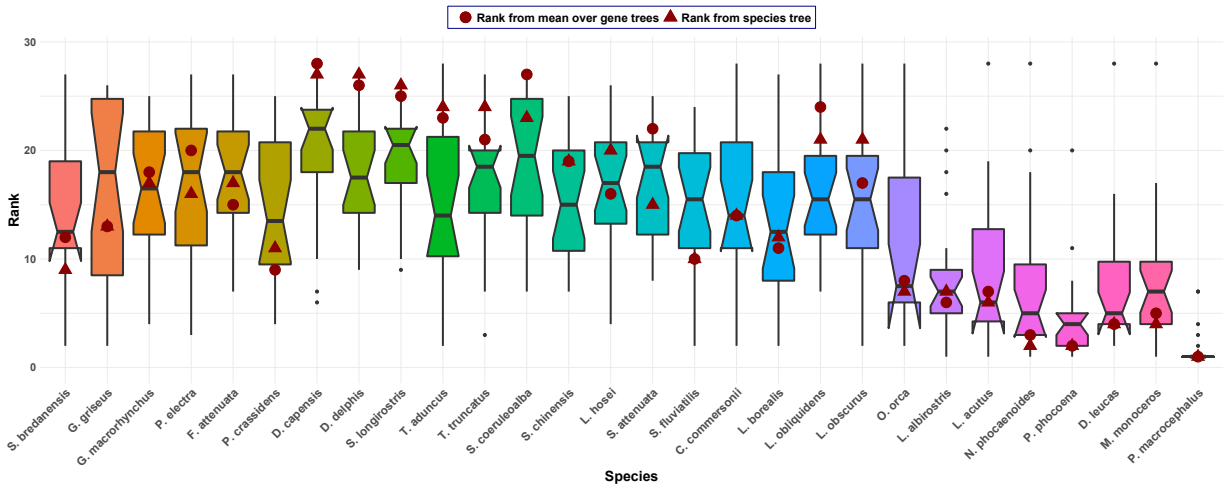


Figure 2: Boxplot of the ranks obtained from the FP index for the Dolphin data set consisting of 28 species and 22 genes. In addition, the ranks obtained from the average FP index across the 22 gene trees (dots) and the ranks on the species tree (triangles) are depicted. The boxplots show large variability in ranks across gene trees. Moreover, the ranks obtained from the average FP index across gene trees and the FP index on the species tree are sometimes ordered differently. For example, *F. attenuata* receives a higher rank than *P. electra* on the species tree, but a smaller rank based on the average FP index.

3 Results

We observed extensive variation in the rankings of taxa induced by the FP index across different gene trees for all data sets analyzed. As a representative example, Fig. 2 shows boxplots of the rankings for each of 28 marine mammal species for a data set with 22 genes [9]. Note in particular that some species (e.g., *G. griseus*) have ranks that span almost the entire range of possibilities, while relatively few have consistent ranks across all genes (as would be indicated by "narrow" boxplots in Fig. 2). The corresponding boxplots for the other data sets, which vary in the type of organism and the number of genes included, show similar behavior (see the SI).

To measure the extent of variation in the rankings of species across genes trees, the IQR of ranks assigned was computed for each species. The mean, minimum, and maximum of the IQRs across species (Table 1, columns 3-5) are relatively stable, with the mammal, plant, and yeast data sets showing the lowest variability in rankings and the primate data set the highest. The large maximum IQR for the dolphin and fungi data

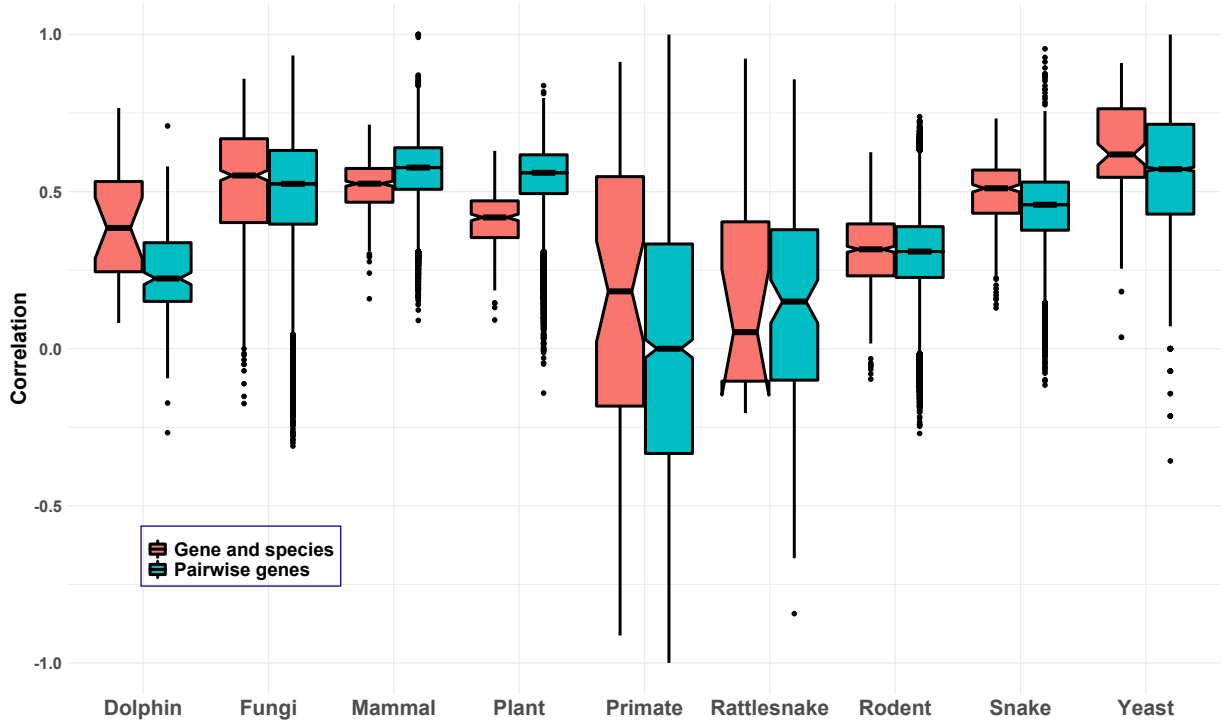


Figure 3: Boxplots of Kendall's τ rank correlation between rankings obtained from individual gene trees and the species tree (red) and rankings obtained from pairs of gene trees (blue) for all data sets considered. Note that for the plant data set, Kendall's τ rank correlation between the rankings obtained from pairs of gene trees was based on the ranks of all species present in both gene trees, respectively (due to missing data, not all gene trees contained the complete set of species).

sets suggests the presence of at least one species for which large variation in ranks across genes is observed.

Fig. 3 summarizes results across all data sets examined. The blue boxplots show the distribution of values of Kendall's τ rank correlation between rankings of species obtained from pairs of gene trees. While there tends to be a positive correlation overall, some boxes are closer to zero than to one, and some negative values (meaning that the corresponding rankings are negatively correlated; in other words, they tend to rank species in opposite order) are present. Thus, using different gene trees for the calculation of the FP index can result in very different prioritization orders.

A similar trend occurs for the correlation between rankings obtained from individual gene trees and the species tree. The red boxplots in Fig. 3 again show a tendency for positive correlations, but in several cases the correlation is closer to zero than to one, and for some data sets (e.g., the primate and rattlesnake data sets) negative values are again observed.

Table 1: Summaries of the species rankings for varying phylogenies. Column 2: Kendall's τ rank correlation between the ranking of species induced by the average FP index across gene trees and that induced by the FP index on the species tree. Columns 3-5: Mean, minimum, and maximum IQR of the ranks obtained in the gene tree analyses (standardized by dividing by the number of taxa).

Data set	Kendall's τ	mean IQR	min IQR	max IQR
Dolphin	0.8465	0.2851	0.0000	0.5804
Fungi	0.6651	0.2206	0.0690	0.5690
Mammal	0.6041	0.1651	0.0270	0.2703
Plant	0.4828	0.1713	0.0192	0.2308
Primate	0.5477	0.3750	0.2500	0.5000
Rattlesnake	0.1502	0.3214	0.1786	0.4643
Rodent	0.5920	0.3084	0.0270	0.5135
Snake	0.6298	0.2360	0.0000	0.3939
Yeast	0.7638	0.1875	0.1250	0.5000

Given the variability in ranks obtained from different gene trees, we finally considered how the FP index averaged across gene trees compared to that obtained from the species tree. For most data sets, we observe some changes in the orderings of ranks when using the average over gene trees vs. the species tree (see Fig. 2 for an example). However, as Table 1 shows, the rankings are all positively correlated and, with the exception of the plant and rattlesnake data sets, the correlation is larger than 0.5. Thus, the average FP index across loci and the FP index based on the species tree tend to result in similar prioritization orders of species in general.

4 Discussion

The FP index is a popular tool for prioritizing species for conservation that assesses their importance to overall biodiversity based on their placement in an underlying phylogenetic tree. As such, it directly depends on a pre-specified phylogeny and its branch lengths, and different phylogenies for the same set of species can lead to different prioritization orders. In analyzing species rankings obtained from the FP index for nine multilocus data sets for a broad range of organisms, we find that different loci/gene trees can induce very different prioritization orders of species (reflected by large variability in the rank a species is assigned across gene trees). However, with few exceptions, taking all loci into account by averaging the FP index across loci or by constructing a species phylogeny based on all loci results in similar prioritization orders.

While we have used the FP index to illustrate the effect of gene tree variation on prioritization, this phenomenon is not limited to the FP index. Because phylogenetic diversity indices rely on specification of an underlying phylogeny, all such indices have the potential for the particular phylogeny chosen to strongly influence the results. When data for multiple loci are available for the species of interest, the question as to how these data should be used to estimate the phylogeny arises. Given the role of evolutionary processes like

incomplete lineage sorting and hybridization in generating variation in the trees for individual loci throughout the genome, the choice of phylogeny to use for evaluating biodiversity is far from straightforward. Somewhat reassuringly, our study finds that two reasonable ways of accommodating across-locus variability – averaging the FP index across loci or computing the FP index from a species tree – lead to similar rankings. On the other hand, our results convincingly demonstrate that use of the phylogeny for a single gene can lead to prioritization orders that are not consistent with those that use information from the entire genome.

Many other questions in evolutionary biology require specification of a phylogeny as a precursor to downstream analyses, and all such settings will be faced with the challenge of determining what data and methods should be used to infer that history. For example, ancestral state reconstruction is crucially reliant on specification of the underlying phylogeny, as has been recently highlighted by studies such as [13]. As another example, analyses of the shifts in diversification rates over time (e.g., [7]) also require specification of a phylogeny prior to inference. Recent work has also shown that analysis of quantitative trait evolution can benefit from the incorporation of gene tree variation [10]. The analyses presented here thus serve to underscore the importance of carefully considering choice of phylogeny for settings such as these.

In addition, our results demonstrate a need for future research into the effect of phylogeny choice on downstream inference. While in many cases use of a species tree might provide a robust estimate of evolutionary relationships, an argument can be made that use of a species tree could obscure relationships of interest to a particular question if such relationships were strongly supported only by a subset of genes. Approaches that more explicitly accommodate variability in the underlying histories of the loci by computing single-locus trees may be promising in such cases, as they would allow an examination of the contribution of individual phylogenies to the overall results. However, this comes at the expense of increased computation, as such approaches would require separate estimation of phylogenies for all individual loci. Another possibility would be to adopt a Bayesian approach to formally integrate over the distribution of gene trees, again incurring a substantial computational cost.

Overall, we hope that our analyses of the FP index serve to both highlight the importance of the issue in the particular application of conservation prioritization as well as provide a call for future research on how to best infer evolutionary relationships for downstream analyses.

Supplementary information

Additional information and figures can be found in the Supplementary Information (SI) available at `https://github.com/lkubatko/FP-Index-Discordance`.

Declarations

- Funding: KW was supported by The Ohio State University's President's Postdoctoral Scholars Program.

- Competing interests: The authors have no relevant financial or non-financial interests to disclose.
- Availability of data and materials: Data from the analyses in this manuscript are available at `https://github.com/lkubatko/FP-Index-Discordance`.
- Authors' contributions: KW and LK designed the research and wrote the manuscript. MRH compiled the data. All authors contributed to the data analysis, and all read and approved the manuscript.

References

[1] F.-C. Chen and W.-H. Li. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, 68:444–456, 2001. doi: 10.1086/318206.

[2] J. Chifman and L. Kubatko. Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, Aug. 2014. doi: 10.1093/bioinformatics/btu530.

[3] D. P. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, 1992. doi: 10.1016/0006-3207(92)91201-3.

[4] N. J. Isaac, S. T. Turvey, B. Collen, C. Waterman, and J. E. Baillie. Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PLoS ONE*, 2(3):e296, Mar 2007. doi: 10.1371/journal.pone.0000296.

[5] IUCN. IUCN SSC phylogenetic diversity task force. `https://www.pdtf.org/`, 2019. Accessed: 11-05-2022.

[6] L. S. Kubatko, H. L. Gibbs, and E. W. Bloomquist. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in sistrurus rattlesnakes. *Systematic Biology*, 60(4):393–409, Mar. 2011. doi: 10.1093/sysbio/syr011.

[7] G. Laudanno, B. Haegeman, D. L. Rabosky, and R. s. Etienne. Detecting lineage-specific shifts in diversification: A proper likelihood approach. *Systematic Biology*, 70(2):389–407, 2021. doi: 10.1093/sysbio/syaa048.

[8] W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, Sept. 1997. doi: 10.1093/sysbio/46.3.523.

[9] M. R. McGowen. Toward the resolution of an explosive radiation—A multilocus phylogeny of oceanic dolphins (Delphinidae). *Molecular Phylogenetics and Evolution*, 60(3):345–357, Sept. 2011. doi: 10.1016/j.ympev.2011.05.003.

[10] F. K. Mendes, J. A. Fuentes-Gonzĺez, J. G. Schraiber, and M. W. Hahn. A multispecies coalescent model for quantitative traits. *Elife*, 7:e36482, 2018. doi: 10.7554/elife.36482.

[11] R. Nichols. Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7):358–364, 2001. doi: https://doi.org/10.1016/S0169-5347(01)02203-0.

[12] P. Pamilo and M. Nei. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583, Sep 1988. doi: 10.1093/oxfordjournals.molbev.a040517.

[13] Y. Park, J. E. J. Patton, G. K. A. Hochberg, and J. W. Thornton. Comment on "ancient origins of allosteric activation in a Ser-Thr kinase". *Science*, 370, 2020. doi: 10.1126/science.abc8301.

[14] R. A. Pyron, C. R. Hendry, V. M. Chou, E. M. Lemmon, A. R. Lemmon, and F. T. Burbrink. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). *Molecular Phylogenetics and Evolution*, 81:221–231, Dec 2014. doi: 10.1016/j.ympev.2014.08.023.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

[16] D. W. Redding. Incorporating genetic distinctness and reserve occupancy into a conservation priorisation approach. Master's thesis, University Of East Anglia, Norwich, UK, 2003.

[17] D. W. Redding, F. Mazel, and A. Ø. Mooers. Measuring evolutionary isolation for conservation. *PLoS ONE*, 9(12):e113490, Dec 2014. doi: 10.1371/journal.pone.0113490.

[18] A. Rokas, B. L. Williams, N. King, and S. B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct. 2003. doi: 10.1038/nature02053.

[19] X.-X. Shen, J. L. Steenwyk, and A. Rokas. Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. *Systematic Biology*, 70(5):997–1014, Feb 2021. doi: 10.1093/sysbio/syab011.

[20] S. Song, L. Liu, S. V. Edwards, and S. Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947, Aug. 2012. doi: 10.1073/pnas.1211733109.

[21] M. S. Springer and J. Gatesy. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33, Jan. 2016. doi: 10.1016/j.ympev.2015.07.018.

[22] A. Stamatakis. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313, 2014. doi: 10.1093/bioinformatics/btu033.

[23] D. Swofford. PAUP*: Phylogenetic analysis using parsimony (*and other methods). *www.paup.phylosolutions.com, version 4a168*, 2021.

[24] M. Vellend, W. K.Cornwell, K. Magnuson-Ford, and A. O.Mooers. Measuring phylogenetic biodiversity. In A. E. Magurran and B. J. McGill, editors, *Biological Diversity: Frontiers in Measurement and Assessment*, chapter 14, pages 194–207. Oxford University Press, Oxford, 2011. ISBN 0199580677.

[25] K. Wicke, M. Fischer, and L. Kubatko. Effects of discordance between species and gene trees on phylogenetic diversity conservation. *Journal of Mathematical Biology*, 86(1), 2023. doi: 10.1007/s00285-022-01845-w.