Supplementary Information for "Effects of phylogenetic variation on prioritization of species for conservation"

Kristina Wicke, Md Rejuan Haque, and Laura Kubatko

In the following, we describe in more detail how the 9 multilocus data sets used in this study were processed prior to computing the Fair Proportion index.

Dolphin data set. This data set obtained from [5] contained DNA sequence data from 24 genes for 47 aquatic mammals. Due to large amounts of missing data, we reduced it to a "complete" data set by excluding 19 species (the species excluded were *D. tropicalis*, *S. frontalis*, *S. clymene*, *S. guianensis*, *L. peronii*, *L. cruciger*, *L. australis*, *C. hectori*, *C. heavisidii*, *C. eutropia*, *G. melas*, *Or_ breviostris*, *Ph_ dalli*, *I. geoffrensis*, *P. blainvillei*, *L. vexillifer*, *Z. cavirostris*, *P. minor*, and *M. novaeangliae*) and 2 genes (the genes excluded were STAT5 and TH), resulting in a total of 22 genes and 28 species. We then estimated gene trees under the GTR+Gamma model and with *P. macroce* as the outgroup using RAxML version 8.2.12 [12]. A species tree estimate was obtained using SVDquartets and maximum likelihood branch lengths were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP* (again using *Ph. macroce* as the outgroup).

Fungi data set. This data set was originally collected by [13] and re-analyzed by [9]. We downloaded the data set from [9] (`https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.9p8cz8wc5`) and used the data found in the directory `2_Shen_et_al_Empirical_datasets/2_fungi/removal_of_inconsistent_genes/ASTRAL` consisting of gene and species tree estimates based on 683 genes for a total of 25 individuals/variants of 18 bipolar budding yeasts and four outgroups. In order to be able to directly use the authors' estimated gene and species trees, we included all individuals in the analysis, even if they came from the same species. However, we re-computed maximum likelihood branch lengths for the species tree under the LG model (using the full amino acid sequence alignment found in file `29taxa_683genes_ML2ASTRAL_CCC.fas` in the directory `2_Shen_et_al_Empirical_datasets/2_fungi/removal_of_inconsistent_genes/Concatenation_ML`) and enforcing a molecular clock in PAUP*. Both gene and species trees were rooted using species *C. jadinii* as the outgroup.

Mammal data set. This data set was originally collated by [10] and re-analyzed by [11]. We obtained the DNA sequence data from [10] (`https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.3629v`) and gene tree estimates based on 447 genes for 33 species of mammals and 4 outgroup species from [11], and directly used the published gene trees (rooting them using *G. gallus* as the outgroup) for our analysis. A species tree estimate was obtained using SVDquartets, and maximum likelihood branch lengths for the species tree were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

*Corresponding author
Email address: `lkubatko@stat.osu.edu`

Plant data set. This data set was originally collated by [1] and re-analyzed by [9]. We downloaded the data set from [9] (`https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.9p8cz8wc5`) and used the data found in the directory `2_Shen_et_al_Empirical_datasets/3_plant/removal_of_inconsistent_genes` consisting of DNA sequence data as well as gene and species tree estimates based on 363 genes for 48 Lamiaceae and 4 outgroup species. We chose taxon *L. tibetica* as the outgroup and rooted the gene trees provided in the subfolder `ASTRAL`. However, 45 of them (corresponding to the following lines in the file `52taxa_363genes_ML2ASTRAL_CCC.trees`: 10, 28, 29, 34, 43, 49, 50, 52, 53, 58, 62, 63, 79, 95, 116, 123, 131, 144, 150, 155, 170, 188, 202, 208, 216, 222, 229, 239, 245, 256, 259, 273, 274, 278, 281, 296, 307, 309, 311, 312, 338, 343, 346, 354, 362) did not contain this outgroup species and were discarded, resulting in a total of 318 gene trees. We remark, however, that not all of the remaining 318 gene trees contained all 52 taxa, i.e., there was a certain amount of missing data. Moreover, we excluded the sequence data corresponding to the 45 removed genes from the full alignment (file `3_plant_52taxa_363genes_ML2ASTRAL_CCC.fas` found in the subfolder `Concatenation_ML`) and estimated a species tree based on the remaining 318 genes using SVDquartets. Finally, maximum likelihood branch lengths were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

Primate data set. This data set obtained from [2] contained DNA sequence data from 52 genes for four primate species. We estimated gene trees under the GTR+Gamma model and a species tree using RAxML version 8.2.12, respectively SVDquartets. In both cases, trees were rooted using *P. pygmaeus* as the outgroup. Finally, maximum likelihood branch lengths for the species tree were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

Rattlesnake data set. This data set from [4] contained DNA sequence data from 19 genes for 24 individuals of six subspecies of *Sistrurus* rattlesnakes and 2 outgroup species. Each diploid individual is represented by 2 sequences. We first randomly picked 1 sequence per subspecies and 1 outgroup sequence, i.e., 7 sequences in total. The individuals used were the outgroup ac1OUTG (*A. contortrix*), sca151WI (*S. c. catenatus*), sced127AZ (*S. c. edwardsii*), scter115K (*S. c. tergeminus*), smc1NC (*S. m. miliarius*), smi100FL (*S. m. barbouri*), and sms10K (*S. m. streckeri*). As 3 out of the 19 genes did not contain data for all species (LAM did not contain data for scter115K and smi100FL, ef did not contain data for sced127AZ, and fgb did not contain data for ac1OUTG), we subsequently estimated gene trees for only 16 genes under the GTR+Gamma model using RAxML version 8.2.12. Using SVDquartets we then estimated a species tree and computed maximum likelihood branch lengths for the species tree were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

Rodent data set. This data set was originally collated by [8] and re-analyzed by [9]. We downloaded the data set from [9] (`https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.9p8cz8wc5`) and used the data found in the directory `2_Shen_et_al_Empirical_datasets/1_animal/removal_of_inconsistent_genes` consisting of DNA sequence data as well as gene and species tree estimates based on 794 genes for 37 rodent species. More precisely, we used the gene trees provided in the subfolder `ASTRAL`, but discarded 33 of them, which were incomplete, i.e., which did not contain all 37 taxa (the gene trees removed correspond to the following lines in the file `37taxa_794genes_Rodent_ML2ASTRAL_CCC.trees`: 23, 94, 95, 105, 145, 155, 242, 251, 259, 272, 275, 315, 320, 332, 366, 368, 405, 513, 549, 579, 605, 608, 614, 618, 637, 673, 676, 681, 688, 716, 747, 757, 793). Thus, we used a total of 761 gene trees. Moreover, we excluded the corresponding regions from the full alignment (file `37taxa_794genes_Rodent_ML2ASTRAL_CCC.fas` found in the subfolder `Concatenation_ML`) and estimated a species tree based on the sequence data corresponding to the remaining 761 genes using SVDquartets [3] as implemented in the PAUP* package [14]. Finally, maximum likelihood branch lengths were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*. Both gene and species trees were rooted using *R. norvegicus* (Murine_Ref_Rat_20161116 (Rat37)) as the outgroup.

Snake data set. This data set obtained from [6] (`https://datadryad.org/stash/dataset/doi:10.5061/dryad.rb5nc`) contained DNA sequence data as well as gene and species tree estimates for 31 caenophidians and 2 outgroup species. For our analysis, we used gene and species trees estimated by the authors under a maximum likelihood framework (files `col33_ML.trees` and `col_ML33.tre`). The trees were rooted using species *A. carolinensis* as the outgroup. Finally, maximum likelihood branch lengths were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

Yeast data set. This data set from [7] contained DNA sequence data from 106 genes for 8 yeast species. We estimated gene trees under the GTR+Gamma model and a species tree using RAxML version 8.2.12, respectively SVDquartets. In both cases, trees were rooted using species *C. albicans* as the outgroup. Finally, maximum likelihood branch lengths for the species tree were computed under the GTR+Gamma model and enforcing a molecular clock in PAUP*.

References

[1] B. Boachon, C. R. Buell, E. Crisovan, N. Dudareva, N. Garcia, G. Godden, L. Henry, M. O. Kamileen, H. R. Kates, M. B. Kilgore, B. R. Lichman, E. V. Mavrodiev, L. Newton, C. Rodriguez-Lopez, S. E. O'Connor, D. Soltis, P. Soltis, B. Vaillancourt, K. Wiegert-Rininger, and D. Zhao. Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in lamiaceae. *Molecular Plant*, 11(8):1084–1096, Aug 2018. doi: 10.1016/j.molp.2018.06.002.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] E. J. Roycroft, A. Moussalli, and K. C. Rowe. Phylogenomics uncovers confidence and conflict in the rapid radiation of australo-papuan rodents. *Systematic Biology*, 69(3):431–444, Sept. 2019. doi: 10.1093/sysbio/syz044.

[9] 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.

[10] 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.

[11] 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.

[12] A. Stamatakis. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30:1312–1313, 2014.

[13] J. L. Steenwyk, D. A. Opulente, J. Kominek, X.-X. Shen, X. Zhou, A. L. Labella, N. P. Bradley, B. F. Eichman, N. Čadež, D. Libkind, J. DeVirgilio, A. B. Hulfachor, C. P. Kurtzman, C. T. Hittinger, and A. Rokas. Extensive loss of cell-cycle and DNA repair genes in an ancient lineage of bipolar budding yeasts. *PLOS Biology*, 17(5):e3000255, May 2019. doi: 10.1371/journal.pbio.3000255.

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