Online Appendix 2

Supplemental Methods

Summary Coalescent Species Tree Inference

Empirical data.— Gene trees were estimated for both the UCE contig alignments and UCE allele alignments with PhyML (Guindon et al. 2010), using the parallelized implementation in CloudForest `https://github.com/ngcrawford/CloudForest`. In order to compute node support in the summary coalescent species tree, we used CloudForest to generate 1000 non-parametric bootstrap replicates of each UCE gene tree dataset (contig and allele data, each consisting of 820 loci). This bootstrap algorithm re-samples nucleotides within the UCE alignments as well as UCE loci from the gene tree set, as described in Seo (2008). This resulted in 1000 replicates of the gene tree set (each replicate n=820 gene trees) for each of the two datasets. All trees were rooted using the `root()` function from the R-package APE (Paradis et al. 2004), assigning the sample of the genus *Florisuga* as the outgroup.

We used the software MP-EST (Yu et al. 2007) to estimate a summary coalescent species tree from the UCE gene tree data. The bootstrap gene tree datasets from the previous step were analyzed in MP-EST, assigning every individual as a separate taxon in the species tree (in the case of the allele dataset both alleles were assigned to the same taxon), in order to observe all individuals in the resulting species tree. We ran a separate MP-EST analysis for each bootstrap replicate. The resulting set of 1000 MP-EST species trees was summarized into one consensus tree with SumTrees v4.1.0 (Sukumaran and

Holder 2010), defining the root at the *Florisuga* sample. The node values on the consensus tree represent bootstrap support of the respective clade.

Simulated data.— We estimated gene trees and the summary coalescent species tree for both simulated datasets (simulated contig and allele data) using methods identical to those we used for the empirical *Topaza* UCE data (PhyML implementation in CloudForest and subsequent analysis in MP-EST). We ran two separate MP-EST analyses for each dataset, one in which we applied the species assignments under which the data were simulated and another where we assigned every individual to a separate taxon (identical to the empirical data), in order to observe the placement of every sample on the resulting species tree. The latter was repeated for each of the ten simulation replicates.

Supplemental Results

Gene Trees and Summary Coalescent Species Tree (MP-EST)

Empirical Topaza data.— The gene trees of both datasets (consensus and allele data) are generally poorly resolved and show a great variation of topologies (Figs. 1a and 1b, top panels), as was expected, due to the low number of phylogenetically informative sites per UCE locus. For the gene trees generated from allele data, we removed one allele of the outgroup taxon *Florisuga*, in order to root all trees using a single and consistent outgroup. In the case of the consensus data, 24% of gene trees return *T. pyra* as monophyletic while 17% of gene trees return *T. pella* as monophyletic. For the allele data, 14% of gene trees return *T. pyra* as monophyletic and 10% return *T. pella* as monophyletic (Supplementary Fig. S9 available on Dryad). These lower clade-frequencies in the allele dataset are most likely a consequence of more possible topologies with 19 terminals versus only 10 terminals in the consensus dataset.

Despite the uninformative gene tree topologies, the resulting summary coalescent species trees (Figs. 1a and 1b, center panels) support both *T. pyra* and *T. pella* with 100% bootstrap support as reciprocally monophyletic. However, the branch lengths in the species trees are quite obscured, with very short inter-nodal distances and long terminal branches. This tree shape has previously been recognized and referred to as "bonsai tree" (Gatesy and Springer 2014) and is typical of summary coalescent species trees that are based on loci with too little phylogenetic signal (Gatesy and Springer 2014; Springer and Gatesy 2014). The "bonsai tree" shape results from cases of inconsistent topological patterns in the gene tree data, because MP-EST arbitrarily assigns branch length values of around 9 coalescent units to terminal branches (Gatesy and Springer 2014), creating the observed long terminal branches.

Simulated data.— Similarly to the empirical data, no predominant topological pattern appears among the gene trees in either of the two simulated datasets (Figs. 1c and 1d, top panels). The two main clades in the simulation species tree, D,E and X,Y,Z (equivalent to *T. pyra* and *T. pella* in the empirical data), are monophyletic in only a small subset of the gene trees: D,E is monophyletic in 21% and X,Y,Z in 13% of the gene trees produced from the contig data, and in 12% and 19% of the gene trees derived from the allele data, respectively (Supplementary Fig. S9 available on Dryad).

The MP-EST species trees resulting from the simulated data show the same "bonsai tree" shape (Figs. 1c and 1d, center panels) as observed for the empirical data. The results reported here remain consistent throughout ten independently simulated datasets (Supplementary Figs. S10 and S11 available on Dryad). For the simulated contig dataset, neither of the two main clades (D,E or X,Y,Z) are supported as monophyletic in the MP-EST species tree, independently of the clade assignment model (Fig. 1c). The simulated allele dataset on the other hand yields more accurate results. For the analysis

without species assignments, the clade X,Y,Z is supported as monophyletic with 100% bootstrap support. The result improves further when applying the proper species assignment model from the simulation input tree, which leads to estimating the correct species tree topology (Fig. 1d).

Supplemental Discussion

MSC versus Summary Coalescent

Differently to the good performance of the sequence datasets under the MSC model, the same data did not perform as well in the summary coalescent approach. When analyzing the empirical data under the summary coalescent model, both morphospecies *T. pyra* and *T. pella* are supported as reciprocally monophyletic (100% bootstrap). However, the inferred topology within the two morphospecies differs from the MSC results and is not consistent between the contig data and the allele data, although well supported by high bootstrap values in both datasets (Figs. 1a and 1b). In deed, the simulated data show, that some clades in the MP-EST species tree are inferred incorrectly with false confidence, in some cases supported by 100% bootstrap support (Figs. 1c and 1d), which explains the contradicting but highly supported topologies in the empirical data. However, if the correct species assignments are applied to the simulated data, the allele sequences produce the correct species tree topology, while the results from the contig data remain incorrect. Another problem when analyzing our data under the summary coalescent approach, are the resulting branch lengths in the inferred species tree, which are non-informative and thus difficult to interpret. This is most likely a result of the rather conservative UCE alignments (few informative sites), which lead to largely unresolved gene trees (Fig. 1, upper panels). The set of rather unresolved gene trees in turn leads to the inference of very

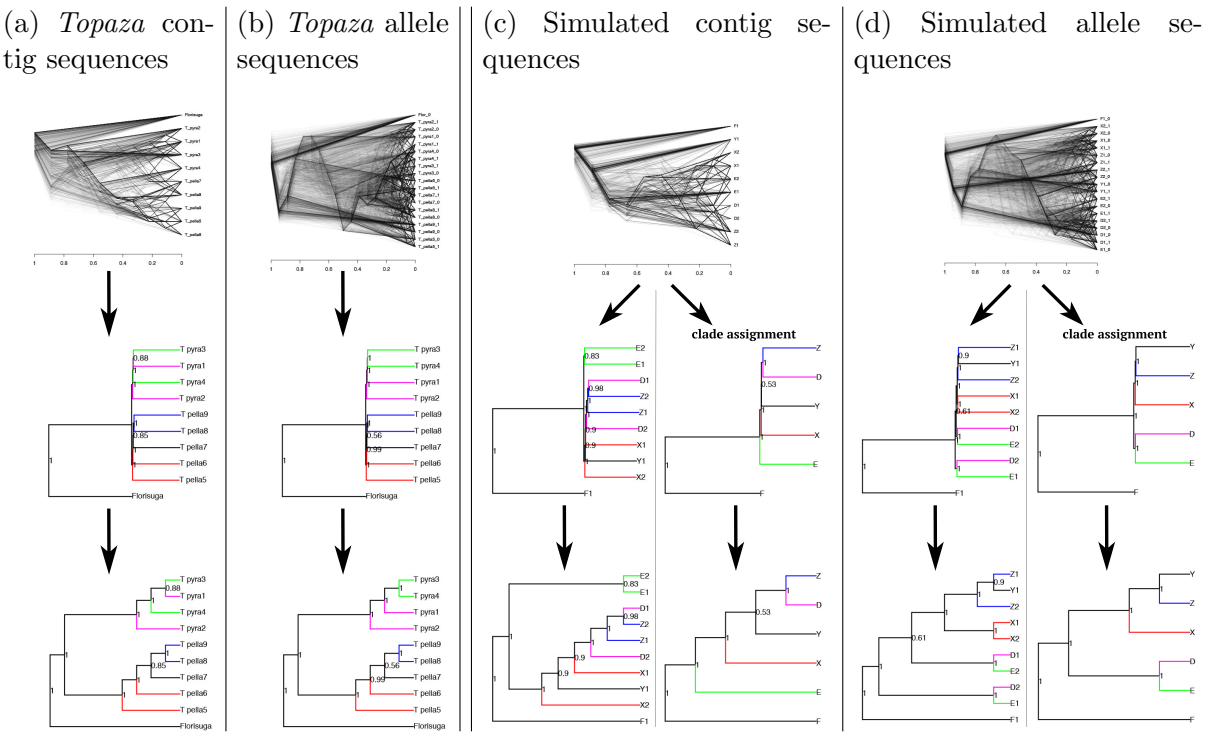

Figure 1: Summary coalescent species tree estimation for (a) empirical contig alignments, (b) empirical allele alignments, (c) simulated contig alignments, and (d) simulated allele alignments. Shown for each dataset (a-d) are all 820 gene tree topologies generated with PhyML (top panels), the resulting summary coalescent species tree as estimated with MP-EST (center panels), and the MP-EST species tree with branch lengths removed in order to improve readability (bottom panels). For the simulated data (c and d), we ran two analyses for each dataset: in the first scenario every individual was assigned as a separate taxon, and in the second scenario we applied the proper species assignments from the simulation input tree. The gene trees (top panels) are plotted without branch lengths, in order to visualize that only the topologies of gene trees are being used as input data when estimating the species tree with MP-EST, thus discarding all branch-length information.

short inter-nodal distances in the species tree, producing the odd "bonsai tree" shape (Gatesy and Springer 2014).

Analyzing the data under a summary coalescent model (using MP-EST) did not return the same intraspecific topology and for the simulated data returned incorrect topologies with erroneously high confidence (Fig. 1). This is mainly attributable to the fact that existing summary coalescent approaches require assigning sequences to species. If two alleles from the same individual are left unassigned (assuming they are separate species under the MSC) we knowingly violate the assumptions of the model. For this reason allele sequences cannot be treated as independent samples of a population in Summary Coalescent methods, but they have to be assigned to one coalescent taxon. Another problem when analyzing our data under the summary coalescent approach, are the resulting branch lengths in the inferred species tree, which are non-informative and thus difficult to interpret. This is most likely a result of the rather conservative UCE alignments (few informative sites), which lead to largely unresolved gene trees (Fig. 1, cloudogram). The set of rather unresolved gene trees in turn leads to the inference of very short inter-nodal distances in the species tree, producing the odd "bonsai tree" shape (Gatesy and Springer 2014).

*

References

Gatesy, J. and M. S. Springer. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Molecular Phylogenetics and Evolution 80:231–266.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010.

New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59:307–21.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Seo, T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Molecular Biology and Evolution 25:960–71.

Springer, M. S. and J. Gatesy. 2014. Land plant origins and coalescence confusion. Trends in Plant Science 19:267–9.

Sukumaran, J. and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–71.

Yu, L., Y.-W. Li, O. a. Ryder, and Y.-P. Zhang. 2007. Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. BMC Evolutionary Biology 7:198.