Syst. Biol. 73(2):308–322, 2024

https://doi.org/10.1093/sysbio/syad076
Advance Access Publication January 3, 2024

Artifactual Orthologs and the Need for Diligent Data Exploration in Complex Phylogenomic Datasets: A Museomic Case Study from the Andean Flora

Laura A. Frost[1,2,*], Ana M. Bedoya[1], and Laura P. Lagomarsino[1,*]

[1]*Shirley C. Tucker Herbarium, Department of Biological Sciences, Louisiana State University, Life Science Annex Building A257, Baton Rouge, LA 70803, USA*
[2]*Biology Department, University of South Alabama, 5871 USA N Dr, Mobile, AL 36688, USA*
[*]*Correspondence to be sent to: Laura Frost, Shirley C. Tucker Herbarium, Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA; Email: lafrost@southalabama.edu; Laura Lagomarsino, Louisiana State University, 202 Life Science Building, Baton Rouge, LA 70803, USA; Email: llagomarsino1@lsu.edu.*
Laura Frost and Ana M. Bedoya contributed equally to this article.

Received 21 October 2022; reviews returned 20 November 2023; accepted 2 January 2024
Associate Editor: Ya Yang

Abstract.—The Andes mountains of western South America are a globally important biodiversity hotspot, yet there is a paucity of resolved phylogenies for plant clades from this region. Filling an important gap in our understanding of the World's richest flora, we present the first phylogeny of *Freziera* (Pentaphylacaceae), an Andean-centered, cloud forest radiation. Our dataset was obtained via hybrid-enriched target sequence capture of Angiosperms353 universal loci for 50 of the ca. 75 spp., obtained almost entirely from herbarium specimens. We identify high phylogenomic complexity in *Freziera*, including the presence of data artifacts. Via by-eye observation of gene trees, detailed examination of warnings from recently improved assembly pipelines, and gene tree filtering, we identified that artifactual orthologs (i.e., the presence of only one copy of a multicopy gene due to differential assembly) were an important source of gene tree heterogeneity that had a negative impact on phylogenetic inference and support. These artifactual orthologs may be common in plant phylogenomic datasets, where multiple instances of genome duplication are common. After accounting for artifactual orthologs as source of gene tree error, we identified a significant, but nonspecific signal of introgression using Patterson's D and f4 statistics. Despite phylogenomic complexity, we were able to resolve *Freziera* into 9 well-supported subclades whose evolution has been shaped by multiple evolutionary processes, including incomplete lineage sorting, historical gene flow, and gene duplication. Our results highlight the complexities of plant phylogenomics, which are heightened in Andean radiations, and show the impact of filtering data processing artifacts and standard filtering approaches on phylogenetic inference. [Andean radiation; Angiosperms353; data artifacts; gene tree filtering; introgression; locus filtering; museomics; paralogy.]

The Andean mountains in South America are one of the most species-rich areas of the world and serve as a center of diversity for many plant groups (Gentry 1982; Mutke and Barthlott 2005). The recent uplift of the Andes has resulted in some of the fastest plant evolutionary radiations reported to date (Madriñán et al. 2013; Hughes 2016), with some greatly phenotypically diverse clades (Hughes and Eastwood 2006; Lagomarsino et al. 2016). A significant portion of Andean plant biodiversity is yet to be described (Ulloa Ulloa et al. 2017) and many, if not most Andean plant species, have never been included in a phylogeny. However, these phylogenies are fundamental toward understanding the evolutionary patterns that contribute to the origin and diversification of biodiversity in the World's richest flora.

Establishing well-supported phylogenies for Andean plant groups is challenging for many reasons. Short divergence times between speciation events, incomplete lineage sorting (ILS), incipient speciation, and introgression all contribute to poor phylogenetic resolution and high gene tree-species tree discordance (Vargas et al. 2017; Morales-Briones et al. 2018; Lagomarsino et al. 2022). This is further complicated by repeated whole genome duplication events throughout the evolutionary history of plants, at both deep and shallow scales (One Thousand Plant Transcriptomes Initiative 2019). There are additional practical limitations for phylogenetic inference in Andean systems. It is difficult to achieve full taxon sampling as species are often narrowly endemic and distributed in remote locations, and members of clades occur in many countries, each with different policies concerning the collection and exportation of samples. As a result, achieving dense sampling of Andean-centered lineages commonly requires the use of herbarium specimens as a source of genetic material, which is associated with lower quality and quantity DNA than in freshly-collected or silica-dried leaf tissue (Bakker et al. 2015).

Improvements in methodology in the past decade bring us closer to achieving resolved phylogenies in previously intractable groups. Advancements in genomic sequencing, including hybrid-enriched target sequence capture, allow for the collection of hundreds to thousands of loci, even from degraded DNA from natural history collections (Bakker et al. 2015; McKain et al. 2018). Further, the development of universal probe sets facilitates the sequencing of hundreds of loci for any system, regardless of the genomic resources available (Johnson et al. 2019). Analytical methods are also increasingly able to accommodate multiple biological sources of gene tree discordance (Ogilvie et al. 2017; Solís-Lemus et al. 2017; Zhang et al. 2018).

Still, a major barrier to phylogenomic inference in many plant clades is the difficulty in efficient identification of orthologous and paralogous sequences from multicopy loci (Yang and Smith 2014; Morales-Briones et al. 2021). The presence of multiple sequences for a single locus within an individual may result from allelic variation or gene/genome duplication. The latter process has repeatedly taken place during the evolutionary history of plant lineages, and results in the presence of nonorthologous gene copies within the same genome (i.e., paralogs) (Li and Barker 2020). Automated detection of multicopy genes in phylogenomic datasets could be hindered when sequence data fails to meet contig number, contig length, or read depth thresholds set during assembly, a scenario that is more likely when is DNA obtained from herbarium specimens (Bakker et al. 2015).

The presence of only one copy of a multicopy gene in a sequenced dataset (i.e., hidden paralogy) may be due to biological processes or to data processing artifacts. "Pseudo-orthologs" are a type of hidden paralogy that results from a biological process: differential loss after gene duplication leads to the presence of a single but nonorthologous copy across species in nature (Smith and Hahn 2021, 2022). While there is some evidence that coalescent-based phylogenetic methods are relatively robust to pseudo-orthologs (Smith and Hahn 2021, 2022), this finding is applicable only under certain conditions that do not include whole genome duplication followed by rediploidization— a phenomenon common in plants (Li et al. 2021). Hidden paralogy can also be artifactual, including when differential assembly of copies in a multicopy gene results in the recovery of nonorthologous sequences for a given locus across taxa. We refer to these as "artifactual orthologs." Although it is well-documented that unrecognized paralogy can have negative impacts on species tree inference (Brown and Thomson 2017), the presence of artifactual orthologs in plant phylogenomic datasets and their impact on phylogenetic inference remains underexplored.

We combat the many challenges of Andean plant phylogenomics to infer the first phylogeny and introgression history of *Freziera* (Pentaphylacaceae), a group that constitutes a cloud forest plant radiation that previously lacked any phylogenetic information. *Freziera* includes 75 spp. of trees and shrubs that are widely distributed throughout montane regions of the Neotropics, from southern Mexico to Bolivia, with a center of diversity (61 spp.) in Andean cloud forests (Santamaría-Aguilar and Monro 2019; Fig. 1). There has been at least one whole genome duplication event in an ancestor of Pentaphylacaceae within the inclusive order Ericales (Larson et al. 2020), and chromosome counts are highly variable in the family (i.e., *Adinandra* [2 spp., n = 42], *Cleyera* [1 spp., n = 45], *Eurya* [4 spp., n = 21, 29, 42], and *Ternstroemia* [2 spp., n = 20, 25]; from Chromosome Counts Database [CCDB;(Rice et al. 2015); ccdb.tau.ac.il]). Ploidy levels in *Freziera* are currently unknown.

Using Angiosperms353 (Johnson et al. 2019) target enrichment data derived almost entirely from herbarium specimens, we establish a phylogenetic baseline of *Freziera* despite widespread artifactual orthology and find evidence for historical introgression. We further explore how various types of data curation help remove artifactual orthologs and their impact on species tree inference. We finally provide suggestions to

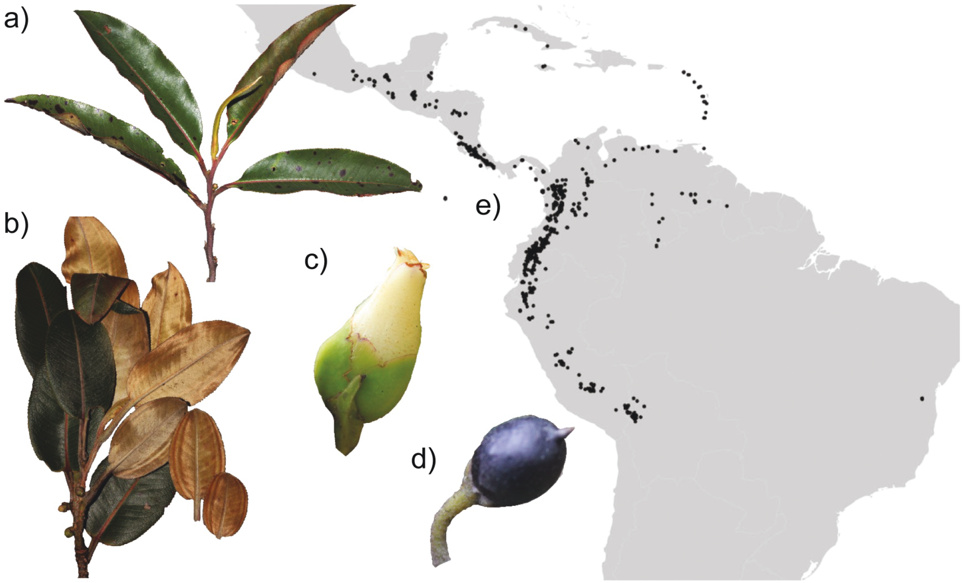

FIGURE 1. Diversity and distribution of *Freziera*. *Freziera* has significant variation in leaf morphology and pubescence, as illustrated by branches of a) *F. candicans* and b) *Freziera sp.*; meanwhile, c) flower and d) fruit morphology are relatively stable. *Freziera* is an Andean radiation; e) most species are distributed in this mountain chain in western South America, with some species in Central America, the Guiana Shield, the Caribbean, and the Atlantic Forest of Brazil. Photos by L. Lagomarsino.

explore complex empirical phylogenomic datasets, especially those with a history of genome duplication and that are obtained with a high reliance on natural history collections.

Materials and Methods

Taxon Sampling

Ninety-four accessions representing 55 *Freziera* species—approximately 73% of the species diversity—were sampled for the ingroup. All but 5 came from herbarium specimens, which had an average age of 31.6 years (range: 7.4–82.8; Supplementary Table 1). Nine accessions from other Pentaphylacaceae were sampled for the outgroup, including *Eurya japonica* (in the genus sister to *Freziera),* *Cleyera albopunctata* (a member of tribe Frezierieae), and 7 species of *Ternstroemia* (belonging to the sister tribe Ternstroemieae; Weitzman et al. 2004; Tsou et al. 2016).

DNA Extraction, Library Prep, Target Enrichment, and Sequencing

Detailed descriptions of laboratory methods are provided in online Appendix 1. Briefly, DNA extraction followed a modified sorbitol extraction protocol (Štorchová et al. 2000). Library preparation used the KAPA Hyper Prep and KAPA HiFi HS Library Amplification kits with iTru i5 and i7 dual-indexing primers. Target enrichment was carried out using the MyBaits Angiosperms353 universal probe set (Johnson et al. 2019; Hale et al. 2020). DNA libraries were sequenced by Novogene in one Illumina Hiseq 3000 lane with 150 bp paired-end reads. Although most samples come from herbarium specimens, DNA concentration met minimum standards for sequencing.

Raw Data Processing

Demultiplexed raw sequence reads were trimmed with illumiprocessor v2.0.9 (Faircloth 2013, 2016), a wrapper for Trimmomatic v0.39 (Bolger et al. 2014). Default settings were used and reads with a minimum length of 40 bp kept. Trimmed reads were assembled into supercontigs (exons and flanking intronic regions) with HybPiper v1.3.1 (Johnson et al. 2016) using the Angiosperms353 target file as reference (Johnson et al. 2019). In addition, we used a taxon-specific target file for Pentaphylacaceae using Easy353 (Zhang et al. 2022), a reference-guided assembly tool for recovery of Angiosperms353 gene sets. We used the transcriptome of *Ternstroemia gymnanthera* (One Thousand Plant Transcriptomes Initiative 2019) for sequence retrieval with Easy353. Additional details of data processing prior to final gene tree inference are available in online Appendix 1.

Gene Tree Inference

Preliminary gene trees from the HybPiper v1.3.1 supercontig alignments were generated from aligned sequences for 322 loci lacking paralog flags with RAxML v.8.2.12 (Stamatakis 2014) under the GTRCAT nucleotide substitution model with 200 rapid bootstraps. The resulting trees were processed with TreeShrink v1.3.3 (Mai and Mirarab 2018) to detect and remove unusually long branches (i.e., potential cross contaminants). Processing in TreeShrink was performed on a "per-gene" and "all-gene" basis. The identified branches (most of which corresponded to samples with few sequenced reads; Supplementary Table S1) were removed from final alignments. We then removed alignments with fewer than 25 ingroup samples from further analyses. Accessions present in <10% of processed gene trees were trimmed using the R package ape (Paradis and Schliep 2019a). Gene trees were inferred with IQ-TREE multicore v2.1.1 (Nguyen et al. 2015) including model selection via ModelFinder (Kalyaanamoorthy et al. 2017), tree inference by ML estimation, ultrafast bootstraps (Hoang et al. 2018), and Shimodaira-Hasegawa-like approximate likelihood-ratio test (SH-aLRT); (Guindon et al. 2010; Anisimova et al. 2011).

Paralog Warnings and Detection of Artifactual Orthologs

Despite assembling a single sequence per sample for the majority of loci and raising paralog warnings for <9% loci with HybPiper v1.3.1 (Supplementary Table S1), visual observation of gene trees and alignments suggested the presence of undetected paralog sequences (i.e., artifactual orthologs) in the dataset. Evidence for this included split clades of outgroup species and distinctive motifs in alignments that reflected higher sequence divergence than expected by allelic variation alone (Fig. 2b). Gene trees without paralog warnings were first examined by-eye in FigTree v1.4.3 (Rambaut 2014) and sorted as putative orthologs or artifactual orthologs. Artifactual orthologs are identified when different accessions for the same ingroup species, each represented by only a single sequence in the alignment, cluster into 2 subclades separated by relatively long internal branches (Fig. 2b). Visual detection of artifactual paralogs from gene trees in phylogenomic datasets thus benefits from sampling multiple individuals per species. We grouped loci with no paralog warnings from HybPiper v1.3.1. and confirmed by-eye as putative orthologs into a gene tree set called *orthologs.by.eye* (Table 1).

We compared our by-eye identification of artifactual orthologs with 2 automated paralogy detection strategies: HybPiper v2.0 (Johnson et al. 2016) and HybPhaser (Nauheimer et al. 2021). Both versions of HybPiper could detect paralogs if multiple assembled contigs each cover ≥75% of a reference sequence; this returns a long paralog warning in HybPiper2. HybPiper2 also raises paralog warnings when multiple contigs are assembled across ≥75% of the reference sequence even when individual contigs are <75% the full sequence length; this returns a paralog-by-depth warning. In addition, HybPiper2 reports stitched contigs (i.e., those derived from multiple SPAdes contigs), which can be

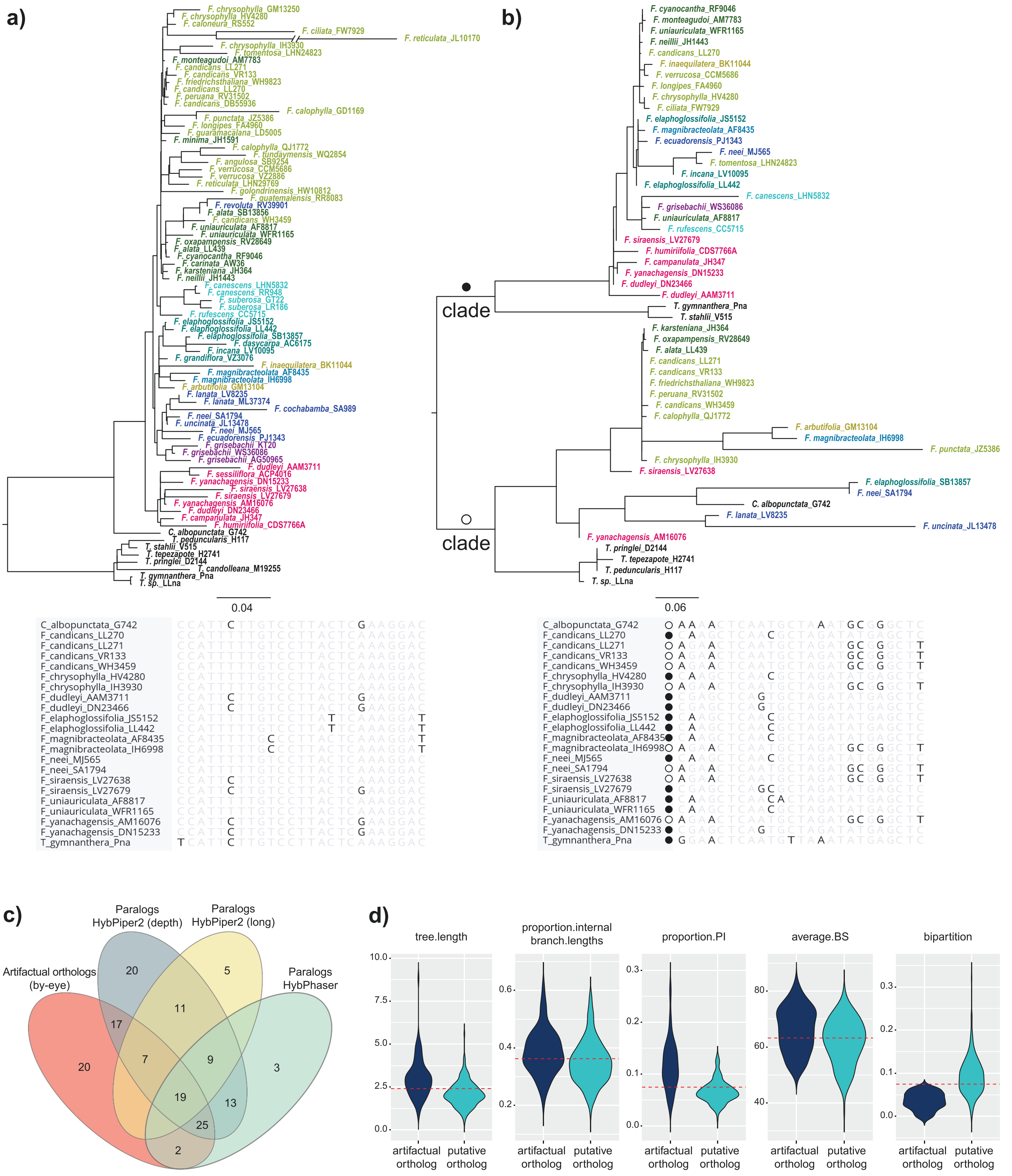

FIGURE 2. Exemplary gene trees and alignment subsets for a) a putative ortholog, and b) an artifactual ortholog. Putative ortholog gene trees exhibit monophyletic genera for outgroups and appropriate relationships to the ingroup, a relatively shallow backbone in the ingroup, and relationships similar to the preliminary species tree. A relatively low level of variation is present in alignments for putative orthologs, which is fitting of a locus in a universal probe set being applied within a genus. Artifactual ortholog gene trees often have deep divergences between subsets of the ingroup, including between samples from the same species. Within those subclades, patterns of relationships from the preliminary species tree are repeated. In some cases, as in b), outgroups are also non-monophyletic and subsets of outgroup samples are recovered as sister to the ingroup subclades. Alignments for artifactual orthologs exhibit relatively high variation with observable motifs between the suspected copies, as indicated by the black and white dots. c) Overlap between artifactual orthologs identified by-eye, and paralog

Table 1. Descriptions of the filtering criteria used, names applied to each of the datasets, number of loci (as a subset of the 313 *orthologs.unfiltered*) selected by each, and the impact of filtering criteria on gene tree discordance, branch support, and species tree topology.

Description of filtering criteria for keeping loci	Putative orthologs dataset name after filtering	No. of loci that passed filter	Normalized quartet score	Average branch support	Number of highly supported branches (***n*** = 77; **PP > 0.97)**	Two clades in Candicans group	Mono-phyletic Elapho-glossifolia group	Branching order of Elapho-glossifolia group	Placement of Arbutifolia clade
Loci not flagged as paralogs by HybPiper v.1.3.1	*orthologs.unfiltered*	313	0.567	0.774	30	+	–	n/a	+
Loci identified as putative orthologs after by-Eye inspection of gene trees and alignments	*orthologs.by.eye*	182	**0.634**	**0.792**	29	+	+	+	+
Loci with 1 contig at 75% reference sequence (no long paralog warnings from HybPiper v.2.0)	*orthologs.hybpiper2.long*	262	**0.582**	**0.78**	29	+	+	+	+
Loci without paralog warnings by depth or length from HybPiper v.2.0	*orthologs.hybpiper2.no.warnings*	187	**0.619**	**0.783**	27	+	+	+	+
Loci that have a lower proportion of SNPs than 1.5x the interquartile range above the 3rd quartile for any given sample	*HybPhaser*	242	**0.597**	**0.783**	28	+	+	+	+
gene trees with a total tree length above the average value across all gene trees	*tree.length*	140	0.524	0.688	20	–	–	n/a	–
Gene trees with the proportion of total tree length comprising internal branch lengths above the average proportion across all gene trees	*proportion.internal.branch.lengths*	149	0.542	0.685	18	–	+	+	+
Alignment contains ^{3}7.5% parsimony informative sites in ingroup (average value across all alignments)	*proportion.PI*	160	0.528	0.715	19	–	–	n/a	+
Gene trees with average bootstrap support across all branches of each gene tree above the average value across all gene trees	*average.BS*	163	0.56	0.727	18	+	+	–	+
Gene trees with above average bipartition support relative to the species tree inferred from the *orthologs.unfiltered* dataset	*bipartition*	166	**0.601**	**0.792**	**30**	+	+	+	+
Orthology inference via gene tree pruning using homologs with monophyletic, nonrepeating outgroups	*MO*	222	**0.662**	0.664	17	+*	+*	–	–

Values for datasets that performed equally or better than the orthologs.unfiltered dataset for each metric are bolded. Major topological conflicts between species trees and a consensus topology of the 11 filtered datasets (Fig. 3b) are shown, indicating instances where the species tree is compatible (+) or in disagreement (–) with the consensus topology. Datasets from which the Elaphoglossifolia group was not resolved as monophyletic are indicated with "n/a." Asterisks mark disagreement from the alternative placement of only one species. The "branching order of the Elaphoglossifolia group" includes conflict in the placement of *F. magnibracteolata*.

warnings raised by automated paralogy detection strategies. d) Performance of gene tree filtering strategies on selection of artifactual and putative orthologs. Orthology was determined by eye. The dashed lines indicate the various thresholds above which gene trees were kept in filtered datasets, as indicated in Table 1. Portions of the violin plots above the threshold were included in the dataset filtered by the indicated criterion; portions below were excluded.

used to identify chimeric sequences (i.e., stitched contigs in which the two separate sequences are derived from different gene copies); HybPiper2 returns chimera warnings when a sufficient number of read pairs from stitched contigs map to different SPAdes contigs and have sequence mismatch above a given threshold for one read but not the other (see HybPiper2 documentation for additional details; https://github.com/mossmatters/HybPiper/). HybPhaser is an automated strategy to identify hybrids, polyploids, contamination, and paralog sequences by identifying samples with a high proportion of heterozygous loci and allele divergence, and loci with a high proportion of SNPs (Nauheimer et al. 2021).

In addition, we applied a method for automated orthology detection (i.e., gene tree pruning) using homologs with monophyletic outgroup (MO; Table 1; Yang and Smith 2014; Morales-Briones et al. 2022). MO identifies clusters in gene trees with monophyletic outgroups and searches from root to tips for duplications. When duplicated taxa are found on either side of a bifurcation, the subtree with the fewest ingroup taxa is pruned. MO does not directly identify artifactual orthologs but has the potential to remove their effect by selecting orthogroups and allowing the inclusion of more loci in a final dataset.

We generated 5 sets of gene trees to compare artifactual orthology identification using the above automated pipelines; a description of filtering criteria and thresholds for each dataset are in Table 1. Loci that received a paralog warning in HybPiper v1.3.1 were excluded to generate the *orthologs.unfiltered* dataset. The remaining filtered datasets (i.e., *orthologs.HybPiper2.long*, *orthologs.HybPiper2.no.warnings, HybPhaser,* and *MO*) are subsets of the *orthologs.unfiltered* dataset. For the *HybPhaser* dataset, we did not filter loci by missing data to ensure that only loci with excess heterozygosity (loci with more than 1.5× the interquartile range above the third quartile), as expected in putative paralogs, were removed. We kept ortholog groups with ≥25 ingroup taxa in the *MO* dataset for comparison with the other filtering strategies.

Impact of Standard Gene Tree Filtering Strategies on the Presence of Artifactual Orthologs

We explored how standard gene tree filtering affects the presence of artifactual orthologs (as identified by-eye) in our dataset. Gene trees were filtered using 5 common empirical criteria for phylogenomic subsampling (Table 1): 1 alignment-based metric (*proportion.PI*), 2 tree length metrics (*tree.length* and *proportion.internal.branch.lengths*), and 2 gene tree support metrics (*average.BS* and *bipartition*). Gene tree filtering is typically applied to minimize gene tree estimation error, gene tree discordance (Molloy and Warnow 2018), and for phylogenomic subsampling (Mongiardino Koch 2021) but applying these metrics may also favor the selection of artifactual orthologs in a dataset. Due to the presence of nonorthologous copies in alignments including artifactual orthologs, a high degree of genetic variation and an increase in the proportion of informative sites is expected compared with true orthologs. We also expected elevated average bootstrap support in gene trees inferred from artifactual orthologs because of well-supported clusters of nonorthologous sequences. Filtering by bootstrap support or collapsing poorly supported nodes in gene trees prior to species tree inference are standard practice to remove gene tree estimation error (Zhang et al. 2018). The 2 tree length criteria were selected because, while a high percentage of internal branch length can signal high phylogenetic signal (Shen et al. 2016), it can also indicate biological pseudo-orthologs (Smith and Hahn 2022) and likely, artifactual orthologs (Fig. 2b). Finally, filtering by bipartition support may remove artifactual orthologs by favoring loci that are more concordant with the inferred species tree (Smith et al. 2018).

Standard gene tree filtering strategies were applied to the *orthologs.unfiltered* dataset. The *proportion.PI* was obtained using AMAS (Borowiec 2016) with outgroups removed. Estimates of *proportion.internal.branch.lengths* were extracted from IQ-TREE ‘*.iqtree’ output files. Bootstrap support values were extracted from IQ-TREE maximum likelihood tree files and averaged to calculate *average.BS*. For *tree.length*, input gene trees were rooted using pxrr in phyx (Brown et al. 2017) and tree length was calculated with the script ‘get_var_length.py’ of SortaDate (Smith et al. 2018) excluding outgroups. *Bipartition* support was calculated with the SortaDate “get_bp_genetrees.py” script against a species tree generated with ASTRAL-III (Zhang et al. 2018) from the *orthologs.unfiltered* dataset. Cutoff thresholds for each filter were selected near the average value in our dataset to keep a similar number of loci for each criterion.

Phylogenetic Relationships and Introgression in Freziera

Impact of filtering criteria on species tree inference.—Species trees were inferred in ASTRAL-III from the 11 datasets listed in Table 1. A majority-rule consensus tree was generated from the species trees for the 11 filtered datasets (*consensus* function of the R package *ape*; Paradis and Schliep 2019b) to identify major clades that were concordant across datasets. Gene tree discordance was assessed for all species tree topologies using the final normalized quartet score (i.e., the proportion of quartet trees in gene trees that are present in the species tree) estimated with ASTRAL-III (Mirarab et al. 2014b). Impact of filtering criteria on branch support was assessed by calculating the average branch support and proportion of well-supported branches (ppl ≥ 0.97; Rabiee and Mirarab 2020) across the 11 resulting species trees. We inferred a second consensus tree from the species trees generated from the datasets that showed the greatest improvements to both normalized quartet score and average branch support (*bipartition, by.eye, HybPhaser,* and *orthologs.HybPiper2.no.warnings*), which also had substantial reductions in the number of artifactual orthologs (Fig. 2c). The resulting tree was compared with the *orthologs.unfiltered* dataset to explore the impact that the removal of paralogs, including artifactual orthologs, has on species tree inference.

Signals of introgression in Freziera.—To assess genomic evidence of introgression in species of *Freziera*, we calculated Patterson's D and f4 statistics using the *Dtrios* function in Dsuite v.0.4 r43 (Malinsky et al. 2021). The input VCF file was generated with dDocent (Puritz et al. 2014) by mapping all *Freziera* trimmed reads to the sequences of *T. tepezapote* for the loci in the *bipartition*, *by.eye*, and *hybiper2.no.warnings* datasets. We called SNPs using default values for all mapping parameters. The resulting VCF files were filtered with vcftools (Danecek et al. 2011). We retained SNPs with <50% missing data and retained only one SNP per 100 bp window to decrease the likelihood of including linked SNPs. The statistical significance of the D and f4 was assessed using block jackknife on windows of 75–78 SNPs followed by Benjamini-Hochberg correction as implemented in R (Benjamini and Hochberg 1995) to assess family-wise error rate following (Malinsky et al. 2018). The D and f4 statistics were estimated for all possible trios across 3 datasets (13,244, 15,180, 14,190 trios for the *bipartition*, *by.eye*, and *hybiper2.no.warnings* respectively). The ASTRAL-III species trees inferred from each of the 3 datasets were specified in Dsuite so that the D and f4 estimated values were arranged according to the tree.

The relatively small number of loci in our dataset limited the power to detect introgression along a phylogeny. The f-branch statistic, which accounts for the correlation of D and f4 due to shared ancestry among multiple potential introgression donor species (Malinsky et al. 2018), allows for a better interpretation of introgression patterns across a tree (Malinsky et al. 2021). However, simulation analyses have shown that for a relatively small number of unlinked SNPs (<10,000), the proportion of cases where the strongest inferred f-branch signal corresponds to the correct simulated gene flow recipient and donor branches is <20% (Malinsky et al. 2021). Due to the limited number of unlinked SNPs in our targeted sequence capture dataset, we did not apply this metric.

Results

Sequence capture efficiency

Voucher information for accessions and per sample data for read trimming and contig assembly are available in Supplementary Table S1. Of the 102 accessions for which target enrichment libraries were prepared and sequenced, 22 (21 *Freziera* and 1 outgroup) were excluded either because no sequences were assembled for any of the 353 target loci, or because they were identified to have exceedingly long branches with TreeShrink (most of the latter corresponded to samples with few sequenced reads; Supplementary Table S1). Target enrichment efficiency across all sequenced samples is shown in Supplementary Figure S1. Seventy-two accessions representing 50 of the 75 species of *Freziera* and 8 members of the outgroup remained (see methods; Supplementary Table S1), which had assembled sequences for 52–348 genes (average: ~296). Standard assembly statistics from both versions of HybPiper are in Supplementary Tables S1-S2. No loci were flagged by HybPiper2 as putatively chimeric. However, all but 13 loci included at least one sequence with stitched contigs (Supplementary Table S3). All sequences without stitched contigs correspond to regions spanning a single exon. Off-target data collection was insufficient to assemble plastomes.

Paralog Warnings and Detection of Artifactual Orthologs

The 11 paralogy detection and gene tree filtering criteria that we applied generated datasets with 140–313 loci (Table 1). Results of artifactual ortholog detection by-eye and from automated strategies are shown in Fig. 2a; Supplementary Table S3. HybPiper v.1.3.1 raised paralog warnings for 31 loci (i.e., >1 contig covering 75% of the target sequence was assembled for at least one sample for a given locus). Of the 322 loci that were not flagged as paralogs, 9 were removed because they contained fewer than 25 ingroup samples. This left 313 loci with a single assembled supercontig (exons plus introns) per species, many of which were composed of multiple short contigs stitched together (see HybPiper documentation for more detailed descriptions of supercontigs and stitched contigs: https://github.com/mossmatters/HybPiper/; Supplementary Table S3). By-eye inspection of these 313 loci resulted in the identification of 90 artifactual orthologs. Loci flagged as potentially paralogous in HybPiper v.1.3.1 also had long paralog warnings and paralog warnings by depth with HybPiper2. Of the 313 loci without paralog warnings from HybPiper v.1.31, 51 had long paralog warnings, of which 46 also received paralog warnings by contig depth with HybPiper2. An additional 75 loci were not flagged as long paralogs but did receive paralog warning by contig depth for at least one sample, totaling 121 loci with warnings by contig depth and 126 loci with any kind of paralog warning issued by HybPiper2. HybPhaser identified 73 loci as putative paralogs (Supplementary Table S3; online Appendix 2), most of which also had paralog warnings by depth. There was substantial overlap of loci identified by these methods (Fig. 2c). Three of the methods we compared—HybPiper2 paralog-by-depth warnings, HybPiper2 paralog-by-length warnings, and HybPhaser—flagged 70 (78%) of the loci that we identified as an artifactual orthologs in our by-eye assessment (Fig. 2c). Of these, HybPiper2 paralog-by-depth warnings had the greatest overlap with our by-eye assessment (Fig. 2c; Supplementary Table S3). Assembly with a taxon-specific target file resulted in similarly low levels of paralog detection as the universal target file: <8% of assembled loci recovered >1 long contig for at least

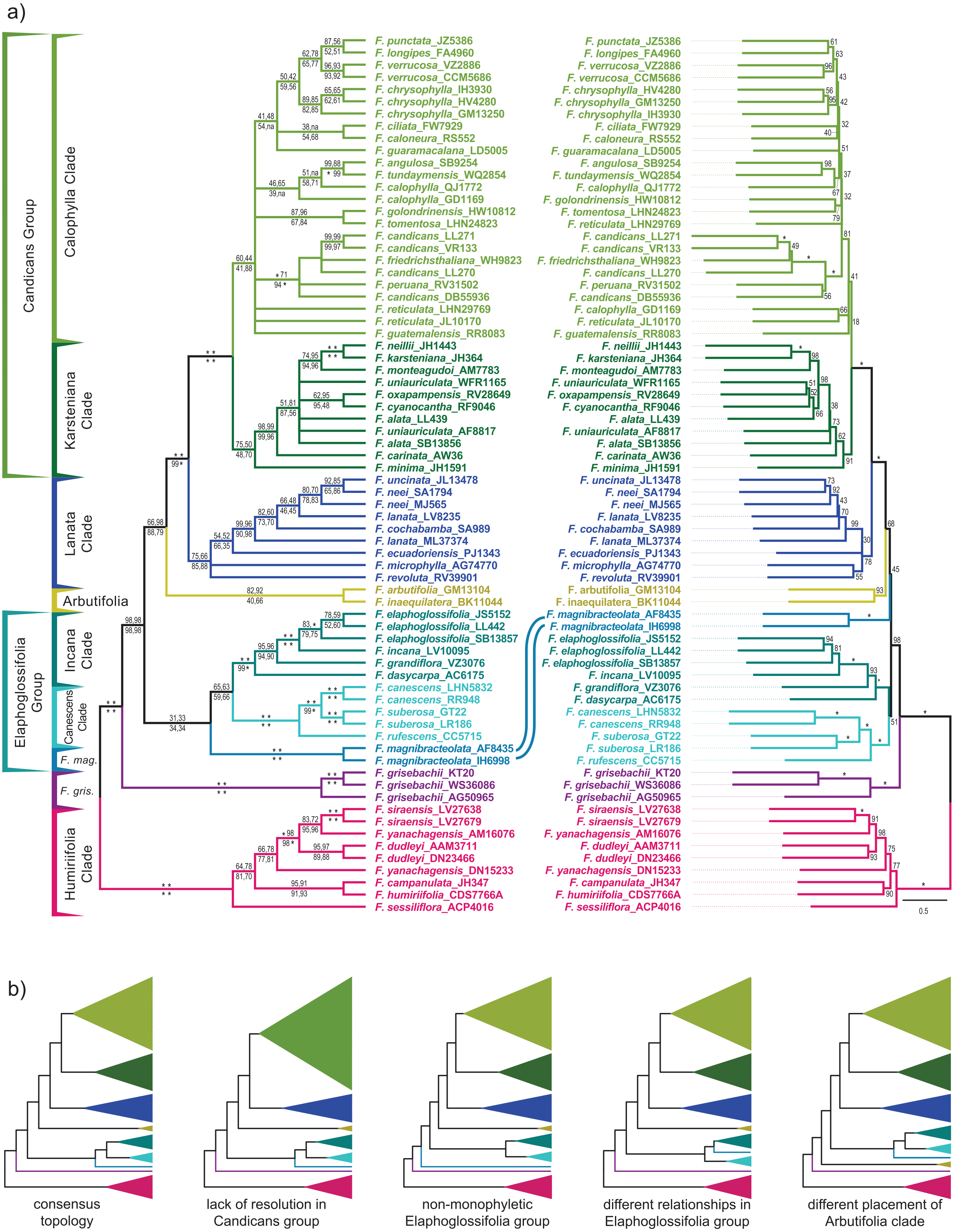

FIGURE 3. Names of clades consistently recovered and species tree topologies for a) The consensus tree of the four datasets in which the proportion of artifactual orthologs was reduced (*bipartition*, *by.eye*, *HybPhaser*, *orthologs.HybPiper.no.warnings*; left) versus the orthologs. unfiltered species tree (right). The placement of *F. magnibracteolata*– the major topological conflict between the two trees– is highlighted with blue lines connecting the tip labels in each. Branches with full support (LPP from ASTRAL = 1) are indicated by an asterisk; elsewhere, the

one sample (Supplementary Table S4). There was a correlation between the number of reads mapped and the number of paralog warnings from HybPiper in a sample (Supplementary Fig. S3).

Impact of Standard Gene Tree Filtering Strategies on the Presence of Artifactual Orthologs

Summary statistics for alignments and gene trees are available in Supplementary Table S3. The ability of gene tree filtering strategies to remove artifactual orthologs (as identified by-eye) is shown in Fig. 2d. A high proportion of putative orthologs were also removed with all filtering strategies, and only filtering gene trees by bipartition support resulted in the removal of most artifactual orthologs. As commonly applied, the remaining four gene tree filtering criteria all resulted in datasets in which there were either similar proportions of putative orthologs and artifactual orthologs in the final datasets (*average.BS, proportion.internal.branch.length*), or higher proportions of artifactual orthologs than putative orthologs (*tree.length*, *proportion.PI*).

Impact of Data Curation on Phylogenetic Performance

We found that removing artifactual orthologs from our datasets improved phylogenetic performance metrics, and that, with the exception of filtering by bipartition support, gene tree filtering did not. Datasets in which curation reduced the proportion of artifactual orthologs (*orthologs.by.eye, orthologs.HybPiper2*, *HybPhaser*, and *bipartition*) had the highest normalized quartet scores and average branch support (Table 1). Except for *bipartition*, these values were consistently lower in datasets resulting from standard gene tree filtering methods (Table 1), likely due to an increased proportion of artifactual orthologs relative to the unfiltered dataset (Fig. 2b). Relative to the unfiltered dataset, the number of highly supported branches was similar in datasets in which artifactual orthologs were filtered out, and lower in datasets that underwent conventional gene tree filtering, again apart from *bipartition* (Table 1). Although there was some impact on removing artifactual orthologs on branch support, it is possible that these results were relatively small due to the high degree of gene tree heterogeneity in our dataset (Table 1).

We also found consistent impacts of removing artifactual orthologs on the species tree topology. Datasets where artifactual orthologs were removed (*by.eye, HybPiper2.long, HybPiper2.depth, HybPhaser, bipartition*) consistently recovered the Elaphoglossifolia group as monophyletic. Contrastingly, species tree topologies generated by gene tree filtering differed from any of those generated from datasets that were curated to remove artifactual orthologs, and sometimes introduced relationships that were found in no other analyses, including of the unfiltered dataset (e.g., nonmonophyletic Callophylla and Karsteniana clades and placement of the Arbutifolia clade; Table 1; Fig. 3b).

Phylogenetic Relationships and Introgression in Freziera

The majority-rule consensus tree of all 11 species trees resulting from data filtering and the consensus tree of the four datasets with the lowest proportions of artifactual orthologs were congruent in the major clades recovered and the relationships between those clades. Seven clades that were frequently inferred across analyses were identified—the Humiriifolia, Canescens, Incana, Arbutifolia, Lanata, Karsteniana, and Calophylla clades—along with *F. grisebachii* and *F. magnibracteolata*, which formed monotypic clades (Fig. 3a). The Elaphoglossifolia group, comprising the Canescens and Incana clades and F. magnibracteolata, was additionally recovered as monophyletic in both consensus trees. Five of the 11 datasets produced species trees consistent with the 9 clades and their branching order in the consensus tree (Table 1; Supplementary Fig. S2). Among species trees consistent with the consensus topology, some branches at deep nodes were inferred with high support (ppl≥0.97) across all analyses: the common ancestor of all *Freziera*, the common ancestor of core *Freziera* (all *Freziera* excluding the Humiriifolia clade), and the successive node within core *Freziera*, the Humiriifolia, Canescens, and Incana clades, and the Candicans group, which comprises the Karsteniana, and Calophylla clades (Fig. 3). Species-level relationships within most subclades were consistent across analyses except for the species-rich Candicans group (Supplementary Fig. S4).

Regions of lowest support in the consensus tree tended to be in conflict between species trees and the consensus topology. These regions often involved 2 notable sets of taxa: the Elaphoglossifolia group, comprising the Incana and Canescens clades and *F. magnibracteolata*, and the Candicans group, which comprises the Calophylla and Karsteniana clades (Table 1, Fig. 3; Supplementary Fig. S4). Although the Elaphoglossifolia group was monophyletic in most datasets, including all in which the proportion of artifactual orthologs was reduced (Table 1), the unstable placement of *F. magnibracteolata* rendered it nonmonophyletic 3 datasets: *orthologs.unfiltered*, *proportion.PI*, and *tree.length* (Table 1; Fig. 3b; Supplementary Fig. S2a,g,j). Contrasting with the Elaphoglossifolia group, the monophyly of the Candicans group was identified in all analyses, generally with high support despite a high degree of gene tree heterogeneity; however, the resolution of its constituent Calophylla and Karsteniana subclades varied

LPP * 100 is provided. Support for each of the 4 datasets in the consensus tree is given clockwise from the top left: *bipartition*, *by.eye*, *HybPhaser*, *orthologs.HybPiper.no.warnings*. b) Cartoon trees displaying the most frequent differences in relationships between clades recovered by different datasets and the consensus tree of all datasets. Cartoons depict one topological scenario but not the only topological outcome.

across analyses (Table 1). Although all datasets in which the proportion of artifactual orthologs was reduced resolved the Calophylla and Karsteniana subclades as monophyletic, species-level relationships within these groups varied (Supplementary Fig. S4).

For the *bipartition*, *by.eye*, and *orthologs.HybPiper2.no.warnings* datasets, *P*-values for the estimated *D* statistics could not be calculated for 75.9%, 85.5%, and 93.6%, respectively, of all evaluated trios due to lack of ABBA-BABA variants (Supplementary Fig. S5). Estimates of Patterson's $D > 0$ were significant (corrected *P*-value ≤ 0.05, *Z*-score >3) for 104 (3.27%), 162 (7.38%), and 46 (5.08%) trios out of the remaining evaluated trios. The number of statistically significant *D* statistics was probably underestimated in our study given the large proportion of trios for which *P*-values could not be calculated due to a lack of ABBA-BABA variants (~76%–94%). Resulting f4-ratio statistics across the 3 datasets also show evidence for multiple instances of introgression in *Freziera* (Supplementary Fig. S5). These results were further supported by the results from HybPhaser, which identified several samples with >80% locus heterozygosity and >1% allele divergence across all major groups recovered (online Appendix 2; Supplementary Fig. S6). This is indicative of prevalence of introgression and polyploidy (Nauheimer et al. 2021) in *Freziera* species included in this study.

Discussion

Relying on DNA extracted primarily from herbarium specimens, we inferred the first phylogeny of *Freziera* (Pentaphylacaceae), an understudied tropical plant lineage. Our museomic dataset, representing hybrid-enriched target sequence capture of Angiosperms353 loci, highlights the many challenges of working with understudied clades even in the genomic era: no a priori phylogenetic hypothesis, a universal bait set, poor-quality DNA, and high proportion of paralogs, many of which we identified as artifactual orthologs. Despite phylogenomic complexity including a high degree of gene tree heterogeneity, we resolve *Freziera* into nine clades whose histories have been shaped by myriad evolutionary processes, including incomplete lineage sorting, introgression, and gene or genome duplication. In the face of these complexities, we identify that the biggest improvements to phylogenetic inference in *Freziera* did not come from filtering gene trees to maximize phylogenetic informativeness. Instead, they came from reducing the noise from artifactual orthologs, which was accomplished in a variety of ways: observing data by-eye, implementing automated pipelines, and identifying gene tree filtering mechanisms that are consistent with reducing this artifact. We offer recommendations on strategies for removal of similar data processing artifacts for phylogenetic inference of groups where multicopy loci are expected to be prevalent.

Identification of Artifactual Orthologs Using Automated Pipelines

We observed a widespread pattern of artifactual orthology in our dataset, in which multicopy genes were recovered as single copy due to errors in the assembly process. These loci had multiple identifiable motifs in their alignments, and their resulting gene trees typically exhibited polyphyly of species and deep divergences between clades, including outgroup taxa and members of those spuriously polyphyletic species (Fig. 2b). As troubling as the prospect of unfiltered paralogs may be for empiricists, artifactual orthologs may be easier to detect than biological pseudo-orthologs (Smith and Hahn 2022) because of these striking, easily identifiable patterns.

Using our visual inspection as a baseline, we were able to assess the ability of automated paralogy detection pipelines to identify artifactual orthologs. There was substantial overlap in the loci that we identified by-eye and those removed through various automated paralog detection mechanisms (Fig. 2a). Although artifactual orthologs result from the assembly of one contig for a truly multicopy locus, they were often identifiable if more than one short contig covered >75% of the target length (the strategy used in *orthologs.HybPiper2.no.warnings*). Similarly, but to a lesser extent, artifactual orthologs were removed by filtering out loci with high proportion of SNPs (*HybPhaser*) and with the assembly of multiple contigs at least 75% of sequence length (*orthologs.HybPiper2.long*). We recommend by-eye identification as an initial strategy to explore whether artifactual orthologs are present in an individual empirical dataset, and emphasize the utility of sampling multiple individuals per species.

Our dataset relied heavily on degraded DNA extracted from herbarium tissue, which we believe contributes to the assembly of artifactual orthologs. Though high-molecular weight DNA is fragmented during library preparation, the desired fragment size for Illumina libraries is 200-500 base pairs (bp; Bronner and Quail 2019). As is common when working with specimens from the wet tropics (Bakker et al. 2015), many of our samples had a high proportion of fragments <200 bp (unpublished). These short fragments restrict data collection in noncoding regions, as fragments are less likely to span the regions flanking targeted exons. Using HybPiper v2.0.1, we identified a high proportion of stitched contigs—separate contigs concatenated into a single sequence. The assembly of multiple contigs belonging to different paralog copies into a single sequence (i.e., chimeric sequences) is an artifact that is possible in the presence of stitched contigs. The preferential selection of the longest contig with the highest coverage for a given gene region during assembly further contributes to decreased paralog detection and increased chimeric assembly. Despite the high proportion of stitched contigs, no genes were flagged by HybPiper v2.0.1 with a chimera warning for any sample (Supplementary Table S2). This is likely due to current limitations and stringency in chimeric sequence detection. Specifically, paired ends must completely map to separate contigs and reads must map entirely within an exon on the separate contigs to flag a chimera warning. Short fragments from degraded DNA reduce the likelihood of finding read pairs that will meet these criteria, and, therefore, the likelihood of detecting chimeric sequences, though they may be present in the dataset.

Effectively identifying chimeric sequences from single or stitched contigs (as even a single assembled contig for a given locus may be a chimera of different alleles) remains a challenge in the assembly of target capture data from short read sequences, especially as assembly pipelines cannot distinguish between reads derived from different paralogs or alleles. Detection of chimeric sequences and its impact on phylogenetic inference remains a fundamental problem in phylogenomics, particularly in groups where polyploidization and hybridization are suspected to be prevalent (Morales-Briones et al. 2018). Although not possible with degraded DNA from herbarium specimens, a potential solution includes generating long-read sequence data at a sufficient depth for accurate phasing of copies.

We believe the lack of data spanning introns, resulting in a high proportion of stitched contigs, and the potential for greater disparity between capture of both copies from herbarium DNA, may be a primary source of artifactual orthologs. Low sequencing depth resulting in poor coverage of noncoding regions is another factor that can increase the proportion of stitched contigs and/or differential success capturing copies. However, sequencing depth is unlikely to be the main source of artifactual orthologs in our dataset as the number of reads mapped to target loci (Supplementary Table S1) in our dataset is higher per sample than studies that have successfully examined paralogs in Angiosperms353 datasets (Johnson et al. 2019; Gardner et al. 2020). Taken together, this highlights caveats that remain with museomic data: long-read sequence data is not an option in many cases and deeper sequencing cannot recover missing data. Herbarium specimens are an invaluable resource for improving sampling and filling gaps in the tree of life, and we do not advocate for the exclusion of herbarium specimens in phylogenomic datasets. Rather, we recommend careful assessment of datasets considering these artifactual complications in assembly.

Impact of Standard Gene Tree Filtering Strategies in the Presence of Artifactual Orthologs

Gene tree filtering is a commonly applied strategy to minimize gene tree estimation error (GTEE) and its impact on species tree estimation (Molloy and Warnow 2018). Without an explicit method to estimate GTEE in empirical data, multiple criteria including alignment length, the number/proportion of variable or parsimony informative sites, total tree length, the proportion of internal branch lengths, and average bootstrap support of gene trees have all been used as implicit proxies to account for GTEE (Leaché et al. 2014; Liu et al. 2015; Shen et al. 2016; Blom et al. 2017). It is argued that these metrics select for "higher quality" gene trees, however, this assumption is violated in the presence of artifactual orthologs, which are associated with higher values of many standard metrics (Fig. 2c).

Only one filtering criterion successfully reduced the proportion of artifactual orthologs in the *Freziera* dataset: bipartition support. The high efficiency of bipartition support (Fig. 2d) is likely due to the major topological differences between the consensus species tree and gene trees of artifactual orthologs (Figs. 2 and 3). However, a significant proportion of putative orthologs were also removed through this filtering mechanism (Fig. 2d). This curation strategy should be used with caution and may require additional assessment of loci falling below the threshold if investigating biological processes such as ILS or introgression, as some highly discordant, single-copy loci will also be filtered. Given that the most successful methods of automated paralog detection require assembly with HybPiper, bipartition support could provide a useful metric by which users can remove artifactual orthologs in datasets assembled using other pipelines.

Both artifactual orthologs and putative orthologs were removed at similar levels from datasets by the remaining 4 gene tree filtering criteria, in particular *tree.length* and *proportion.PI.* This demonstrates that filtering by alignment or gene tree characteristics using standard techniques may actually result in a relative increase of data artifacts in complex phylogenomic datasets.

Although gene tree filtering metrics have the potential to significantly improve phylogenetic support (Doyle et al. 2015) and clarify relationships, it may be an inappropriate strategy for some studies (Molloy and Warnow 2018). This is true in the case of *Freziera*, in which gene tree filtering resulted in lower branch support and spurious topological inference for all gene tree filtering criteria except bipartition support (Table 1, Supplementary Fig. S2). This is likely the result of the removal of a large proportion of single-copy loci relative to artifactual orthologs (Fig. 2b). The impact of the presence of artifactual orthologs and gene tree filtering strategies on species tree inference will likely vary across datasets. We recommend careful observation of phylogenomic datasets before applying these criteria of data curation.

Impact of Data Curation on Phylogenetic Performance in the Face of Artifactual Orthologs

Species tree inference of *Freziera* with a 2-step, coalescent-aware species tree inference algorithm (i.e., ASTRAL-III) was relatively robust to artifactual orthology. This is consistent with recent studies demonstrating that species tree methods are robust to the presence of paralogs (Yan et al. 2022), though undetected paralogs have also been documented to mislead species tree inference (Brown and Thomson 2017; Siu-Ting et al. 2019). Across analyses of our 11 datasets, backbone relationships were largely consistent and topological differences were primarily concentrated in a few portions of the phylogeny, at least one of which (the Candicans group) corresponds to a rapid radiation where levels of ILS and introgression are likely high. We found that the unfiltered dataset (i.e., *orthologs.unfiltered*) had a slightly higher proportion of well-supported nodes, though this number was close in absolute value to those from curated datasets with a reduced proportion of artifactual orthologs (Table 1). This may be simply a result of a larger dataset, as coalescent-based phylogenetic inference algorithms require a large number of independent loci to resolve challenging relationships (Leaché and Rannala 2011; Mirarab et al. 2014a).

Reducing artifactual orthologs had an overall positive impact on phylogenetic inference and support (Table 1; Fig. 3). Curated datasets where artifactual orthologs were removed (*orthologs.by.eye, orthologs.HybPiper2, HybPhaser, bipartition*; Table 1) had higher normalized quartet scores and average bootstrap support relative to the unfiltered dataset and consistently recovered the monophyly of the Elaphoglossifolia group via the placement of *F. magnibracteolata*, a relationship that was not present in the unfiltered dataset (Table 1). Improvements in these datasets are likely the result of the reduction in gene tree heterogeneity with the removal of the noise introduced by artifactual orthologs, though gene tree heterogeneity persists in curated datasets and likely reflects biological processes that have shaped *Freziera*'s evolutionary history (Table 1). Ensuring that paralogs, including artifactual orthologs, are appropriately handled in phylogenetic analyses is not only essential for accurate phylogenetic estimation, it is also crucial to downstream analyses that require accurate branch lengths, including divergence date estimation (Siu-Ting et al. 2019).

Orthology inference using monophyletic outgroup (*MO*) was not associated with improved phylogenetic performance, despite being successfully applied in datasets where paralogs are represented as multiple copy loci in assembled datasets (Morales-Briones et al. 2022). The *MO* species tree had reduced gene tree discordance (i.e., higher normalized quartet sampling score) relative to the unfiltered dataset, likely from the removal of nonorthologous copies. However, it failed to recover some of the major clades identified from curated data where artifactual orthologs were removed. This may be due to information loss since trees may be extensively pruned by MO, resulting in smaller subtrees with fewer taxa. MO is a valuable tool to identify orthologous clusters in complex phylogenomic datasets with multicopy paralogs and can increase phylogenetic resolution in groups with a history of polyploidy (Morales-Briones et al. 2022); however, it cannot be used to identify artifactual orthologs and did not improve phylogenetic performance in *Freziera*, where paralogs were predominantly single-copy.

Phylogenetic Relationships and Introgression in Freziera

One of the barriers to the study of neotropical diversification is the difficulty resolving phylogenies of recent, rapid Andean radiations. Despite *Freziera*'s low species richness compared with many cloud forest plant clades, we find similar hallmarks of explosive radiation in our phylogenetic results. In the face of phylogenomic complexity—including the presence of multiple copy paralogs, artifactual orthologs, rapid radiations, and DNA extracted from herbarium specimens—relationships among species and subclades of *Freziera* were consistently inferred across curated datasets. Most topological differences across species trees in datasets with a reduced proportion of artifactual orthologs were within the Candicans group, a rapid radiation with short internal branch lengths. We resolve *Freziera* into nine subclades, most of which are moderate to well supported (Fig. 3). Despite their monophyly, these clades generally lack morphological synapomorphies and are not geographically structured. The Humiriifolia clade, which is sister to core *Freziera*, best represents the wide morphological diversity in each subclade: its species have among the largest (*F. humiriifolia*) and smallest (*F. yanachagensis*) leaves in the genus, despite occurring in close proximity in the Cordillera del Cóndor region of southern Ecuador. Although our phylogenetic backbone is generally well-supported, there are 3 nonmutually exclusive biological sources that explain the very high levels of gene tree discordance in our dataset: introgression, ILS, and gene (or genome) duplication. Despite a genome-wide phylogenomic dataset, we are unable to pinpoint exactly when and where along the phylogeny each of these processes has occurred. However, our resulting phylogenetic framework provides a robust starting point from which to understand the nonbifurcating nature of diversification of this Andean shrub clade.

Incomplete lineage sorting is common in rapid Andean radiations (Morales-Briones et al. 2018; Murillo-A et al. 2022). Within *Freziera*, the Candicans group, especially its substituent Calophylla clade, carries the hallmarks of ILS due to rapid radiation, including conflicting species relationships with short branch lengths between close relatives (Fig 3a; Supplementary Fig. S4). Not only is gene tree-species tree discordance higher in this clade compared with other regions of *Freziera*'s phylogeny, but species relationships also differ across analyses (Supplementary Fig. S4). A further indication of ILS in this clade is the fact that the Calophylla clade is the most widespread of *Freziera*'s subclades and includes species with some of the broadest distributions in the genus. These distributional patterns allow a greater possibility that ancestral populations were large and widespread, contributing to ILS, amplifying the effect of short times between speciation events in this group.

Even after applying data curation that may inflate support for bifurcating relationships, a strong signal of introgression was identifiable in *Freziera*. This was evidenced by multiple significant f4 and D statistics (Supplementary Fig. S5). Due to the limited number of unlinked SNPs in our target capture data and the correlation of D and f4 introgression statistics when trios share internal branches, we were not able to isolate the exact branches along which introgression has occurred—a challenge even among the most complete datasets (Tricou et al. 2022). However, our resulting phylogenetic framework provides a robust starting point from which to understand the nonbifurcating nature of *Freziera*'s diversification. Areas of conflict across species trees in our curated datasets are strong candidates for lineages that have been directly shaped by past introgression (Fig. 3, Supplementary Fig. S4). A species that is particularly promising for future investigation is *F. magnibracteolata*, whose unstable placement along the backbone of *Freziera* (Table 1. Fig. 3; Supplementary Figs. S2,S4) suggests a potential history of introgression (MacGuigan and Near 2019; Cai et al. 2021). Notably, the placement of this species was the only major relationship to be impacted by removing artifactual orthologs relative to the unfiltered dataset (Table 1, *Monophyletic Elaphoglossifolia group*). To further assess the extent of gene flow and identify the branches involved in introgression events in this rapid radiation, future research will target a much larger portion of the genome (Malinsky et al. 2021) and include deeper taxons with multiple individuals per species.

Finally, either gene or genome duplication has resulted in paralogs in *Freziera*. We identify both paralogs for which multiple copies are identifiable during assembly and artifactual orthologs that are represented by only a single copy in our dataset (Fig. 2b, Supplementary Table S3). Although the processes that gave rise to these paralogs are yet to be examined in detail, it is likely that they are the product of allopolyploidy, especially considering the extensive history of genome duplication via polyploidy in Ericales, the order to which *Freziera* belongs (Larson et al. 2020), as well as chromosome count variation within Pentaphylacaceae (Rice et al. 2015).

Conclusion

A major current challenge in phylogenomics is the difficulty in teasing apart specific sources of gene tree discordance in empirical datasets and accounting for these in phylogenetic inference (Morales-Briones et al. 2022; Tricou et al. 2022). It is a significant challenge to accurately identify paralogs, pinpoint specific instances of introgression, disentangle incomplete lineage sorting from historical gene flow, and reduce the impact of gene tree estimation error in a single empirical phylogenomic dataset in which all of these sources of discordance are present. In addition to these challenges, phylogenomic data have the potential to be very complex, particularly for clades that are well-understood to be recalcitrant like Andean plant radiations (Pease et al. 2016; Vargas et al. 2017). Here, we showed that careful data curation allowed us to detect a high proportion of artifactual orthologs, which we were able to reduce with multiple, nonmutually exclusive methods: heeding paralog warnings, removing gene trees with a high proportion of heterozygous sites, and filtering gene trees using bipartition support. These data curation strategies were subsequently associated with higher support, lower gene tree conflict, and a more stable species tree— the first for an understudied tropical plant clade that previously lacked any phylogenetic information. We advocate for the observation of empirical phylogenomic data, including gene tree alignments and topologies, and that data curation be tailored to unique properties of individual datasets to better address the above-mentioned complexities in phylogenetic inference.

Although commonly used filtering techniques, assembly parameters, and other automated aspects of phylogenomics are powerful tools for improving phylogenetic inference, we have shown that they can also increase the proportion of data artifacts (i.e., artifactual paralogs) and have negative impacts on phylogenetic support and inference. Automated filtering techniques are not a replacement for a deep understanding of a dataset. Selecting filtering strategies for individual datasets should be informed by the latter because the decision is likely to be a balance between minimizing the presence of data artifacts while maximizing the number of loci useful for phylogenetic inference. Although targeted sequence capture of universal loci offers potential, especially for phylogenetic studies relying heavily on DNA from natural history collections, these datasets are not without limitations related to the nature of the data themselves and to the algorithms we use to process and analyze them. Combining the exploration of datasets with deep knowledge of the organismal biology of targeted clades is crucial towards overcoming these limitations and inferring robust phylogenetic hypotheses.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.v9s4mw72k

Acknowledgments

We would like to thank the Missouri Botanical Garden (MO) for access to their collections. We thank Brant Faircloth, Matthew Johnson, Carl Oliveros, and Jessie Salter for their guidance in library preparation, and Brant Faircloth for access to laboratory equipment. Computational analyses were performed on LSU High Performance Computing's SuperMike cluster. We thank three anonymous reviewers for their comments and criticism, which greatly improved previous versions of this work. This manuscript benefited from feedback from Laymon Ball, Janet Mansaray, and Diego Paredes-Burneo. The taxonomic expertise of Daniel Santamaría-Aguilar was critical throughout the design and implementation of this research.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. DEB-2055525. Additional funding was provided by a Louisiana Board of Regents Research Competitiveness Subprogram grant and the Louisiana State University College of Science and Office of Research and Economic Development.

References

Anisimova M., Gil M., Dufayard J.-F., Dessimoz C., Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60:685–699.

Bakker F.T., Lei D., Yu J., Mohammadin S., Wei Z., van de Kerke S., Gravendeel B., Nieuwenhuis M., Staats M., Alquezar-Planas D.E., Holmer R. 2015. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol. J. Linn. Soc. 117:33–43.

Benjamini Y., Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc.: Series B. Stat. Methodol. 57:289–300.

Blom M.P.K., Bragg J.G., Potter S., Moritz C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66:352–366.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ 4:e1660.

Bronner I.F., Quail M.A. 2019. Best practices for Illumina library preparation. Curr Prot. Hum. Genet. s 102:e86.

Brown J.M., Thomson R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Brown J.W., Walker J.F., Smith S.A. 2017. Phyx: phylogenetic tools for unix. Bioinformatics 33:1886–1888.

Cai L., Xi Z., Lemmon E.M., Lemmon A.R., Mast A., Buddenhagen C.E., Liu L., Davis C.C. 2021. The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, malpighiales. Syst. Biol. 70:491–507.

Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824–837.

Faircloth B.C. 2013. Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. http://dx.doi.org/10.6079/J9ILL

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786–788.

Gardner E.M., Johnson M.G., Pereira J.T., Puad A.S.A., Arifiani D., Wickett N.J., Zerega N.J.C. 2020. Paralogs and off-target sequences improve phylogenetic resolution in a densely-sampled study of the breadfruit genus (Artocarpus, Moraceae). Syst. Biol. 70:558–575.

Gentry A.H. 1982. Neotropical floristic diversity: phytogeographical connections between Central and South America, Pleistocene climatic fluctuations, or an accident of the Andean orogeny? Ann. Mo. Bot. Gard. 69:557–593.

Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 30. Syst. Biol. 59:307–321.

Hale H., Gardner E.M., Viruel J., Pokorny L., Johnson M.G. 2020. Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. Appl. Plant Sci. 8:e11337.

Hoang, D.P., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35: 518–522.

Hughes C.E. 2016. The tropical Andean plant diversity powerhouse. New Phytol. 210:1152–1154.

Hughes C.E., Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of plant diversification after uplift of the Andes. Proc. Natl. Acad. Sci. U.S.A. 103:10334–10339.

Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett N.J. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4:1600016.

Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis D.E., Soltis P.S., Wong G.K.-S., Baker W.J., Wickett N.J. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-Medoids clustering. Syst. Biol. 68:594–606.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

Lagomarsino L.P., Condamine F.L., Antonelli A., Mulch A., Davis C.C. 2016. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). New Phytol. 210:1430–1442.

Lagomarsino L.P., Frankel L., Uribe-Convers S., Antonelli A., Muchhala N. 2022. Increased resolution in the face of conflict: phylogenomics of the Neotropical bellflowers (Campanulaceae: Lobelioideae), a rapid plant radiation. Ann. Bot 129:723–736.

Larson D.A., Walker J.F., Vargas O.M., Smith S.A. 2020. A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. Am. J. Bot. 107:773–789.

Leaché A.D., Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60:126–137.

Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R., Bauer A.M., Nielsen S.V., Greenbaum E., Rödel M.-O., Schmitz A., LeBreton M., Ineich I., Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T. 2014. A hybrid phylogenetic–phylogenomic approach for species tree estimation in African *Agama* lizards with applications to biogeography, character evolution, and diversification. Mol. Phylogenet. Evol. 79:215–230.

Li Z., Barker M.S. 2020. Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. GigaScience 9:giaa004.

Li Z., McKibben M.T.W., Finch G.S., Blischak P.D., Sutherland B.L., Barker M.S. 2021. Patterns and processes of diploidization in land plants. Annu. Rev. Plant Biol. 72:387–410.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360:36–53.

MacGuigan D.J., Near T.J. 2019. Phylogenomic signatures of ancient introgression in a rogue lineage of darters (Teleostei: Percidae). Syst. Biol. 68:329–346.

Madriñán S., Cortés A.J., Richardson J.E. 2013. Páramo is the world's fastest evolving and coolest biodiversity hotspot. Front. Genet. 4:192.

Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. BMC Genomics 19:272.

Malinsky M., Matschiner M., Svardal H. 2021. Dsuite—Fast *D*-statistics and related admixture evidence from VCF files. Mol. Ecol. Resour. 21:584–595.

Malinsky M., Svardal H., Tyers A.M., Miska E.A., Genner M.J., Turner G.F., Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat. Ecol. Evol. 2:1940–1955.

McKain M.R., Johnson M.G., Uribe-Convers S., Eaton D., Yang Y. 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6:e1038.

Mirarab S., Bayzid M.S., Warnow T. 2014a. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65:366–380.

Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Mongiardino Koch N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol. Biol. Evol. 38:4025–4038.

Morales-Briones D.F., Gehrke B., Huang C.-H., Liston A., Ma H., Marx H.E., Tank D.C., Yang Y. 2022. Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in Alchemilla sl. (Rosaceae). Syst. Biol. 71:190–207.

Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling sources of gene tree discordance in phylogenomic data sets: testing ancient hybridizations in Amaranthaceae sl. Syst. Biol. 70:219–235.

Morales-Briones D.F., Liston A., Tank D.C. 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytol. 218:1668–1684.

Murillo-A J., Valencia-D J., Orozco C.I., Parra-O C., Neubig K.M. 2022. Incomplete lineage sorting and reticulate evolution mask species relationships in Brunelliaceae, an Andean family with rapid, recent diversification. Am. J. Bot. 109:1139–1156.

Mutke J., Barthlott W. 2005. Patterns of vascular plant diversity at continental to global scales. Biol. Skr 55:521–531.

Nauheimer L., Weigner N., Joyce E., Crayn D., Clarke C., Nargar K. 2021. HybPhaser: a workflow for the detection and phasing of hybrids in target capture data sets. Appl. Plant Sci. 9:e11441.

Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34:2101–2114.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574:679–685.

Paradis E., Schliep K. 2019a. ape 50: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Paradis E., Schliep K. 2019b. ape 50: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 14:e1002379.

Puritz J.B., Hollenbeck C.M., Gold J.R. 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431.

Rabiee M., Mirarab S. 2020. Forcing external constraints on tree inference using ASTRAL. BMC Genomics 21:218.

Rambaut. 2014. FigTree v1. 4.2, a graphical viewer of phylogenetic trees. http://tree.bio.ed.ac.uk/software/figtree/.

Rice A., Glick L., Abadi S., Einhorn M., Kopelman N.M., Salman-Minkov A., Mayzel J., Chay O., Mayrose I. 2015. The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytol. 206:19–26.

Santamaría-Aguilar D., Monro A.K. 2019. Compendium of *Freziera* (Pentaphylacaceae) of South America including eleven new species and the typification of 22 names. Kew Bull. 74:14.

Shen X.-X., Salichos L., Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol 8:2565–2580.

Siu-Ting K., Torres-Sánchez M., San Mauro D., Wilcockson D., Wilkinson M., Pisani D., O'Connell M.J., Creevey C.J. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. Mol. Biol. Evol. 36:1344–1356.

Smith M.L., Hahn M.W. 2022. The frequency and topology of pseudoorthologs. Syst. Biol. 71: 649–659.

Smith M.L., Hahn M.W. 2021. New approaches for inferring phylogenies in the presence of paralogs. Trends Genet. 37:174–187.

Smith S.A., Brown J.W., Walker J.F. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PLoS One 13:e0197433.

Solís-Lemus C., Bastide P., Ané C. 2017. PhyloNetworks: a package for phylogenetic networks. Mol. Biol. Evol. 34:3292–3298.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Štorchová H., Hrdličková R., Chrtek J. Jr, Tetera M., Fitze D., Fehrer J. 2000. An improved method of DNA isolation from plants collected in the field and conserved in saturated NaCl/CTAB solution. Taxon 49:79–84.

Tricou T., Tannier E., de Vienne D.M. 2022. Ghost lineages highly influence the interpretation of introgression tests. Syst. Biol. 71:1147–1158.

Tsou C.-H., Li L., Vijayan K. 2016. The intra-familial relationships of Pentaphylacaceae sl as revealed by DNA sequence analysis. Biochem. Genet. 54:270–282.

Ulloa Ulloa C., Acevedo-Rodríguez P., Beck S., Belgrano M.J., Bernal R., Berry P.E., Brako L., Celis M., Davidse G., Forzza R.C., Gradstein S.R., Hokche O., León B., León-Yánez S., Magill R.E., Neill D.A., Nee M., Raven P.H., Stimmel H., Strong M.T., Villaseñor J.L., Zarucchi J.L., Zuloaga F.O., Jørgensen P.M. 2017. An integrated assessment of the vascular plant species of the Americas. Science 358:1614–1617.

Vargas O.M., Ortiz E.M., Simpson B.B. 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol. 214:1736–1750.

Weitzman A.L., Dressler S., Stevens P.F. 2004. Ternstroemiaceae. In: Kubitzki K., editor. Flowering Plants. Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 450–460.

Yan Z., Smith M.L., Du P., Hahn M.W., Nakhleh L. 2022. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst. Biol. 71:367–381.

Yang Y., Smith S.A. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31:3081–3092.

Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19:153.

Zhang Z., Xie P., Guo Y., Zhou W., Liu E., Yu Y. 2022. Easy353: a tool to get angiosperms353 genes for phylogenomic research. Mol. Biol. Evol. 39:msac261.