Strong phylogenetic signal despite high phylogenomic complexity in an Andean plant radiation (*Freziera,* Pentaphylacaceae)

Laura Frost[1,2,3*], Ana M. Bedoya[1,3], & Laura Lagomarsino[1*]

[1]Shirley C. Tucker Herbarium, Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70808, United States

[2]Biology Department, University of South Alabama, Mobile, AL 36688, United States

[3] These co-first authors contributed equally

*Corresponding author:

Laura Frost

Tel: +1 2565205813

Email: lafrost@southalabama.edu

Laura Lagomarsino

Tel: +1 2255788555

Email: llagomarsino1@lsu.edu

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 to 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 multi-copy 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 non-specific signal of introgression using Patterson's D and f4 statistics. Despite phylogenomic complexity, we were able to resolve *Freziera* into nine 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.

KEYWORDS: 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 results in 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 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).

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

81 Still, a major barrier to phylogenomic inference in many plant clades is the difficulty in
82 efficient identification of orthologous and paralogous sequences from multi-copy loci (Yang and
83 Smith 2014; Morales-Briones et al. 2022). The presence of multiple sequences for a single locus
84 within an individual may result from allelic variation or gene/genome duplication. The latter
85 process has repeatedly taken place during the evolutionary history of plant lineages, and results
86 in the presence of non-orthologous gene copies within the same genome (i.e., paralogs) (Li and
87 Barker 2020). Automated detection of multi-copy genes in phylogenomic datasets could be
88 hindered when sequence data fails to meet contig number, contig length, or read depth thresholds
89 set during assembly, a scenario that is more likely when is DNA obtained from herbarium
90 specimens (Bakker et al. 2015).

91 The presence of only one copy of a multi-copy gene in a sequenced dataset (i.e., hidden
92 paralogy) may be due to biological processes or to data processing artifacts. “Pseudo-orthologs”
93 are a type of hidden paralogy that results from a biological process: differential loss after gene
94 duplication leads to the presence of a single but non-orthologous copy across species in nature
95 (Smith and Hahn 2021a, 2021b). While there is some evidence that coalescent-based

96 phylogenetic methods are relatively robust to pseudo-orthologs (Smith and Hahn 2021a, 2021b),
97 this finding is applicable only under certain conditions that do not include whole genome
98 duplication followed by rediploidization— a phenomenon common in plants (Li et al. 2021).
99 Hidden paralogy can also be artifactual, including when differential assembly of copies in a
100 multi-copy gene results in the recovery of non-orthologous sequences for a given locus across
101 taxa. We refer to these as "artifactual orthologs." While it is well-documented that unrecognized
102 paralogy can have negative impacts on species tree inference (Brown and Thomson 2017), the
103 presence of artifactual orthologs in plant phylogenomic datasets and their impact on phylogenetic
104 inference remains underexplored.

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

117 Using Angiosperms353 (Johnson et al. 2019) target enrichment data derived almost
118 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 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.

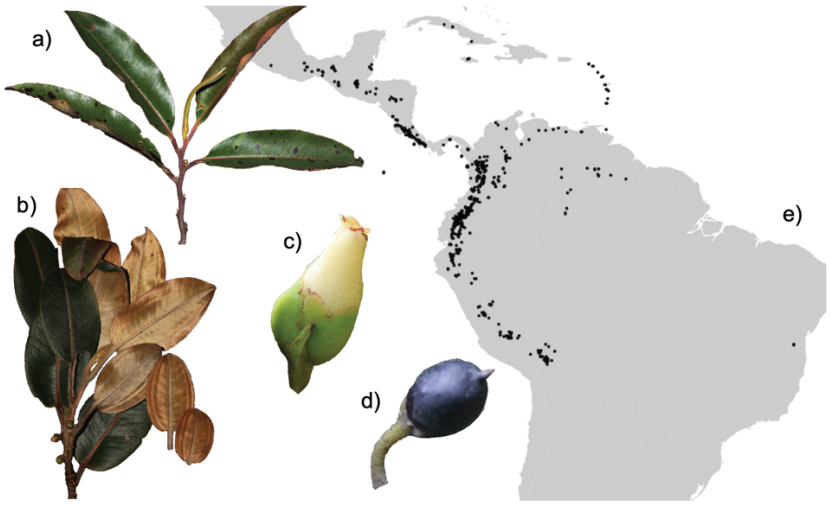

Figure 1. Diversity and distribution of *Freziera*. *Freziera* has significant variation in leaf morphology and pubescence, as illustrated by branches of a) *F. candicans* and b) *Freziera sp.*; meanwhile, c) flower and d) fruit morphology are relatively stable. *Freziera* is an Andean radiation; e) 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.

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 five 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),

139 and 7 species of *Ternstroemia* (belonging to the sister tribe Ternstroemieae; (Weitzman et al.
140 2004; Tsou et al. 2016).
141
142 *DNA extraction, library prep, target enrichment, and sequencing*
143 Detailed descriptions of laboratory methods are provided in online Appendix 1. Briefly,
144 DNA extraction followed a modified sorbitol extraction protocol (Štorchová et al. 2000). Library
145 preparation used the KAPA Hyper Prep and KAPA HiFi HS Library Amplification kits with iTru
146 i5 and i7 dual-indexing primers. Target enrichment was carried out using the MyBaits
147 Angiosperms353 universal probe set (Johnson et al. 2019; Hale et al. 2020). DNA libraries were
148 sequenced by Novogene in one Illumina Hiseq 3000 lane with 150 bp paired-end reads.
149 Although most samples come from herbarium specimens, DNA concentration met minimum
150 standards for sequencing.
151
152 *Raw data processing*
153 Demultiplexed raw sequence reads were trimmed with illumiprocessor v2.0.9 (Faircloth
154 2013, 2016), a wrapper for Trimmomatic v0.39 (Bolger et al. 2014). Default settings were used
155 and reads with a minimum length of 40 bp kept. Trimmed reads were assembled into
156 supercontigs (exons and flanking intronic regions) with HybPiper v1.3.1 (Johnson et al. 2016)
157 using the Angiosperm353 target file as reference (Johnson et al. 2019). In addition, we used a
158 taxon-specific target file for Pentaphylacaceae using Easy353 (Zhang et al. 2022), a reference-
159 guided assembly tool for recovery of Angiosperms353 gene sets. We used the transcriptome of
160 *Ternstroemia gymnanthera* (One Thousand Plant Transcriptomes Initiative 2019) for sequence

161 retrieval with Easy353. Additional details of data processing prior to final gene tree inference are
162 available in online Appendix 1.
163
164 *Gene tree inference*
165 Preliminary gene trees from the HybPiper v1.3.1 supercontig alignments were generated
166 from aligned sequences for the 322 loci lacking paralog flags with RAxML v.8.2.12 (Stamatakis
167 2014) under the GTRCAT nucleotide substitution model with 200 rapid bootstraps. The resulting
168 trees were processed with TreeShrink v1.3.3 (Mai and Mirarab 2018) to detect and remove
169 unusually long branches (i.e., potential cross contaminants). Processing in TreeShrink was
170 performed on a "per-gene" and "all-gene" basis. The identified branches (most of which
171 corresponded to samples with few sequenced reads; Supplementary Table S1) were removed
172 from final alignments. We then removed alignments with fewer than 25 ingroup samples from
173 further analyses. Accessions present in <10% of processed gene trees were trimmed using the R
174 package ape (Paradis and Schliep 2019a). Gene trees were inferred with IQ-TREE multicore
175 v2.1.1 (Nguyen et al. 2015) including model section via ModelFinder (Kalyaanamoorthy et al.
176 2017), tree search, ultrafast bootstraps (Hoang et al. 2018), and Shimodaira-Hasegawa-like
177 approximate likelihood-ratio test (SH-aLRT; (Guindon et al. 2010; Anisimova et al. 2011).
178
179 *Paralog warnings and detection of artifactual orthologs*
180 Despite assembling a single sequence per sample for the majority of loci and raising
181 paralog warnings for <9% loci with HybPiper v1.3.1 (Supplementary Table S1), visual
182 observation of gene trees and alignments suggested the presence of undetected paralog sequences
183 (i.e., artifactual orthologs) in the dataset. Evidence for this included split clades of outgroup

184 species and distinctive motifs in alignments that reflected higher sequence divergence than
185 expected by allelic variation alone (Fig. 2b). Gene trees without paralog warnings were first
186 examined by-eye in FigTree v1.4.3 (Rambaut 2014) and sorted as putative orthologs or
187 artifactual orthologs. Artifactual orthologs are identified when different accessions for the same
188 ingroup species, each represented by only a single sequence in the alignment, cluster into two
189 subclades separated by relatively long internal branches (Fig. 2b). Visual detection of artifactual
190 paralogs from gene trees in phylogenomic datasets thus benefits from sampling multiple
191 individuals per species. We grouped loci with no paralog warnings from HybPiper v1.3.1. and
192 confirmed by-eye as putative orthologs into a gene tree set called *orthologs.by.eye* (Table 1).
193 We compared our by-eye identification of artifactual orthologs with two automated
194 paralogy detection strategies: HybPiper v2.0 (Johnson et al. 2016) and HybPhaser (Nauheimer et
195 al. 2021). Both versions of HybPiper could detect paralogs if multiple assembled contigs each
196 cover ≥75% of a reference sequence; this returns a long paralog warning in HybPiper2.
197 HybPiper2 also raises paralog warnings when multiple contigs are assembled across ≥75% of the
198 reference sequence even when individual contigs are <75% the full sequence length; this returns
199 a paralog-by-depth warning. In addition, HybPiper2 reports stitched contigs (i.e., those derived
200 from multiple SPAdes contigs), which can be used to identify chimeric sequences (i.e., stitched
201 contigs in which the two separate sequences are derived from different gene copies); HybPiper2
202 returns chimera warnings when a sufficient number of read pairs from stitched contigs map to
203 different SPAdes contigs and have sequence mismatch above a given threshold for one read but
204 not the other (see HybPiper2 documentation for additional details;
205 https://github.com/mossmatters/HybPiper/). HybPhaser is an automated strategy to identify
206 hybrids, polyploids, contamination, and paralog sequences by identifying samples with a high

207 proportion of heterozygous loci and allele divergence, and loci with a high proportion of SNPs
208 (Nauheimer et al. 2021).

209 In addition, we applied a method for automated orthology detection (i.e., gene tree
210 pruning) using homologs with monophyletic outgroup (MO; Table 1; Yang and Smith 2014;
211 Morales-Briones et al. 2021a). MO identifies clusters in gene trees with monophyletic outgroups
212 and searches from root to tips for duplications. When duplicated taxa are found on either side of
213 a bifurcation, the subtree with the fewest ingroup taxa is pruned. MO does not directly identify
214 artifactual orthologs but has the potential to remove their effect by selecting orthogroups and
215 allowing the inclusion of more loci in a final dataset.

216 We generated five sets of gene trees to compare artifactual orthology identification using
217 the above automated pipelines; a description of filtering criteria and thresholds for each dataset
218 are in Table 1. Loci that received a paralog warning in HybPiper v1.3.1 were excluded to
219 generate the *orthologs.unfiltered* dataset. The remaining filtered datasets (i.e.,
220 *orthologs.HybPiper2.long*, *orthologs.HybPiper2.no.warnings, HybPhaser,* and *MO*) are subsets
221 of the *orthologs.unfiltered* dataset. For the *HybPhaser* dataset, we did not filter loci by missing
222 data (we kept loci recovered in at least 0.01 proportion of taxa) to ensure that only loci with
223 excess heterozygosity (loci with more than 1.5× the interquartile range above the third quartile),
224 as expected in putative paralogs, were removed. We kept ortholog groups with ≥25 ingroup taxa
225 in the *MO* dataset for comparison with the other filtering strategies.

226

227 *Impact of standard gene tree filtering strategies on the presence of artifactual orthologs*

228 We explored how standard gene tree filtering affects the presence of artifactual orthologs
229 (as identified by-eye) in our dataset. Gene tree filtering is typically applied to minimize gene tree

230 estimation error, gene tree discordance (Molloy and Warnow 2018), and for phylogenomic
231 subsampling (Mongiardino Koch 2021) but applying these metrics may also favor the selection
232 of artifactual orthologs in a dataset (Fig. 2d,e). Gene trees were filtered using five common
233 empirical criteria for phylogenomic subsampling (Table 1): one alignment-based metric
234 (*proportion.PI*), two tree length metrics (*tree.length* and *proportion.internal.branch.lengths*), and
235 two gene tree support metrics (*average.BS* and *bipartition*). Due to the presence of non-
236 orthologous copies in alignments including artifactual orthologs, a high degree of genetic
237 variation is expected relative to true orthologs, increasing the proportion of informative sites and,
238 because of well-supported clusters of non-orthologous sequences, we expected an elevated
239 average bootstrap support in gene trees (Fig. 2d). Filtering by bootstrap support is standard
240 practice, collapsing low supported nodes in gene trees prior to species tree inference to remove
241 gene tree error (Zhang et al. 2018). The two tree length criteria were selected because, while a
242 high percentage of internal branch length can signal high phylogenetic signal (Shen et al. 2016),
243 it can also indicate biological pseudo-orthologs (Smith and Hahn, 2016b) and likely, artifactual
244 orthologs (Fig. 2). Finally, filtering by bipartition support may remove artifactual orthologs by
245 favoring loci that are more concordant with the inferred species tree (Smith et al. 2018; Fig. 2e).
246 Standard gene tree filtering strategies were applied to the *orthologs.unfiltered* dataset.
247 The *proportion.PI* was obtained using AMAS (Borowiec 2016) with outgroups removed.
248 Estimates of *proportion.internal.branch.lengths* were extracted from IQ-TREE '*.iqtree' output
249 files. Bootstrap support values were extracted from IQ-TREE maximum likelihood tree files and
250 averaged to calculate *average.BS*. For *tree.length*, input gene trees were rooted using pxrr in
251 phyx (Brown et al. 2017) and tree length was calculated with the script 'get_var_length.py' of
252 SortaDate (Smith et al. 2018) excluding outgroups. *Bipartition* support was calculated with

253 SortaDate (Smith et al. 2018)'get_bp_genetrees.py' script(Smith et al. 2018) against a species
254 tree generated with ASTRAL-III (Zhang et al. 2018) from the *orthologs.unfiltered* dataset.
255 Cutoff thresholds for each filter were selected near the average value in our dataset to keep a
256 similar number of loci for each criterion.

257

258 **Table 1**. Descriptions of the filtering criteria used, names applied to each of the datasets, number of loci removed
259 and selected by each, and the impact of filtering criteria on gene tree discordance, branch support, and species tree
260 topology. Values for datasets that performed equally or better than the *orthologs.unfiltered* dataset for each metric
261 are bolded. Major topological conflicts between species trees and a consensus topology of the 11 filtered datasets
262 (Fig. 3b) are shown, including instances where the species tree is compatible (+) or in disagreement (-) with the
263 consensus topology. Datasets from which the Elaphoglossifolia group was not resolved as monophyletic are
264 indicated with "n/a". Asterisks mark disagreement from the alternative placement of only one species. The
265 "Branching order of the Elaphoglossifolia group" includes conflict in the placement of *F. magnibracteolata.*

Description of filtering criteria (for keeping loci)	Putative orthologs dataset name after filtering	No. of loci removed by filtering criteria	No. of loci that passed filtering criteria	Normalized Quartet Score	Average branch support	Proportion of highly supported branches (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*	31	313	0.567	0.774	0.390	+	–	n/a	+
loci identified as putative orthologs after by-eye inspection of gene trees and alignments	*orthologs.by.eye*	131 (90)	182	**0.634**	**0.792**	0.377	+	+	+	+
loci with 1 contig at ≥75% reference sequence (no long paralog warnings from HybPiper v.2.0)	*orthologs.HybPiper2.long*	51	262	**0.582**	**0.780**	0.377	+	+	+	+
loci without paralog warnings by depth or length from HybPiper v.2.0	*orthologs.HybPiper2.no.warnings*	126	187	**0.619**	**0.783**	0.351	+	+	+	+
loci that have a lower proportion of SNPs than 1.5x the interquartile range above the 3rd quartile for any given sample	*HybPhaser*	71	242	**0.597**	**0.783**	0.364	+	+	+	+
orthology inference via gene tree pruning using homologs with monophyletic, non-repeating outgroups	*MO*	91	222	**0.662**	0.664	0.221	+*	+*	–	–
alignment contains ≥ 7.5% parsimony informative sites in ingroup (average value across all alignments)	*proportion.PI*	153	160	0.528	0.715	0.247	–	–	n/a	+
gene trees with internal branch lengths comprising an above average proportion of the total tree length across all gene trees	*proportion.internal.branch.lengths*	164	148	0.542	0.685	0.234	–	+	+	+
gene trees with average bootstrap support across all nodes above the average value across all gene trees	*average.BS*	150	163	0.560	0.727	0.234	+	+	–	+
gene tree with total tree length above the average value across all gene trees	*tree.length*	173	139	0.524	0.688	0.260	–	–	n/a	–
gene trees with above average bipartition support relative to the species tree inferred from the orthologs.unfiltered dataset	*bipartition*	147	166	**0.601**	**0.792**	**0.390**	+	+	+	+

266

267

268 *Phylogenetic relationships and introgression in Freziera*

269 *Impact of filtering criteria on species tree inference.*—Species trees were inferred in
270 ASTRAL-III from the 11 datasets listed in Table 1. A majority-rule consensus tree was
271 generated from the species trees for the 11 filtered datasets (*consensus* function of the R package
272 *ape*; Paradis and Schliep 2019b) to identify major clades that were concordant across datasets.
273 Gene tree discordance was assessed for all species tree topologies using the final normalized
274 quartet score (NQS; proportion of quartet trees in gene trees that are present in the species tree

275 (Mirarab et al. 2014b) estimated with ASTRAL-III. Impact of filtering criteria on branch support
276 was assessed by calculating the average branch support and proportion of well-supported
277 branches (ppl≥0.97; Rabiee and Mirarab 2020) across the 11 resulting species trees. We inferred
278 a second consensus tree from the species trees generated from the datasets that showed the
279 greatest improvements to both normalized quartet score and average branch support (*bipartition,*
280 *by.eye, HybPhaser,* and *orthologs.HybPiper2.no.warnings*), which also had substantial
281 reductions in the number of artifactual orthologs (Fig. 2c). The resulting tree was compared to
282 the *orthologs.unfiltered* dataset to explore the impact that the removal of paralogs, including
283 artifactual orthologs, has on species tree inference.

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

297 three datasets were specified in Dsuite so that the D and f4 estimated values were arranged
298 according to the tree.

299 The relatively small number of loci in our dataset limited the power to detect introgression
300 along a phylogeny. The f-branch statistic, which accounts for the correlation of D and f4 due to
301 shared ancestry among multiple potential introgression donor species (Malinsky et al. 2018),
302 allows for a better interpretation of introgression patterns across a tree (Malinsky et al. 2021).
303 However, simulation analyses have shown that for a relatively small number of unlinked SNPs
304 (<10,000), the proportion of cases where the strongest inferred f-branch signal corresponds to the
305 correct simulated gene flow recipient and donor branches is <20% (Malinsky et al. 2021). Due to
306 the limited number of unlinked SNPs in our targeted sequence capture dataset, we did not apply
307 this metric.

308

309 **RESULTS**

310 *Sequence capture efficiency*

311 Voucher information for accessions and per sample data for read trimming and contig
312 assembly are available in Supplementary Table S1. Of the 102 accessions for which target
313 enrichment libraries were prepared and sequenced, 22 (21 *Freziera* and 1 outgroup) were
314 excluded either because no sequences were assembled for any of the 353 target loci, or because
315 they were identified to have exceedingly long branches with TreeShrink (most of the latter
316 corresponded to samples with few sequenced reads; Supplementary Table S1). Target
317 enrichment efficiency across all sequenced samples is shown in Supplementary Figure S1.
318 Seventy-two accessions representing 50 of the 75 species of *Freziera* and 8 members of the
319 outgroup remained (see methods; Supplementary Table S1), which had assembled sequences for
320 52–348 genes (average: ~296). Standard assembly statistics from both versions of HybPiper are

321 in Supplementary Tables S1-S2. No loci were flagged by HybPiper2 as putatively chimeric.
322 However, all but 13 loci included at least one sequence with stitched contigs (Supplementary
323 Table S3). All sequences without stitched contigs correspond to regions spanning a single exon.
324 Off-target data collection was insufficient to assemble plastomes.
325
326 *Paralog warnings and detection of artifactual orthologs*
327 The 11 paralogy detection and gene tree filtering criteria that we applied generated
328 datasets with 140–313 loci (Table 1). Results of artifactual ortholog detection by-eye and from
329 automated strategies are shown in Fig. 2a and Supplementary Table S3. HybPiper v.1.3.1 raised
330 paralog warnings for 31 loci (i.e., >1 contig was assembled for at least one sample for a given
331 locus). Of the 322 loci that were not flagged as paralogs, nine were removed because they
332 contained fewer than 25 ingroup samples, leaving 313 with a single assembled contig per
333 species. By-eye inspection of these 313 loci resulted in the identification of 90 artifactual
334 orthologs. Loci flagged as potentially paralogous in HybPiper v.1.3.1 also had long paralog
335 warnings and paralog warnings by depth with HybPiper2. Of the 313 loci without paralog
336 warnings from HybPiper v.1.31, 51 had long paralog warnings, of which 46 also received
337 paralog warnings by contig depth with HybPiper2. An additional 75 loci were not flagged as
338 long paralogs but did receive paralog warning by contig depth for at least one sample, totaling
339 121 loci with warnings by contig depth and 126 loci with any kind of paralog warning issued by
340 HybPiper2. HybPhaser identified 73 loci as putative paralogs (Supplementary Table S3; online
341 Appendix 2), most of which also had paralog warnings by depth. There was substantial overlap
342 of loci identified by these methods (Fig. 2c). Three of the methods we compared— HybPiper2
343 paralog-by-depth warnings, HybPiper2 paralog-by-length warnings, and HybPhaser— flagged

344 70 (78%) of the loci that we identified as a artifactual orthologs in our by-eye assessment (Fig.
345 2c). Of these, HybPiper2 paralog-by-depth warnings had the greatest overlap with our by-eye
346 assessment (Fig. 2c; Supplementary Table S3). Assembly with a taxon-specific target file
347 resulted in similarly low levels of paralog detection as the universal target file: <8% of
348 assembled loci recovered >1 long contig for at least one sample (Supplementary Table S4).
349 There was a correlation between the number of reads mapped and the number of paralog
350 warnings from HybPiper in a sample (Supplementary Fig. S3).
351
352 *Impact of standard gene tree filtering strategies on the presence of artifactual orthologs*
353 Summary statistics for alignments and gene trees are available in Supplementary Table
354 S3. The impact of gene tree filtering strategies on removal of artifactual and putative orthologs
355 (as identified by-eye) is shown in Fig. 2e. Putative orthologs are removed with all filtering
356 strategies, and only filtering gene trees by bipartition support resulted in the removal of most
357 artifactual orthologs. As commonly applied, the remaining four gene tree filtering criteria all
358 resulted in datasets in which there were either similar proportions of putative orthologs and
359 artifactual orthologs in the final datasets (*average.BS, proportion.internal.branch.length*), or that
360 favored the inclusion of artifactual orthologs (*tree.length*, *proportion.PI*).

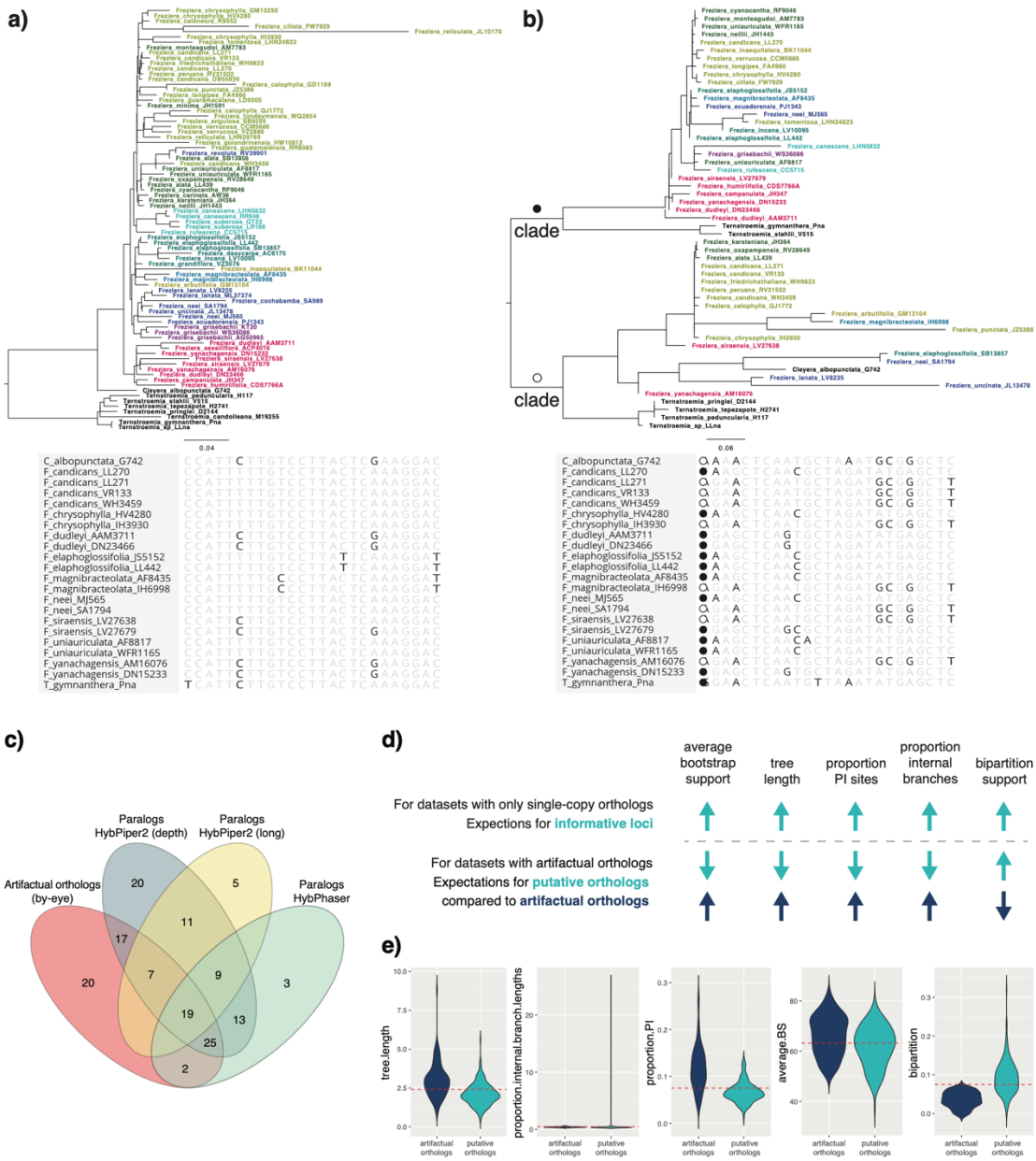

Figure 2. Exemplary gene trees and alignment subsets for a) a putative ortholog, and b) an artifactual ortholog. Putative ortholog gene trees exhibit monophyletic genera for outgroups and appropriate relationships to the ingroup, a relatively shallow backbone in the ingroup, and relationships similar 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 warnings raised by automated paralogy detection strategies. d) Expectations for informative, single-copy orthologs change for many gene tree filtering metrics when data sets include artifactual orthologs; arrows indicate whether expected values for metrics are high or low. e) Impact of gene tree filtering strategies on removal of artifactual and putative orthologs as identified by-eye.

375 The dashed lines indicate the various thresholds above which gene trees were kept in filtered datasets, as indicated in
376 Table 1.
377

378 *Impact of data curation on phylogenetic performance*

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

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

399

400 *Phylogenetic relationships and introgression in Freziera*

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

417 Regions of lowest support in the consensus tree tended to be in conflict between species
418 trees and the consensus topology. These regions often involved two notable sets of taxa: the
419 Elaphoglossifolia group, comprising the Incana and Canescens clades and *F. magnibracteolata*,
420 and the Candicans group, which comprises the Calophylla and Karsteniana clades (Table 1, Fig.
421 3; Supplementary Fig. S4). While the Elaphoglossifolia group was monophyletic in most
422 datasets, including all in which the proportion of artifactual orthologs was reduced (Table 1), the

unstable placement of *F. magnibracteolata* rendered it non-monophyletic three 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 across analyses (Table 1). While 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 three 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 (LH) and >1% allele divergence (AD) 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.

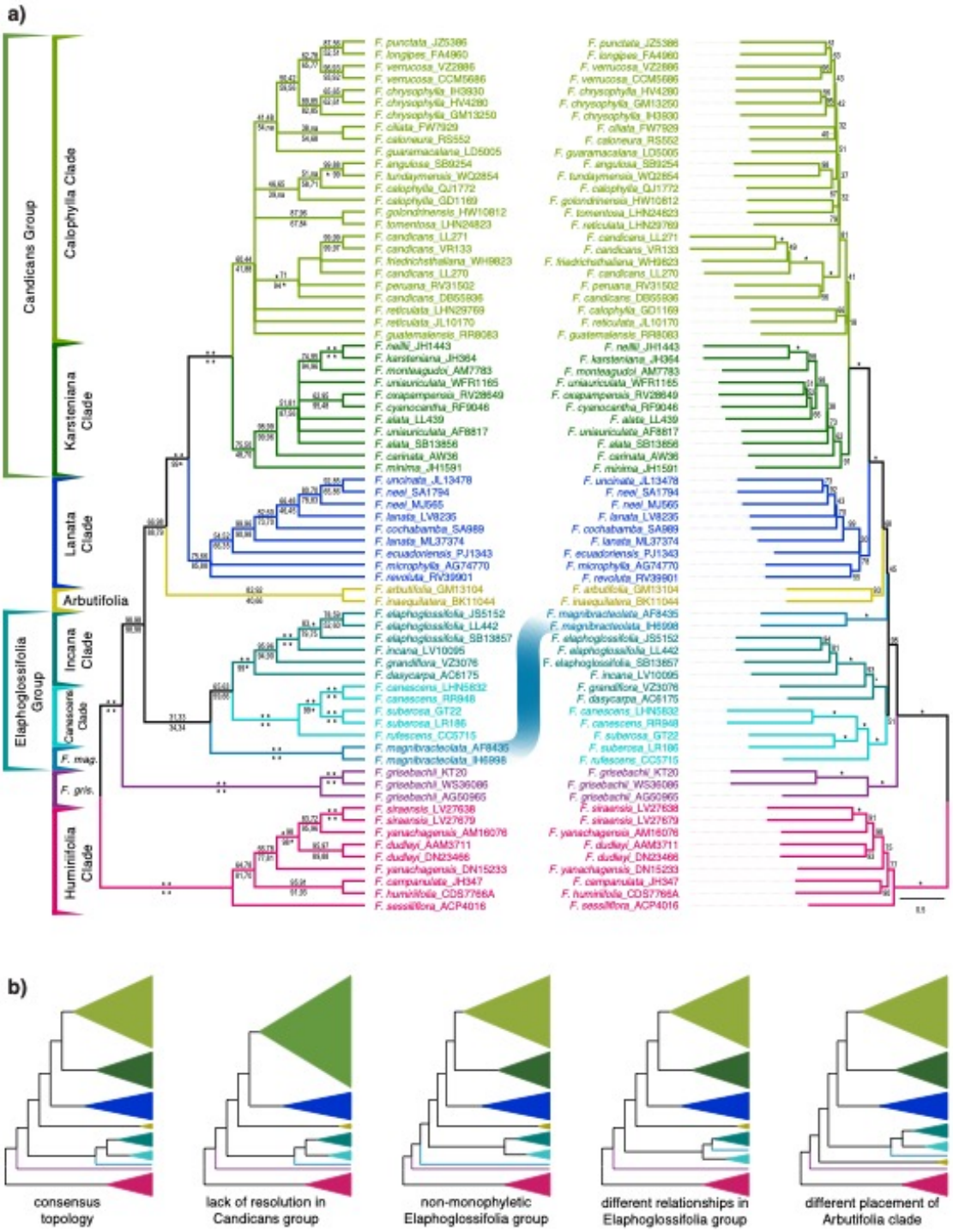

Figure 3. Names of clades consistently recovered and species tree topologies for a) The consensus tree of the four datasets in which the proportion of artifactual orthologs was reduced (*bipartition*, *by.eye*, *HybPhaser*, *orthologs.HybPiper.no.warnings*; left) versus the orthologs.unfiltered species tree (right). The placement of *F. magnibracteolata*– the major topological conflict between the two trees– is highlighted with the blue line connecting the tip labels in each. Branches with full support (LPP from ASTRAL = 1) are indicated by an asterisk; elsewhere, the LPP×100 is provided. Support for each of the four datasets in the consensus tree are 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.

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 multi-copy 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 multi-copy 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

477 polyphyly of species and deep divergences between clades, including outgroup taxa and
478 members of those spuriously polyphyletic species (Fig. 2b). As troubling as the prospect of
479 unfiltered paralogs may be for empiricists, artifactual orthologs may be easier to detect than
480 biological pseudo-orthologs (Smith and Hahn 2021b) because of these striking, easily
481 identifiable patterns.

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

493 Our dataset relied heavily on degraded DNA extracted from herbarium tissue, which we
494 believe contributes to the assembly of artifactual orthologs. Though high-molecular weight DNA
495 is fragmented during library preparation, the desired fragment size for Illumina libraries is 200-
496 500 base pairs (bp; Bronner and Quail 2019). As is common when working with specimens from
497 the wet tropics (Bakker et al. 2015), many of our samples had a high proportion of fragments
498 <200 bp (unpublished). These short fragments restrict data collection in non-coding regions, as
499 fragments are less likely to span the regions flanking targeted exons. Using HybPiper v2.0.1, we

500 identified a high proportion of stitched contigs—separate contigs concatenated into a single
501 sequence. The assembly of multiple contigs belonging to different paralog copies into a single
502 sequence (i.e., chimeric sequences) is an artifact that is possible in the presence of stitched
503 contigs. The preferential selection of the longest contig with the highest coverage for a given
504 gene region during assembly further contributes to decreased paralog detection and increased
505 chimeric assembly. Despite the high proportion of stitched contigs, no genes were flagged by
506 HybPiper v2.0.1 with a chimera warning for any sample (Supplementary Table S2). This is
507 likely due to current limitations and stringency in chimeric sequence detection. Specifically,
508 paired ends must completely map to separate contigs and reads must map entirely within an exon
509 on the separate contigs to flag a chimera warning. Short fragments from degraded DNA reduce
510 the likelihood of finding read pairs that will meet these criteria, and, therefore, the likelihood of
511 detecting chimeric sequences, though they may be present in the dataset.

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

521 We believe the lack of data spanning introns, resulting in a high proportion of stitched
522 contigs, and the potential for greater disparity between capture of both copies from herbarium

523 DNA, is a primary source of artifactual orthologs. Low sequencing depth resulting in poor
524 coverage of non-coding regions is another factor that can increase the proportion of stitched
525 contigs and/or differential success capturing copies. However, sequencing depth is unlikely to be
526 the main source of artifactual orthologs in our dataset as the number of reads mapped to target
527 loci (Supplementary Table S1) in our dataset is higher per sample than studies that have
528 successfully examined paralogs in Angiosperms353 datasets (Johnson et al. 2019; Gardner et al.
529 2020). Taken together, this highlights caveats that remain with museomic data: long-read
530 sequence data is not an option in many cases and deeper sequencing cannot recover missing data.
531 Herbarium specimens are an invaluable resource for improving sampling and filling gaps in the
532 tree of life, and we do not advocate for the exclusion of herbarium specimens in phylogenomic
533 datasets. Rather, we recommend careful assessment of datasets considering these artifactual
534 complications in assembly.
535
536 *Impact of standard gene tree filtering strategies in the presence of artifactual orthologs*
537 Gene tree filtering is a commonly applied strategy to minimize gene tree estimation error
538 (GTEE) and its impact on species tree estimation (Molloy and Warnow 2018). Without an
539 explicit method to estimate GTEE in empirical data, multiple criteria including alignment length,
540 the number/proportion of variable or parsimony informative sites, total tree length, the
541 proportion of internal branch lengths, and average bootstrap support of gene trees have all been
542 used as implicit proxies to account for GTEE (Leaché et al. 2014; Liu et al. 2015; Shen et al.
543 2016; Blom et al. 2017). It is argued that these metrics select for "higher quality" gene trees,
544 however, this assumption is violated in the presence of artifactual orthologs, which are
545 associated with higher values of many standard metrics (Fig. 2c).

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

556 Both artifactual orthologs and putative orthologs were at removed similar levels from
557 datasets by the remaining four gene tree filtering criteria, in particular *tree.length* and
558 *proportion.PI.* This is consistent with our predictions about characteristics of artifactual
559 orthologs (Fig. 2d,e) and demonstrates that filtering by alignment or gene tree characteristics
560 using standard techniques may actually result in a relative increase of data artifacts in complex
561 phylogenomic datasets.

562 While gene tree filtering metrics have the potential to significantly improve phylogenetic
563 support (Doyle et al. 2015) and clarify relationships, it may be an inappropriate strategy for some
564 studies (Molloy and Warnow 2018). This is true in this case for *Freziera*, in which gene tree
565 filtering resulted in lower branch support and spurious topological inference for all gene tree
566 filtering criteria except bipartition support (Table 1, Supplementary Fig. S2). This is likely the
567 result of removal of a large proportion of single-copy loci relative to artifactual orthologs (Fig.
568 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 two-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 eleven 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.*

592 *magnibracteolata*, a relationship that was not present in the unfiltered dataset (Table 1).
593 Improvements in these datasets are likely the result of the reduction in gene tree heterogeneity
594 with the removal of the noise introduced by artifactual orthologs, though gene tree heterogeneity
595 persists in curated datasets and likely reflects biological processes that have shaped *Freziera*'s
596 evolutionary history (Table 1). Ensuring that paralogs, including artifactual orthologs, are
597 appropriately handled in phylogenetic analyses is not only essential for accurate phylogenetic
598 estimation, it is also crucial to downstream analyses that require accurate branch lengths,
599 including divergence date estimation (Siu-Ting et al. 2019).

600 Orthology inference using monophyletic outgroup (*MO*) was not associated with
601 improved phylogenetic performance, despite being successfully applied in datasets where
602 paralogs are represented as multiple copy loci in assembled datasets (Morales-Briones et al.
603 2022). The *MO* species tree had reduced gene tree discordance (i.e., higher normalized quartet
604 sampling score) relative to the unfiltered dataset, likely from the removal of non-orthologous
605 copies. However, it failed to recover some of the major clades identified from curated data where
606 artifactual orthologs were removed. This may be due to information loss: the high number of
607 multicopy loci in our dataset resulted in many trees being extensively pruned by MO, resulting in
608 smaller subtrees with fewer taxa. MO is a valuable tool to identify orthologous clusters in
609 complex phylogenomic datasets with multi-copy paralogs and can increase phylogenetic
610 resolution in groups with a history of polyploidy (Morales-Briones et al. 2022); however, it
611 cannot be used to identify artifactual orthologs and did not improve phylogenetic performance in
612 *Freziera*, where paralogs were predominantly single-copy.
613

614 *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 to 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 moderately 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. While our phylogenetic backbone is generally well-supported, there are three non-mutually 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 non-bifurcating 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

638 Calophylla clade, carries the hallmarks of ILS due to rapid radiation, including conflicting
639 species relationships with short branch lengths between close relatives (Fig 3a; Supplementary
640 Fig. S4). Not only is gene tree-species tree discordance higher in this clade compared to other
641 regions of *Freziera*'s phylogeny, but species relationships also differ across analyses
642 (Supplementary Fig. S4). Further indication of ILS in this clade is the fact that the Calophylla
643 clade is the most widespread of *Freziera*'s subclades and includes species with some of the
644 broadest distributions in the genus. These distributional patterns allow a greater possibility that
645 ancestral populations were large and widespread, contributing to ILS, amplifying the effect of
646 short times between speciation events in this group.

647 Even after applying data curation that may inflate support for bifurcating relationships, a
648 strong signal of introgression was identifiable in *Freziera*. This was evidenced by multiple
649 significant f4 and D statistics (Supplementary Fig. S5). Due to the limited number of unlinked
650 SNPs in our target capture data and the correlation of D and f4 introgression statistics when trios
651 share internal branches, we were not able to isolate the exact branches along which introgression
652 has occurred— a challenge even among the most complete datasets (Tricou et al. 2022).
653 However, our resulting phylogenetic framework provides a robust starting point from which to
654 understand the non-bifurcating nature of *Freziera*'s diversification. Areas of conflict across
655 species trees in our curated datasets are strong candidates for lineages that have been directly
656 shaped by past introgression (Fig. 3, Supplementary Fig. S4). A species that is particularly
657 promising for future investigation is *F. magnibracteolata*, whose unstable placement along the
658 backbone of *Freziera* (Table 1. Fig. 3; Supplementary Figs. S2,S4) suggests a potential history of
659 introgression (MacGuigan and Near 2019; Cai et al. 2021). Notably, the placement of this
660 species was the only major relationship to be impacted by removing artifactual orthologs relative

661 to the unfiltered dataset (Table 1, *Monophyletic Elaphoglossifolia group*). To further assess the
662 extent of gene flow and identify the branches involved in introgression events in this rapid
663 radiation, future research will target a much larger portion of the genome (Malinsky et al. 2021)
664 and include deeper taxon with multiple individuals per species.

665 Finally, either gene or genome duplication has resulted in paralogs in *Freziera*. We
666 identify both paralogs for which multiple copies are identifiable during assembly and artifactual
667 orthologs that are represented by only a single copy in our dataset (Fig. 2b, Supplementary Table
668 S3). While the processes that gave rise to these paralogs are yet to be examined in detail, it is
669 likely that they are product of allopolyploidy, especially considering the extensive history of
670 genome duplication via polyploidy in Ericales, the order to which *Freziera* belongs (Larson et al.
671 2020), as well as chromosome count variation within Pentaphylacaceae (Rice et al. 2015).

672

673 **CONCLUSION**

674 A major current challenge in phylogenomics is the difficulty in teasing apart specific
675 sources of gene tree discordance in empirical datasets, and to account for these in phylogenetic
676 inference (Morales-Briones et al. 2022; Tricou et al. 2022). It is a significant challenge to
677 accurately identify paralogs, pinpoint specific instances of introgression, disentangle incomplete
678 lineage sorting from historical gene flow, and reduce the impact of gene tree estimation error in a
679 single empirical phylogenomic dataset in which all of these sources of discordance are present.
680 In addition to these challenges, phylogenomic data have the potential to be very complex,
681 particularly for clades that are well-understood to be recalcitrant like Andean plant radiations
682 (Pease et al. 2016; Vargas et al. 2017). Here, we showed that careful data curation allowed us to
683 detect a high proportion of artifactual orthologs, which we were able to reduce with multiple,

684 non-mutually exclusive methods: heeding paralog warnings, removing gene trees with a high
685 proportion of heterozygous sites, and filtering gene trees using bipartition support. These data
686 curation strategies were subsequently associated with higher support, lower gene tree conflict,
687 and a more stable species tree— the first for an understudied tropical plant clade that previously
688 lacked any phylogenetic information. We advocate for the observation of empirical
689 phylogenomic data, including gene tree alignments and topologies, and that data curation be
690 tailored to unique properties of individual datasets to better address the above-mentioned
691 complexities in phylogenetic inference.

692 While commonly used filtering techniques, assembly parameters, and other automated
693 aspects of phylogenomics are powerful tools to improving phylogenetic inference, we have
694 shown that they can also increase the proportion of data artifacts (i.e., artifactual paralogs) and
695 have negative impacts on phylogenetic support and inference. Automated filtering techniques are
696 not a replacement for a deep understanding of a dataset. Selecting filtering strategies for
697 individual datasets should be informed by the latter since the decision is likely to be a balance
698 between minimizing the presence of data artifacts while maximizing the number of loci useful
699 for phylogenetic inference. While targeted sequence capture of universal loci offers potential,
700 especially for phylogenetic studies relying heavily on DNA from natural history collections,
701 these datasets are not without limitations related to the nature of the data themselves and to the
702 algorithms we use to process and analyze them. Combining exploration of datasets with deep
703 knowledge of the organismal biology of targeted clades is crucial toward overcoming these
704 limitations and inferring robust phylogenetic hypotheses.
705
706

DATA AVAILABILITY

All data generated in this study are publicly accessible. Raw sequence reads are deposited on NCBI's Sequence Read Archive: (*web address added upon acceptance*). Supplementary materials cited in this paper and processed data are available from the Dryad Digital Repository (*https://datadryad.org/stash/share/X6HXB23fQmPqODdRRdXNJnGePrw29vUPbm1JgIijjFo*).

ACKNOWLEDGEMENTS

This research was funded by a Louisiana Board of Regents Research Competitiveness Subprogram grant, the LSU College of Science and Office of Research and Economic Development, and NSF Award DEB-2055525. We would like to thank the Missouri Botanical Garden (MO) for their 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.

725 REFERENCES

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

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

733 Benjamini Y., Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and
734 Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B Stat. Methodol. 57:289–
735 300.

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

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

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

743 Bronner I.F., Quail, M.A. 2019. Best practices for Illumina library preparation. *Curr. Prot. Hum.*
744 *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-

765 sampled study of the breadfruit genus (Artocarpus, Moraceae). Syst. Biol. 70:558–575.

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

769 Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New
770 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
771 performance of PhyML 3.0. Syst. Biol. 59:307–321.

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

775 Hughes C.E. 2016. The tropical Andean plant diversity powerhouse. New Phytol. 210:1152–
776 1154.

777 Hughes C.E., Eastwood R. 2006. Island radiation on a continental scale: exceptional rates of
778 plant diversification after uplift of the Andes. Proc. Natl. Acad. Sci. U. S. A. 103:10334–
779 10339.

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

783 Johnson M.G., Pokorny L., Dodsworth S., Botigué L.R., Cowan R.S., Devault A., Eiserhardt
784 W.L., Epitawalage N., Forest F., Kim J.T., Leebens-Mack J.H., Leitch I.J., Maurin O., Soltis
785 D.E., Soltis P.S., Wong G.K.-S., Baker W.J., Wickett N.J. 2019. A universal probe set for

786 targeted sequencing of 353 nuclear genes from any flowering plant designed using k-
787 Medoids clustering. Syst. Biol. 68:594–606.

788 Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017.
789 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods.
790 14:587–589.

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

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

796 Larson D.A., Walker J.F., Vargas O.M., Smith S.A. 2020. A consensus phylogenomic approach
797 highlights paleopolyploid and rapid radiation in the history of Ericales. Am. J. Bot.
798 107:773–789.

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

801 Leaché A.D., Wagner P., Linkem C.W., Böhme W., Papenfuss T.J., Chong R.A., Lavin B.R.,
802 Bauer A.M., Nielsen S.V., Greenbaum E., Rödel M.-O., Schmitz A., LeBreton M., Ineich I.,
803 Chirio L., Ofori-Boateng C., Eniang E.A., Baha El Din S., Lemmon A.R., Burbrink F.T.
804 2014. A hybrid phylogenetic–phylogenomic approach for species tree estimation in African
805 *Agama* lizards with applications to biogeography, character evolution, and diversification.
806 Mol. Phylogenet. Evol. 79:215–230.

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

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

812 Li Z., McKibben M.T.W., Finch G.S., Blischak P.D., Sutherland B.L., Barker M.S. 2021.
813 Patterns and Processes of Diploidization in Land Plants. Annu. Rev. Plant Biol. 72:387–410.

814 MacGuigan D.J., Near T.J. 2019. Phylogenomic Signatures of Ancient Introgression in a Rogue
815 Lineage of Darters (Teleostei: Percidae). Syst. Biol. 68:329–346.

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

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

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

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

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

827 Mirarab S., Bayzid M.S., Warnow T. 2014a. Evaluating Summary Methods for Multilocus
828 Species Tree Estimation in the Presence of Incomplete Lineage Sorting. Syst. Biol. 65:366–
829 380.

830 Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014b.
831 ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 30:i541–
832 8.

833 Molloy E.K., Warnow T. 2018. To Include or Not to Include: The Impact of Gene Filtering on
834 Species Tree Estimation Methods. Systematic Biology. 67:285–303.

835 Mongiardino Koch N. 2021. Phylogenomic Subsampling and the Search for Phylogenetically
836 Reliable Loci. Mol. Biol. Evol. 38:4025–4038.

837 Morales-Briones D.F., Gehrke B., Huang C.-H., Liston A., Ma H., Marx H.E., Tank D.C., Yang
838 Y. 2021a. Analysis of paralogs in target enrichment data pinpoints multiple ancient
839 polyploidy events in Alchemilla s.L. (Rosaceae). Syst. Biol. 71:190–207.

840 Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F.,
841 Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021b. Disentangling Sources of Gene
842 Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in
843 Amaranthaceae s.l. Syst. Biol. 70:219–235.

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

847 Murillo-A J., Valencia-D J., Orozco C. I., Parra-O C., Neubig, K. M. 2022. Incomplete lineage

848 sorting and reticulate evolution mask species relationships in Brunelliaceae, an Andean
849 family with rapid, recent diversification. *Am. J. Bot.* 109:1139-1156.

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

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

855 Nguyen L.-T., Schmidt H.A., von Haeseler A., Minh B.Q. 2015. IQ-TREE: a fast and effective
856 stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.
857 32:268–274.

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

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

862 Paradis E., Schliep K. 2019a. ape 5.0: an environment for modern phylogenetics and
863 evolutionary analyses in R. Bioinformatics. 35:526–528.

864 Paradis E., Schliep K. 2019b. ape 5.0: an environment for modern phylogenetics and
865 evolutionary analyses in R. Bioinformatics. 35:526–528.

866 Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources of
867 Adaptive Variation during a Rapid Radiation. PLoS Biol. 14:e1002379.

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

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

872 Rambaut. 2014. FigTree v1. 4.2, a graphical viewer of phylogenetic trees. //tree. bio. ed. ac.
873 uk/software/figtree/right angle bracket.

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

877 Santamaría-Aguilar D., Monro A.K. 2019. Compendium of *Freziera* (Pentaphylacaceae) of
878 South America including eleven new species and the typification of 22 names. Kew Bull.
879 74:14.

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

883 Siu-Ting K., Torres-Sánchez M., San Mauro D., Wilcockson D., Wilkinson M., Pisani D.,
884 O'Connell M.J., Creevey C.J. 2019. Inadvertent Paralog Inclusion Drives Artifactual
885 Topologies and Timetree Estimates in Phylogenomics. Mol. Biol. Evol. 36:1344–1356.

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

Smith M.L., Hahn M.W. 2021b. The frequency and topology of pseudoorthologs. bioRxiv.:2021.02.17.431499.

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 s.l. 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.

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

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

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

917 Yan Z., Smith M.L., Du P., Hahn M.W., Nakhleh L. 2022. Species Tree Inference Methods
918 Intended to Deal with Incomplete Lineage Sorting Are Robust to the Presence of Paralogs.
919 Syst. Biol. 71:367–381.

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

922 Zhang Z., Xie P., Guo Y., Zhou W., Liu E., Yu Y. 2022. Easy353: A Tool to Get
923 Angiosperms353 Genes for Phylogenomic Research. Mol. Biol. Evol. 39.

924

925 FIGURE AND TABLE CAPTIONS

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

932
933 FIGURE 2**.** Exemplary gene trees and alignment subsets for a) a putative ortholog, and b) an
934 artifactual ortholog. Putative ortholog gene trees exhibit monophyletic genera for outgroups and
935 appropriate relationships to the ingroup, a relatively shallow backbone in the ingroup, and
936 relationships similar to the preliminary species tree. A relatively low level of variation is present
937 in alignments for putative orthologs, which is fitting of a locus in a universal probe set being
938 applied within a genus. Artifactual ortholog gene trees often have deep divergences between
939 subsets of the ingroup, including between samples from the same species. Within those
940 subclades, patterns of relationships from the preliminary species tree are repeated. In some cases,
941 as in b), outgroups are also non-monophyletic and subsets of outgroup samples are recovered as
942 sister to the ingroup subclades. Alignments for artifactual orthologs exhibit relatively high
943 variation with observable motifs between the suspected copies, as indicated by the black and
944 white dots. c) Overlap between artifactual orthologs identified by-eye, and paralog warnings
945 raised by automated paralogy detection strategies. d) Expectations for informative, single-copy
946 orthologs change for many gene tree filtering metrics when data sets include artifactual
947 orthologs; arrows indicate whether expected values for metrics are high or low. e) Impact of gene
948 tree filtering strategies on removal of artifactual and putative orthologs as identified by-eye. The

949 dashed lines indicate the various thresholds above which gene trees were kept in filtered datasets,
950 as indicated in Table 1.
951
952 FIGURE 3. Names of clades consistently recovered and species tree topologies for a) The
953 consensus tree of the four datasets in which the proportion of artifactual orthologs was reduced
954 (*bipartition*, *by.eye*, *HybPhaser*, *orthologs.HybPiper.no.warnings*; left) versus the
955 orthologs.unfiltered species tree (right). The placement of *F. magnibracteolata*– the major
956 topological conflict between the two trees– is highlighted with the blue line connecting the tip
957 labels in each. Branches with full support (LPP from ASTRAL = 1) are indicated by an asterisk;
958 elsewhere, the LPP * 100 is provided. Support for each of the four datasets in the consensus tree
959 are given clockwise from the top left: *bipartition*, *by.eye*, *HybPhaser*,
960 *orthologs.HybPiper.no.warnings*. b) Cartoon trees displaying the most frequent differences in
961 relationships between clades recovered by different datasets and the consensus tree of all
962 datasets. Cartoons depict one topological scenario but not the only topological outcome.
963
964 TABLE 1. Descriptions of the filtering criteria used, names applied to each of the datasets,
965 number of loci removed and selected by each, and the impact of filtering criteria on gene tree
966 discordance, branch support, and species tree topology. Values for datasets that performed
967 equally or better than the orthologs.unfiltered dataset for each metric are bolded. Major
968 topological conflicts between species trees and a consensus topology of the 11 filtered datasets
969 (Fig. 3b) are shown, including instances where the species tree is compatible (+) or in
970 disagreement (-) with the consensus topology. Datasets from which the Elaphoglossifolia group
971 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.

ONLINE APPENDIX 1. Supplementary Materials and Methods

ONLINE APPENDIX 2. Paralogs identified by HybPhaser for all samples and for each sample.

SUPPLEMENTARY FIGURE S1. Heat map showing gene recovery efficiency for all submitted samples (rows) across the 353 targeted genes (columns).

SUPPLEMENTARY FIGURE S2. ASTRAL-III species trees for the 11 datasets: a) *orthologs.unfiltered*, b) *orthologs.by.eye*, c) *orthologs.HybPiper.long*, d) *orthologs.HybPiper.no.warnings*, e) *HybPhaser*, f) *MO*, g) *proportion.PI*, h) *proportion.internal.branch.lengths*, i) *average.BS*, j) *tree.length*, k) *bipartition*

SUPPLEMENTARY FIGURE S3. Number of reads mapped to target regions in silica (blue) and herbarium (red) samples. There is a correlation between the number of reads mapped and the number of paralog warnings from HybPiper in a sample ($R^2 = 0.55$, $F(df=1) = 123.6$, $p < 0.0001$). Assembly and paralog detection are improved with increased sequencing effort.

SUPPLEMENTARY FIGURE S4. Tanglegram plots showing pairwise comparisons between ASTRAL-III species tree topologies for a) *consensus* vs. *orthologs.unfiltered*, b) *consensus* vs. *orthologs.by.eye*, c) *consensus* vs. *orthologs.HybPiper2.long*, d) *consensus* vs. *orthologs.HybPiper2.no.warnings*, e) *consensus* vs. *HybPhaser*, f) *consensus* vs. *MO*, g) *consensus* vs. *proportion.PI*, h) *consensus* vs. *proportion.internal.branch.lengths*, i) *consensus* vs. *average.BS*, j) *consensus* vs. *tree.length*, k) *consensus* vs. *bipartition*, l) *orthologs.by.eye* vs.

998 *orthologs.unfiltered*, m) *orthologs.by.eye* vs. *orthologs.HybPiper2.long*, n) *orthologs.by.eye* vs.
999 *orthologs.HybPiper2.no.warnings*, o) *orthologs.by.eye* vs. *HybPhaser*, p) *orthologs.by.eye* vs.
1000 *MO*, q) *orthologs.by.eye* vs. *proportion.PI*, r) *orthologs.by.eye* vs.
1001 *proportion.internal.branch.lengths*, s) *orthologs.by.eye* vs. *average.BS*, t) *orthologs.by.eye* vs.
1002 *tree.length*, u) *orthologs.by.eye* vs. *bipartition*, v) *orthologs.HybPiper2.no.warnings* vs.
1003 *orthologs.unfiltered*, w) *orthologs.HybPiper2.no.warnings* vs. *orthologs.HybPiper2.long*, x)
1004 *orthologs.HybPiper2.no.warnings* vs. *HybPhaser*, y) *orthologs.HybPiper2.no.warnings* vs. MO,
1005 z) *orthologs.HybPiper2.no.warnings* vs. *proportion.PI*, aa) *orthologs.HybPiper2.no.warnings* vs.
1006 *proportion.internal.branch.lengths*, bb) *orthologs.HybPiper2.no.warnings* vs. *average.BS*, cc)
1007 *orthologs.HybPiper2.no.warnings* vs. *tree.length*, dd) *orthologs.HybPiper2.no.warnings* vs.
1008 *bipartition*, ee) *HybPhaser* vs. *orthologs.unfiltered*, ff) *HybPhaser* vs. *orthologs.HybPiper2.long*,
1009 gg) *HybPhaser* vs. MO, hh) *HybPhaser* vs. *proportion.PI*, ii) *HybPhaser* vs.
1010 *proportion.internal.branch.lengths*, jj) *HybPhaser* vs. *average.BS*, kk) *HybPhaser* vs.
1011 *tree.length*, ll) *HybPhaser* vs. *bipartition*, mm) *bipartition* vs. *orthologs.unfiltered*, nn)
1012 *bipartition* vs. *orthologs.HybPiper2.long*, oo) *bipartition* vs. *MO*, pp) *bipartition* vs.
1013 *proportion.PI*, qq) *bipartition* vs. *proportion.internal.branch.lengths*, rr) *bipartition* vs.
1014 *average.BS*, ss) *bipartition* vs. *tree.length*, tt) *orthologs.HybPiper2.long* vs. *orthologs.unfiltered*,
1015 uu) *MO* vs. *orthologs.unfiltered*, vv) *proportion.PI* vs. *orthologs.unfiltered*, xx)
1016 *proportion.internal.branch.lengths* vs. *orthologs.unfiltered*, yy) *average.BS* vs.
1017 *orthologs.unfiltered*, zz) *tree.length* vs. *orthologs.unfiltered.*
1018
1019 SUPPLEMENTARY FIGURE S5. Results of D and f4 statistics calculated with Dsuite on the
1020 “bipartition support”, “by eye”, and “HybPiper2 no warnings" datasets. The columns indicate 1)
1021 values of Patterson’s D-statistic for evaluated trios, 2) highly significant p-values associated with

the D-statistics after Benjamini-Hochberg correction, and 3) f4 ratio statistics. Introgression is detected across all datasets even after assessing family-wise error rate.

SUPPLEMENTARY FIGURE S6. Scatterplot showing locus heterozygosity (proportion of loci with SNPs) versus allele divergence calculated with HybPhaser for all included accessions.

SUPPLEMENTARY TABLE S1. Per sample data--species name, voucher information and herbarium code for tissue gathered from specimens (codes follow Index Herbariorum: http://sweetgum.nybg.org/science/ih/), collection date, provenance, and sample ID used in phylogenetic analyses--and summary statistics generated with HybPiper v.1.3.1.

SUPPLEMENTARY TABLE S2. HybPiper v.2.0.1 assembly statistics for all samples included in this study. Paralog warnings per individual are shown.

SUPPLEMENTARY TABLE S3. Paralog identification results from the by-eye and automated strategies per locus. The proportion of stitched contigs, paralog warning by depth, and long paralog warnings from HybPiper v.2.0.1, paralog warnings from HybPiper v.1.3.1, removed loci with HybPhaser as well as per locus statistics for alignments and gene trees used to explore how standard gene tree filtering affects the presence of artifactual orthologs in our dataset are shown. Details of by-eye identification of paralogs, and the handling of reports from HybPiper v,2.0.1 are in Appendix 1.

SUPPLEMENTARY TABLE S4. HybPiper v.2.0.1 paralog report from assembly using a taxon-specific target file generated with Easy353. Reference sequences were extracted from transcriptome data of Ternstroemia gymnanthera (Pentaphylacaceae). Numbers on the table correspond to the number of contigs assembled per locus.

1048
1049 SUPPLEMENTARY TABLE S5. Per sample summary statistics from HybPhaser, including the
1050 recovered sequence length (bp), the recovered sequence length as the proportion of the target
1051 sequences from HybPiper (bpoftarget), number of paralogs found in all samples, number of
1052 additional paralogs found for each sample, number of loci recovered, allele divergence, locus
1053 heterozygosity, loci with >0.5% SNPs, loci with >1% SNPs, loci with >2% SNPs