The danger zone: the joint trap of incomplete lineage sorting and long-branch attraction in resolving the Gondwanan origin of Rafflesiaceae and Apodanthaceae

Liming Cai[1,2], Natalia Pabón-Mora[3], Juan F. Alzate[4], Liang Liu[5], Charles C. Davis[2]

[1]Department of Biology, University of Florida, Gainesville, FL 32611, USA
[2]Department of Organismal Biology, Harvard University Herbaria, Harvard University, Cambridge, MA 02138, USA
[3] Instituto de Biología, Universidad de Antioquia, Medellín, Colombia
[4] Facultad de Medicina, Centro Nacional de Secuenciación Genómica, Sede de Investigación Universitaria, Universidad de Antioquia, Medellín, Colombia
[5]Department of Statistics, the University of Georgia, Athens, GA 30602, USA

Abstract

Two key factors have been implicated as major impediments to phylogenomic inference: incomplete lineage sorting (ILS)—especially in cases where clades are in the so-called anomaly zone—and erroneous gene tree estimation—commonly manifested by long-branch attraction in the Felsenstein zone. Rafflesiaceae (Malpighiales) is an iconic parasitic plant clade whose species occur west of Wallace's line in tropical Southeast Asia. This clade has been notoriously difficult to place phylogenetically and is nested within an explosive ancient radiation, rendering the Malpighiales one of the thorniest nodes across the angiosperm Tree of Life. The parasitic family Apodanthaceae has recently been positioned as sister to Rafflesiaceae, offering the hope to stabilize their placements. Here, using a dataset of 2,135 genes and complex species tree inference methods, the monophyletic Rafflesiaceae+Apodanthaceae clade is variously placed with Euphorbiaceae, Peraceae, Putranjivaceae, and Pandaceae. Such unstable placements appear to be the result of excessive levels of rate heterogeneity and ILS, which contribute to a phylogenetic "danger zone" where simulation suggests that current methods and genomic data will never provide a tidy phylogenetic resolution for species trees alike. Despite the topological uncertainty, however, our divergence time estimation identifies a mid-Cretaceous origin of

stem group Rafflesiaceae and Apodanthaceae, not only making them the oldest parasitic plant lineage reported to date, but also suggests a likely Gondwana vicariance scenario to explain their disjunct distribution in South America, Africa, Australia, and Northern India.

Keywords

Adaptive radiation, long branch attraction, anomaly zone, composition bias, heterotachy, PhyloBayes, divergence time, biogeography

Introduction

Adaptive radiation is one of the most fascinating phenomena in evolutionary ecology; it occurs when a single lineage rapidly diverges into many descendant lineages each with substantial changes in their morphology, behavior, and physiology (Schluter 2000; Losos and Mahler 2010). Well-known examples include three-spined sticklebacks, Galapagos finches, Caribbean anoles, and Hawaiian silverswords and lobeliads (Schluter and McPhail 1992; Baldwin 1997; Grant and Grant 2007; Givnish et al. 2009; Losos 2011). Despite the intriguing nature of adaptive radiations, such clades often present challenges for phylogenetic inference because they frequently exhibit a combination of short and long branches resulting from extensive accumulation of mutations (Glor 2010; Walker and Smith 2023). Species trees involving adaptive radiations are likely to fall within the "anomaly zone" in which the most common gene tree differs from the species tree, resulting in a potentially high amount of incomplete lineage sorting (ILS) (Degnan and Rosenberg 2006). In addition, mutational saturation and rate heterogeneity among descendant branches in these radiations can lead to various systematic errors, including long-branch attraction and compositional bias (Kapli et al. 2020).

These factors are further confounded when the radiation is ancient, as in the case of Malpighiales—an apparent explosive Cretaceous radiation that has produced plants with diverse morphologies ranging from tropical mangroves (Rhizophoraceae) to submerged aquatics (Podostemaceae) (Davis et al. 2005; Wurdack and Davis 2009; Cai et al. 2020). Among the strangest of all malpighialean members is the parasitic family, Rafflesiaceae, whose species produce the world's largest flowers (Davis et al. 2007; Nikolov and Davis

2017). Until recently, the large-flowered Rafflesiaceae *sensu stricto* (up to 1 m) were thought to be the only parasites within Malpighiales (Barkman et al. 2004; Davis and Wurdack 2004; Davis et al. 2007; Wurdack and Davis 2009), but Alzate et al. (2024) demonstrated the surprising placement of the small-flowered endoparasite Apodanthaceae species (less than 0.5 cm) within Malpighiales as sister to Rafflesiaceae. Their hypothesized synapomorphies include most notably the endophytic habit (endoparasite) where the plant persists as strands of cells with no distinguishable vegetative organs embedded in their hosts; until their flowers emerge briefly for sexual reproduction (Kuijt 1969; Nikolov et al. 2013; Thorogood et al. 2021; Teixeira-Costa and Davis 2021; González and Pabón-Mora 2017; Alzate et al. 2024). However, the two families have very distinct geographic distributions and host preferences. Rafflesiaceae parasitize vines of *Tetrastigma* (Vitaceae) and grow west of Wallace's line in tropical Southeast Asia (Kuijt 1969; Nikolov et al. 2013). In contrast, Apodanthaceae exhibits a pantropical distribution with disjunct occurrences in subtropics and the Mediterranean. Members of Apodanthaceae parasitize Salicaceae and diverse Fabaceae shrubs and trees (Bellot and Renner 2014).

Reconstructing the early evolutionary history of these two parasitic families, including their phylogenetic placements and divergence time, will provide important insights into the biogeographical and ecological processes shaping these extraordinary parasites. Yet the path to their resolution has been slow. Rafflesiaceae has been proposed as sister to Passifloraceae and Euphorbiaceae *sensu stricto* within Malpighiales (Barkman et al. 2004; Nickrent et al. 2004; Davis et al. 2007; Bendiksby et al. 2010) and Apodanthaceae was a widely accepted member of Cucurbitales (Filipowicz and Renner 2010). The lack of genomic resources, excessive rate heterogeneity, and horizontal gene transfer have substantially impeded these previous investigations. The recently available genomic and transcriptomic resources in both families provide a promising opportunity to expand the toolkit for resolving their placement. In particular, the addition of Apodanthaceae offers hope of breaking up the long branch associated with stem group Rafflesiaceae. Despite this promise, there are four genomic aspects that will likely confound phylogenomic investigations. First, nearly half of the conserved genes in Rafflesiaceae (and likely in Apodanthaceae) have been lost due to the evolution of endoparasitism (Cai et al. 2021a),

95 which greatly restricts the repertoire of genes for phylogenomic research. Moreover, these
96 gene losses are overrepresented in housekeeping functions such as photosynthesis and
97 plastid organization, which may lead to incorrect ortholog assessments during
98 phylogenetic analyses (Xiong et al. 2022). Second, parasitic plants are known to acquire
99 foreign genetic materials from their hosts via horizontal gene transfer (HGT) (Davis and
100 Wurdack 2004; Xi et al. 2012; Xi et al. 2013). The incorrect placement of Apodanthaceae in
101 Cucurbitales is likely an outcome of horizontally transferred *matR* and nuclear 18S from
102 ancestral hosts (Filipowicz and Renner 2010). Third, the well-characterized high
103 substitution rate in Rafflesiaceae may lead to spurious species tree reconstruction because
104 of long-branch attraction (Nickrent and Starr 1994; Nickrent and Duff 1996). This
105 phenomenon was shown in all three cellular genomes using representative genes across
106 multiple independently evolved parasitic lineages (Young and dePamphilis 2005;
107 Bromham et al. 2013) and has been demonstrated to span nearly every gene across the
108 entire nuclear genome in Rafflesiaceae (Cai et al. 2021). Fourth, the unusual AT-rich (GC%
109 = 24%) genome content of Rafflesiaceae likely contributes to compositional bias that may
110 incorrectly group distantly related taxa as close relatives (Foster and Hickey 1999; Knight
111 et al. 2001; Wang et al. 2004). Little is known regarding the substitution rates and AT
112 content in Apodanthaceae. These confounding factors make the placement of this enigmatic
113 clade incredibly challenging, but this is further compounded by the nature of their
114 inclusion in the large and diverse order Malpighiales.
115
116 The ancient radiation of Malpighiales includes nine of the ten most phylogenetically
117 recalcitrant angiosperm clades (Smith et al. 2013; Cai et al. 2019; Cai et al. 2020), easily
118 placing its members as the most extraordinary and difficult phylogenomic problem to
119 resolve in angiosperms. In addition, the potentially greatly accelerated substitution rates in
120 Rafflesiaceae and Apodanthaceae combined with the previously summarized genomic
121 anomalies attributable to their parasitic lifestyle make this an even greater intractable
122 phylogenomic problem. This parasitic plant clade falls into a joint trap that we term "the
123 danger zone" of ILS and long-branch attraction. The "danger zone" represents a perturbing
124 combination not conducive to achieving phylogenetic resolution. But we envision that a
125 careful assessment of these factors will yield insights into how to tackle an enormous

126 phylogenomic challenge, which is likely to be found elsewhere in the tree of life. Here, using
127 2,135 genes including a broad sampling of Malpighiales, we tackle the joint problem of ILS
128 and long-branch attraction using a combination of approaches that accommodate
129 genealogical heterogeneity among loci and rate heterogeneity among sites. Our divergence
130 time estimation demonstrated an ancient Gondwanan divergence of Rafflesiaceae and
131 Apodanthaceae providing insight into their disjunct distribution and distinct host
132 preference.
133
134 **RESULTS**
135 **Phylogenomic datasets**
136 By combining data from new RNA sequencing data from Apodanthaceae and two published
137 genome and transcriptome studies within Malpighiales (Cai et al. 2019; Cai et al. 2021), we
138 generated a phylogenomic dataset of 2,135 loci (referred to as G2135 hereafter; Data S1),
139 each containing at least 10 out of 45 species in Malpighiales (Tables S1-2). A phylogenetic
140 based approach from Yang and Smith (2014) was applied to remove paralogs and potential
141 HGTs (Fig. S1). These 45 species included all three genera of Rafflesiaceae (*Sapria, Rafflesia,*
142 and *Rhizanthes*), both genera from Apodanthaceae (*Apodanthes* and *Pilostyles*), and three
143 outgroup species from Oxalidales and Celastrales. Individual DNA and protein alignments
144 were filtered by the profile hidden Markov models implemented in HmmCleaner (Di
145 Franco et al. 2019). The final concatenated matrix consisted of 3,099,966 DNA codon sites
146 (1,033,322 protein sites) with a gene and character occupancy of 70.3% and 75.5%,
147 respectively. Each locus on average contained 32 species and the median length of DNA
148 alignment was 1,260 bp after trimming. The mean branch support for maximum likelihood
149 (ML) gene trees was 72.0 based on ultrafast bootstrap replicates (UFBP) (Hoang et al.
150 2018), but the mean support was only 57.4 UFBP for the placement of Rafflesiaceae and
151 36.7 UFBP for Apodanthaceae (Table S2). In addition to this complete dataset, we compiled
152 two subsampled datasets with 446 (G446) and 829 loci (G829). The G446 dataset
153 contained loci where Rafflesiaceae and Apodanthaceae were monophyletic in individual
154 gene trees (Table S3; Data S1). The G829 dataset was a subset of lock-like loci suitable for
155 divergence time estimation and the application of computationally intensive mixture
156 models (Data S1). See Methods below for details on subsampling criteria. All curated

157 alignments and gene trees are available as supplementary data on Zenodo
158 (10.5281/zenodo.11643025).
159
160 **Identifying suitable genes for placing Rafflesiaceae**
161 To identify suitable genes for phylogenetic reconstruction, especially regarding the
162 placement of Rafflesiaceae and Apodanthaceae, we used fourteen metrics to quantify
163 various aspects of phylogenetic properties (Table S2). These metrics included six tree-wise
164 features such as number of species, number of parsimony informative sites, average branch
165 support, total tree length, root-to-tip distance variance, and gene tree–species tree
166 congruence; and eight metrics specific to Rafflesiaceae and Apodanthaceae, including their
167 DNA compositional bias, lineage-specific branch lengths, and branch support. These
168 fourteen metrics were significantly correlated in 50 out of the 91 comparisons (Pearson
169 correlation p-value <0.05; Fig. S2). For example, the mean branch support, which is a
170 widely applied metric for selecting high-quality phylogenetic markers, was positively
171 correlated with the number of parsimony informative sites (p-value = 8.9e-16; Pearson
172 correlation coefficient = 0.225) and gene tree–species tree congruence (p-value = 0;
173 Pearson correlation coefficient = 0.303). Rafflesiaceae- and Apodanthaceae-specific
174 features were strongly correlated with each other, but relatively independent of the tree-
175 wise metrics. For example, the branch support of Rafflesiaceae and Apodanthaceae was
176 only significantly positively correlated with each other and their branch lengths, but not
177 any tree-wise features (Fig. S2). The only exceptions involved the correlated branch length
178 between Rafflesiaceae, Apodanthaceae, and the total tree length and variation because the
179 long terminal branches of these parasites had a substantial influence on the global mean.
180
181 **Initial phylogenetic reconstructions yielded weakly supported species trees**
182 Explorative phylogenetic reconstructions using concatenation and coalescent methods
183 unanimously supported the monophyly of Rafflesiaceae and Apodanthaceae (100 UFBP or
184 1.0 posterior probability (PP) in all 22 analyses). Despite the effort to mitigate homoplasy,
185 rate heterogeneity, and ILS, the relative replacement of Rafflesiaceae and Apodanthaceae
186 within Malpighiales remained elusive (Table 1; Table S4; Data S2).
187

188 Concatenated DNA analyses of the G2135 dataset placed Rafflesiaceae and Apodanthaceae
189 as sister to the MRCA of Rhizophoraceae+Erythroxylaceae and Ixonanthaceae with
190 moderate to high support (89–99 UFBP; Analysis 1 and 2 in Table 1). However, reduced
191 support and alternative topologies were obtained when attempting to mitigate saturation
192 by removing the third codon position or using protein sequences. For example, after
193 removing the third codon position and applying the more stringent alignment site
194 homology filtering criteria, Rafflesiaceae and Apodanthaceae were placed as sister to
195 Euphorbiaceae+Peraceae with low support (33-38 UFBP) based on DNA and protein
196 alignments (Analysis 4 and 7 in Table 1). The application of complex heterotachy and
197 profile mixture models did not improve the result either. The DNA analysis using the
198 GHOST heterotachy model with four mixture classes placed Rafflesiaceae and
199 Apodanthaceae with Linaceae with a moderate support of 63 UFBP. Application of the
200 protein profile mixture model placed Rafflesiaceae and Apodanthaceae with the MRCA of
201 Euphorbiaceae+Peraceae, Putranjivaceae, and Pandaceae (69 UFBP). Finally, we explored
202 the performance of an *a posteriori* set of 446 loci (G446) where Rafflesiaceae and
203 Apodanthaceae were monophyletic in gene trees. Both DNA and protein analyses of G446
204 placed Rafflesiaceae and Apodanthaceae as sister to all other members of the Malpighiales
205 with low support (44–55 UFBP; Analysis 6 and 10 in Table 1).
206
207 Compared to concatenation, coalescent analyses recovered more consistent results. Ten
208 out of the twelve coalescent analyses placed Rafflesiaceae and Apodanthaceae variously
209 with Euphorbiaceae+Peraceae, Putranjivaceae, Pandaceae, and Rhizophoraceae+
210 Erythroxylaceae with moderate to low support (0.35–0.99 PP; 7–20 BP; Analysis 11-22 in
211 Table 1). Even when using the G446 dataset, both ASTRAL and MP-EST placed Rafflesiaceae
212 and Apodanthaceae as sister to the MRCA of Putranjivaceae, Pandaceae, and
213 Euphorbiaceae+Peraceae. However, in two analyses using the gene trees inferred from the
214 GTR model *without* rate heterogeneity across sites, both ASTRAL and MP-EST placed
215 Rafflesiaceae and Apodanthaceae with Linaceae (0.99 PP in ASTRAL and 4 BP in MP-EST).
216 In particular, gene trees from this enforced GTR model recovered an exceptionally high
217 proportion of monophyletic Rafflesiaceae and Apodanthaceae (n = 653) among individual

218 gene trees compared to models allowing for rate heterogeneity across sites (n = 423–446;
219 Table S3).
220
221 Despite these conflicts, initial analyses using concatenation and coalescent frequently
222 placed Rafflesiaceae and Apodanthaceae with four clades consisting of
223 Euphorbiaceae+Peraceae, Putranjivaceae, Pandaceae, and
224 Erythroxylaceae+Rhizophoraceae. The divergence between these clades was characterized
225 by a series of short internal branches, leading to extensive phylogenetic conflicts (Fig.1).
226 The concordant gene frequencies (gCF) among these four clades plus Rafflesiaceae and
227 Apodanthaceae, defined as the percentage of decisive gene trees supporting that branch
228 (Minh, Hahn, et al. 2020), ranged from 0.41% to 6.39%, with most genes being
229 uninformative (86.22–96.09%). The gCFs for other early-diverging lineages in Malpighiales
230 were similarly low and ranged from 0.02% to 3.85%, reflecting the well-recognized
231 recalcitrant relationship in Malpighiales. In sharp contrast, internal branches within other
232 well-recognized clades such as the parietal clade (MRCA of Salicaceae and Violaceae) and
233 the clusioids (MRCA of Clusiaceae and Calophyllaceae) had gCFs between 20.00–86.32%.
234 Site concordance factors (sCF) showed similar patterns. The sCF ranged from 18.4% to
235 33.9% for DNA and from 28.6% to 34.5% for protein alignments among clades closely
236 related to Rafflesiaceae and Apodanthaceae (Fig. 1). The GC content within the coding
237 regions of Rafflesiaceae and Apodanthaceae was comparable to other free-living
238 Malpighiales and was unlikely to be responsible for the low concordance factors observed
239 above (Fig. S3).

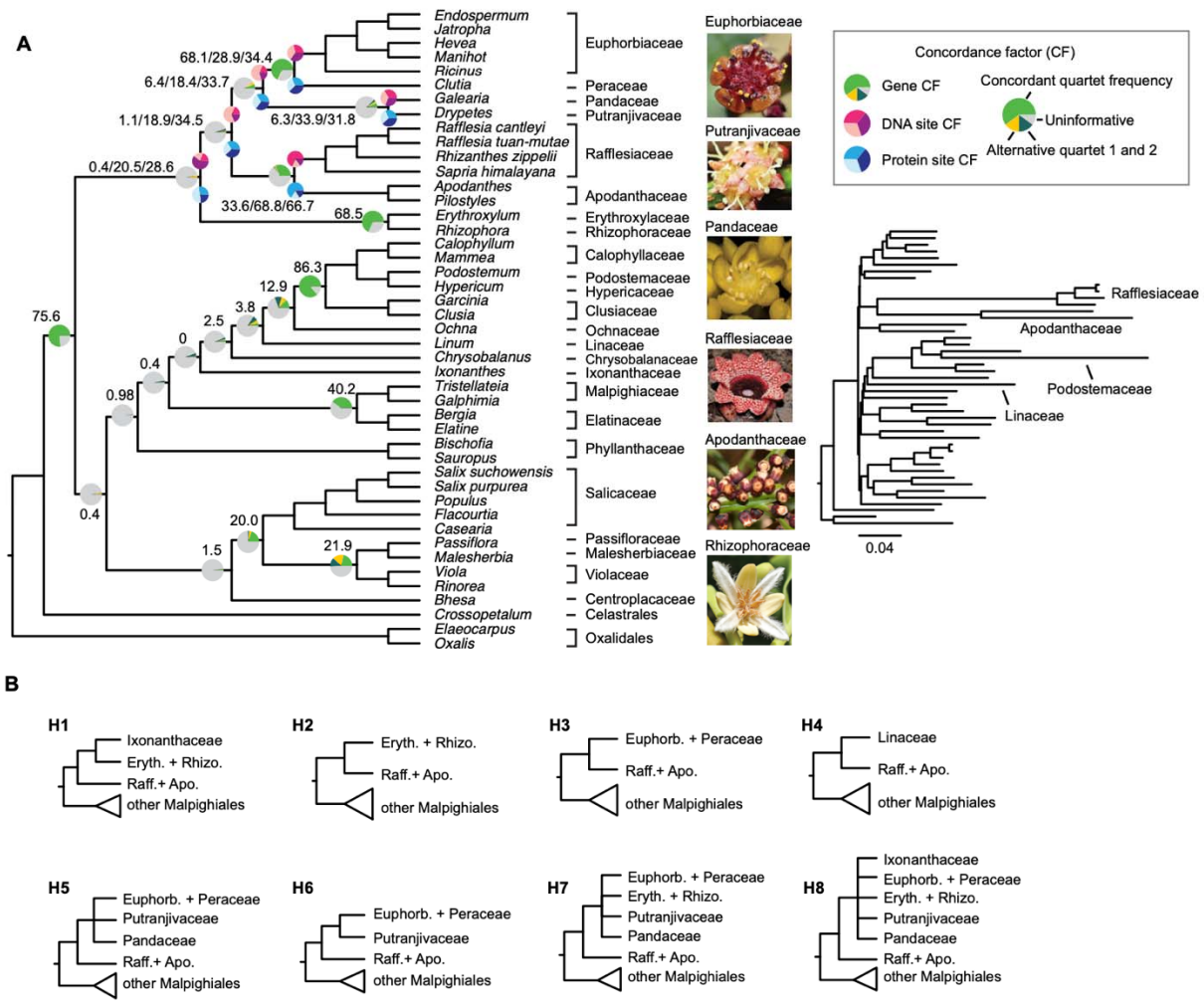


Fig 1. Conflicting phylogenetic signals evidenced by gene and site concordance factors lead to weakly supported conflicting placements of Rafflesiaceae and Apodanthaceae. **A**. The species tree cladogram is inferred by ASTRAL from the complete dataset (Analysis 15 in Table 1). The phylogram with branch lengths in mutational units shown on the right is estimated from IQ-TREE using concatenated DNA alignments. For close relatives of Rafflesiaceae and Apodanthaceae, the gene, DNA site, and protein site concordance factors (CFs) are labeled as numbers and pie charts on each node. For the rest of the tree, only gene concordance factors are labeled and visualized. **B**. Alternative placement of Rafflesiaceae and Apodanthaceae (Raff. + Apo.) across various data subsampling and species tree inference schemes. The 8 hypothesized placements of Rafflesiaceae and Apodanthaceae are variously supported by coalescent or concatenation analyses detailed in Table 1 and Table S4. Family names abbreviated for illustrative purposes include Rhizophoraceae (Rhizo.), Erythroxylaceae (Eryth.), and Euphorbiaceae (Euphorb.).

Table 1 Phylogenetic analysis schemes for 22 concatenation and coalescent methods. The placement and support for Rafflesiaceae and Apodanthaceae are listed. See Fig. 1B for the definition of topologies H1–H8 and branch support is evaluated by 1000 ultrafast bootstrap

(UFBP) in IQ-TREE, local posterior probability (PP) in ASTRAL, and non-parametric bootstrapping in MP-EST (BP). Details of the parameter settings of each analysis, including gene and site filtering strategies, are provided in Table S4.

	ID	Data type	Gene set	Starting partition	Substitution model	Program	Placement	Support
Concatenation	1	DNA	G2135	Gene	ModelFinder	IQ-TREE	H1	84 UFBP
	2	DNA	G2135	Codon	ModelFinder	IQ-TREE	H1	99 UFBP
	3	DNA	G2135	Codon	ModelFinder	IQ-TREE	H2	41 UFBP
	4	DNA	G2135	Codon	ModelFinder	IQ-TREE	H3	38 UFBP
	5	DNA	G2135	Site	GHOST	IQ-TREE	H4	63 UFBP
	6	DNA	G446	Gene	ModelFinder	IQ-TREE	sister to all other Malpighiales	55 UFBP
	7	Protein	G2135	Gene	ModelFinder	IQ-TREE	H3	33 UFBP
	8	Protein	G2135	Gene	ModelFinder	IQ-TREE	Centroplacaceae	27 UFBP
	9	Protein	G829	Site	PMSF	IQ-TREE	H5	69 UFBP
	10	Protein	G446	Gene	ModelFinder	IQ-TREE	sister to all other Malpighiales	44 UFBP
Coalescent	11	DNA	G2135	Codon	GTR (no +I/R/G)	ASTRAL IV	H4	0.99 PP
	12	DNA	G2135	Codon	ModelFinder	ASTRAL IV	H6	0.48 PP
	13	DNA	G2135	Gene	GTR (with +I/R/G)	ASTRAL IV	H7	0.67 PP
	14	DNA	G2135	Codon	GTR (with +I/R/G)	ASTRAL IV	H8	0.35 PP
	15	DNA	G2135	Gene	ModelFinder	ASTRAL IV	H5	0.51 PP
	16	DNA	G446	Gene	ModelFinder	ASTRAL IV	H5	0.52 PP
	17	DNA	G2135	Codon	GTR (no +I/R/G)	MP-EST	H4	4 BP
	18	DNA	G2135	Codon	ModelFinder	MP-EST	H7	20 BP
	19	DNA	G2135	Gene	GTR (with +I/R/G)	MP-EST	H5	14 BP
	20	DNA	G2135	Codon	GTR (with +I/R/G)	MP-EST	H8	17 BP
	21	DNA	G2135	Gene	ModelFinder	MP-EST	H7	8 BP
	22	DNA	G446	Gene	ModelFinder	MP-EST	H5	7 BP

Fast-evolving genes and sites drive biased support

We evaluated the goodness-of-fit of individual genes and sites to eight alternative placements of Rafflesiaceae and Apodanthaceae derived from explorative phylogenetic analyses (defined in Table 1) using the method from Shen et al. (2017). In particular, we

268 examined how spurious relationships might arise from fast-evolving genes and sites. The
269 results turned out to be sensitive to codon position, substitution rate, and type of input
270 data (i.e., DNA versus protein).
271
272 We first examined the log-likelihoods (LL) of individual sites and confirmed that the
273 differences of log-likelihoods (ΔLL) among topologies were similar and there was no site
274 with outstanding ΔLL values (Fig. S4). However, the site rate among the three codon
275 positions varied substantially, with the rates in the third codon being nearly five times
276 (median = 1.954) compared to those in the first (median = 0.483) and second codon
277 positions (median = 0.327; Fig. S5). When evaluating topologies based on concatenated
278 codon positions, both the first and third codons strongly rejected H5, where Rafflesiaceae
279 and Apodanthaceae were placed with Euphorbiaceae+Peraceae, Putranjivaceae, and
280 Pandaceae (approximately unbiased (AU) test *p*-value < 9.02e-56; Table S5). In contrast,
281 the second codon supported H5 as a feasible topology (AU test *p*-value = 0.116; Table S5).
282 The first codons also showed a slight preference for H1 compared to H3 (bp-$RELL_{H1}$ =
283 0.641; bp-$RELL_{H3}$ = 0.162; Table S5) whereas the third codon showed a slight preference
284 for H3 instead (bp-$RELL_{H1}$ = 0.187; bp-$RELL_{H3}$ = 0.720) based on the resampling estimated
285 log-likelihoods (RELL) (Kishino et al. 1990).
286
287 Next, we used ΔLL to identify any outlier gene showing exceptionally high support for one
288 topology over another. For DNA alignments, the ΔLL between the best-supported topology
289 H1 and most other alternative topologies was within a similar range (Fig. S6). However,
290 one gene (locus ID = 313) had exceptionally large ΔLL for topology H3, H7, and H8
291 compared to H1 (Fig. S6), the former of which generally placed the parasites with
292 Euphorbiaceae+Peraceae (H3) or their close relatives (H7, H8) instead of
293 Rhizophoraceae+Erythroxylaceae and Ixonanthaceae (H1). In addition, the same gene
294 showed a reverse trend by strongly favoring H3 over H1 based on protein alignment. We
295 thus removed this gene before investigating the impact of substitution rate on topology
296 support.
297

298 To further explore how fast-evolving genes prone to long-branch attraction may contribute
299 to the conflicting placements of Rafflesiaceae and Apodanthaceae, we ordered genes based
300 on the substitution rate estimated by IQ-TREE, which ranged from 0.523 to 1.897 for DNA
301 alignments and 0.078 to 3.056 for protein alignments. We examined how the support for
302 optimal topology changed with the addition of genes with increasing rates. As a negative
303 control where a topology is universally rejected regardless of evolutionary rates, we
304 compared the ΔLL of the optimal topology (H1 in DNA alignments and H3 in protein
305 alignments) to H4, which placed the parasites with Linaceae. Both DNA and protein data
306 demonstrated a consistent and nearly linear increase in ΔLL as more rapidly evolving genes
307 were added (Fig. 2C, F). The trend also aligned with the simulated mean where genes were
308 randomly ordered regardless of rates. The decisive lead in the final ΔLL (>1000) plus the
309 alignment between the empirical and simulated curve of ΔLL demonstrated an example
310 when the alternative topology (H4) could be rejected with confidence (AU test *p*-value
311 <0.002; Table S5) and the impact of substitution rate was minimal. In contrast, support for
312 H1 (Rhizophoraceae+Erythroxylaceae, Ixonanthaceae) over H3 (Euphorbiaceae+Peraceae)
313 in DNA alignments was clearly driven by a small group of fast-evolving genes (Fig. 2A, B).
314 The cumulative ΔLL of H1–H3 was below zero or near zero for the first 1,735 (81.2%)
315 slower-evolving genes based on DNA alignments (Fig. 2B). However, the 9.8% fastest-
316 evolving genes significantly favored the H1 hypothesis over H3 (AU test *p*-value = 0.147
317 versus 0.0114; Table S8) such that they completely reverse the trend (Fig. 2B). This
318 phenomenon was mitigated in protein-based analyses where the empirical curve of ΔLL
319 aligned better with the simulated average (Fig. 2E). But for both DNA and protein data,
320 slower-evolving genes seem to support a closer affinity between the parasites and
321 Euphorbiaceae+Peraceae (H3) instead of Rhizophoraceae+Erythroxylaceae, Ixonanthaceae
322 (H1). Finally, we examined the changes of ΔLL regarding the H5 topology
323 (Euphorbiaceae+Peraceae, Putranjivaceae, Pandaceae), which is favored by many
324 coalescent analyses and one protein-based concatenation analysis using the profile mixture
325 model. DNA alignments from slower-evolving genes similarly showed a preference for this
326 alternative topology H5 than H1 compared to fast-evolving genes (Fig. 2A). Although this is
327 less obvious in the protein analysis (Fig. 2D).

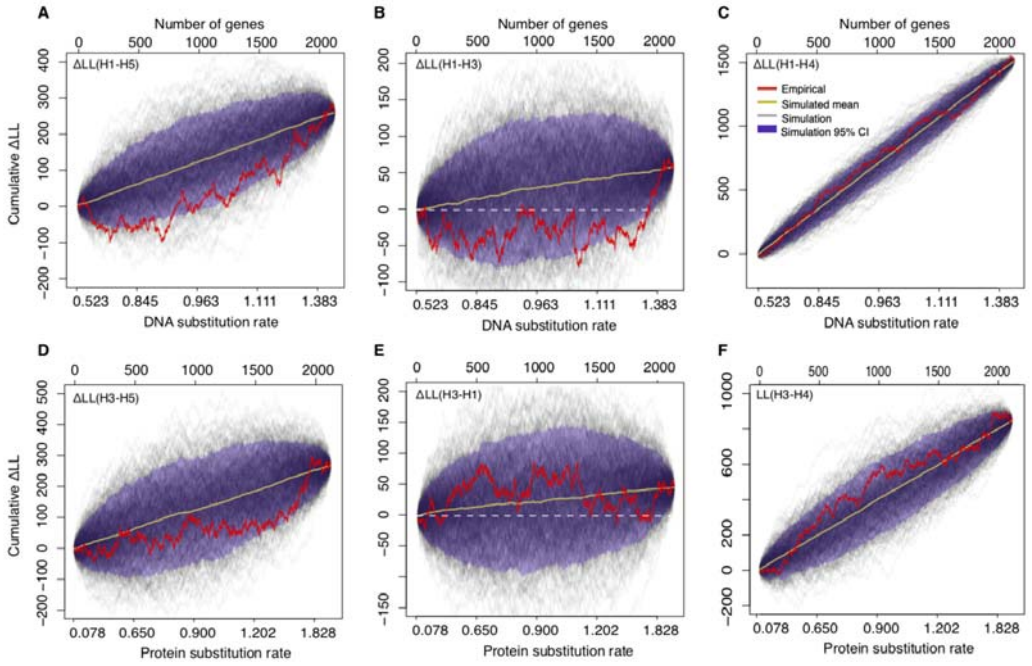
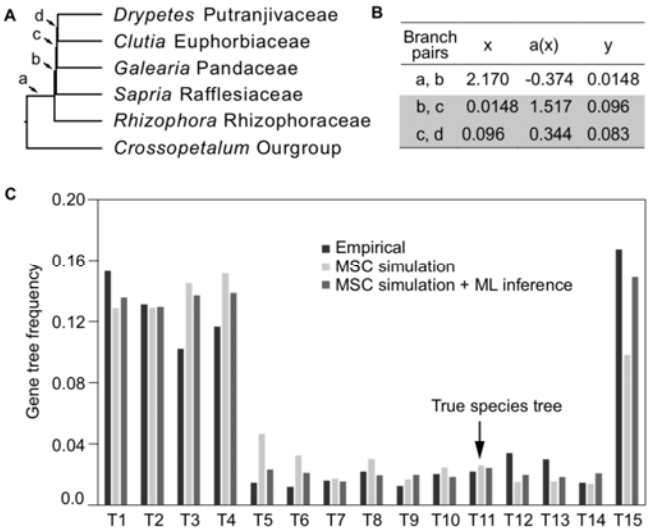


Fig. 2 Changes in cumulative differences of log-likelihood (ΔLL) when adding more rapidly evolving genes. **A–C** ΔLL of the optimum topology H1 compared to three alternative topologies H5 (**A**), H3 (**B**), and H4 (**C**) based on DNA sequences. **D–F** ΔLL of the optimum topology H3 compared to three alternative topologies H5 (**D**), H1 (**E**), and H**4 (F**) based on protein sequences. Topologies H1-H5 are defined in Fig. 1B. All 2,135 genes are ordered on the x-axis based on DNA (A-C) or protein substitution rates (D-E). The rate and likelihoods for individual genes were inferred using the –wpr and -wpl flags under the same optimal substitution model determined by ModelFinder and per-gene partition in IQ-TREE. The red line represents the cumulative ΔLL as genes with higher rates are added. The dark gray lines represent the cumulative ΔLL from 1000 simulated bootstrap replications where genes are randomly ordered. The yellow line represents the simulated average. The gene and site rate and likelihood estimates supporting this figure are provided in Data S3.

The anomaly zone and coalescent simulations

To explore the effects of ILS, we first calculated the branch length in coalescent units for focal clades. The three successive short internal branches among the parasites and the other four focal clades ranged from 0.015 to 0.096 in coalescent units (Fig. 3A). The estimated minimum branch length to escape the anomaly zone was 0.344, which is four

347 times longer than the actual branch lengths (Fig. 3B). This result suggests that these rapid
348 radiations involving Rafflesiaceae+Apodanthaceae are clearly subject to the anomaly zone.
349 Our gene tree simulation with six taxa also corroborated this conclusion; the most common
350 gene tree was discordant from the true species tree (Fig. 3C), which ranked 11th in
351 frequency. The highly consistent frequencies between the empirical and simulated gene
352 trees suggested that most gene tree variations can be explained by ILS. However, a
353 significant difference in gene tree frequencies was observed in topology T15, in which
354 Rafflesiaceae was placed sister to all other Malpighiales (Fig. S7). This topology was
355 present in only 11% of the true gene trees in the simulation but increased to 17% when
356 adding in the ML inference step (Fig. 3C).
357



Branch pairs	x	a(x)	y
a, b	2.170	-0.374	0.0148
b, c	0.0148	1.517	0.096
c, d	0.096	0.344	0.083

358

359 **Fig 3**. The anomaly zone and coalescent simulation. **A.** Reference species tree for coalescent
360 simulation with labeled branches. **B**. For each internal branch pair in a four-taxon tree,
361 boundaries of the anomaly zone a(x) calculated from (Degnan and Rosenberg 2006) are
362 compared to the empirical branch length (y). Two out of three branch pairs (b-c and c-d)
363 fall within the anomaly zone as y< a(x). **C.** Gene tree frequency for the 15 alternative
364 placements of Rafflesiaceae (T1–T15) in the empirical gene trees (dark grey), simulated
365 true gene trees under the multispecies coalescent model (MSC; medium grey), and inferred
366 maximum likelihood (ML) gene trees based on the simulated sequences (light grey).
367 Topologies for T1 to T15 are defined in Fig. S7. In all three cases, gene trees consistent with
368 the true species tree topology have very low frequency.
369

370 When inferring the species tree from simulated alignments, none of the concatenated
371 analyses recovered the true species tree. Specifically, when more than 4,000 loci were used,

372 the true topology was strongly rejected by the AU test, suggesting statistical inconsistency
373 (Table S6). Instead, the incorrect species topologies were variously supported by AU tests.
374 T5 and T15 were most frequently favored, which placed Rafflesiaceae with (Putranjivaceae,
375 Euphorbiaceae) or the MRCA of Rhizophoraceae and Euphorbiaceae, respectively. In
376 contrast, summary coalescent methods consistently recovered the true species when at
377 least 4,000 loci were used (Table S7). The gene tree estimation error was as high as 78.1%
378 for 500-bp-alignments and 70.0% for 2000-bp-alignments in terms of the placement of
379 Rafflesiaceae. Such high estimation error led to incorrect species tree estimation when less
380 than 3,000 genes were used. But even when enough loci were used and the true species
381 tree was recovered, the BP support for Rafflesiaceae was still moderate to low (35–84 BP)
382 and did not increase with the number of genes used. Moreover, alternative topologies could
383 not be rejected based on the likelihood ratio test. For example, the difference in
384 pseudolikelihoods between the true species tree H11 and an alternative topology H5 was
385 0.71 when using 1,000 genes (pseudolikelihood = -10968.5) and 20.3 when using 10,000
386 genes (pseudolikelihood = -109763.3), which were not statistically significant (*p*–
387 value >0.51, Table S7).
388
389 **Divergence time estimation**
390 Finally, based on a species tree topology from the ASTRAL analysis, we inferred the
391 divergence time of Malpighiales using the Bayesian method implemented in PhyloBayes
392 (Lartillot et al. 2009) and the penalized likelihood method implemented in treePL (Smith
393 and O'Meara 2012). Both approaches supported a mid-Cretaceous origin of Malpighiales,
394 followed by radiation in the Albian that gave rise to the majority of the family-level
395 diversity within Malpighiales (Fig. 4; Fig. S8). The Phylobayes analysis was based on a
396 subset of 829 clock-like loci with the third codon position removed. Crown group
397 Malpighiales was inferred to originate during the Albian at 111.1 Ma (95% confidence
398 interval CI = 109.5–112.1; Fig. 4) and the MRCA of Rafflesiaceae and Apodanthaceae
399 diverged from its sister clade, the MRCA of Euphorbiaceae and Pandaceae, during the
400 Albian at 110.9 Ma (95% CI = 109.2–111.9). Rafflesiaceae diverged from Apodanthaceae at
401 106.5 Ma (95% CI = 102.8–108.7) and crown group Rafflesiaceae arose very recently
402 during the Miocene at 15.1 Ma (95% CI = 9.2–28.1). Crown group Apodanthaceae was

inferred to originate in the early Eocene at 53.1 Ma (95% CI = 35.9–66.4). The treePL analysis was based on the concatenated DNA alignment of the complete G2135 dataset without the third codon. The MRCA of Rafflesiaceae and Apodanthaceae diverged from its sister clade at 106.4 Ma (Fig. S8). Stem group Rafflesiaceae diverged from Apodanthaceae in the Cenomanian at 98.8 Ma and its crown group age is inferred to be 19.0 Ma. Crown group Apodanthaceae, on the other hand, originated around the Paleocene-Eocene transition at 55.6 Ma.

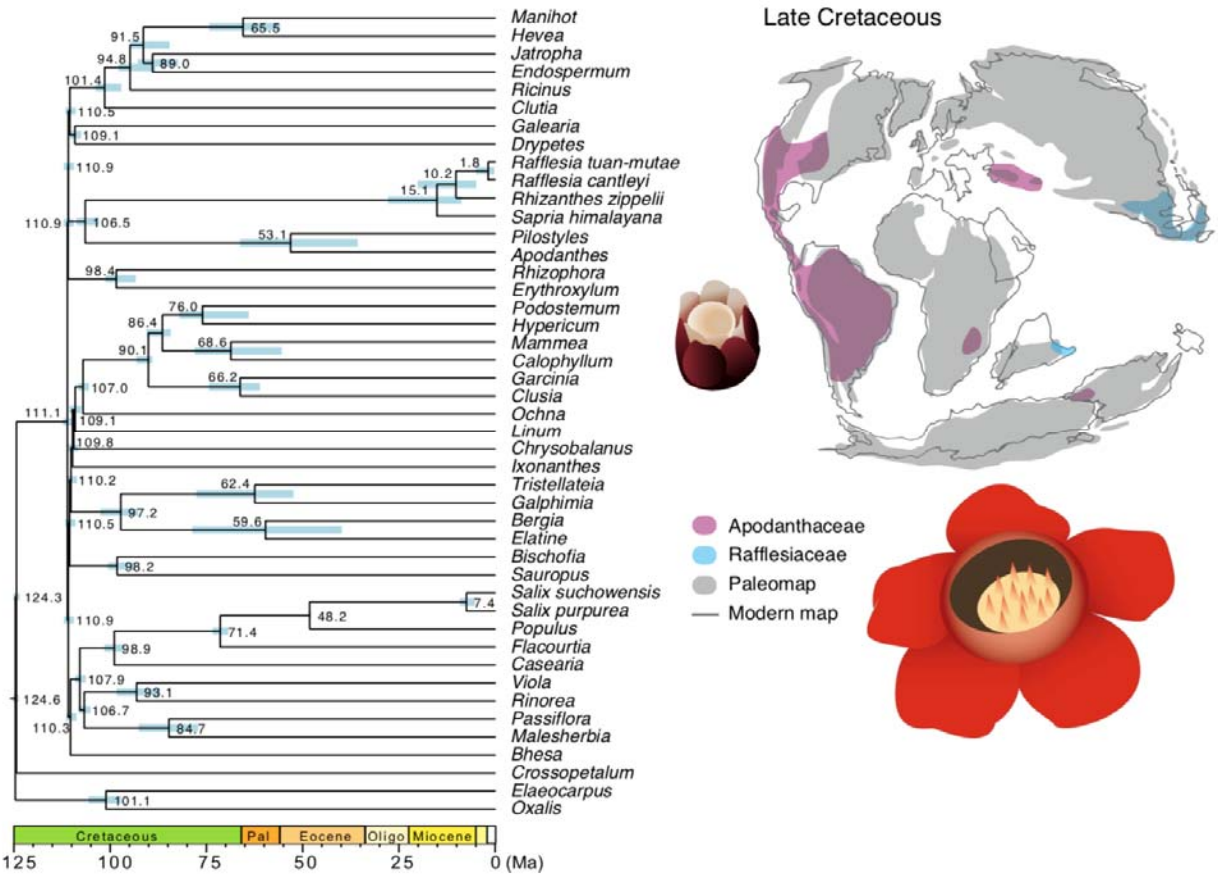


Fig. 4 Bayesian divergence time estimation of Malpighiales. The dated phylogeny is inferred under the CAT-GTR model based on the concatenated DNA sequences from 829 clock-like genes in PhyloBayes. Divergence time measured in million years (Ma) is labeled on each node and the 95% confidence intervals are visualized as bars. Distribution of the extant species in Apodanthaceae and Rafflesiaceae is projected to the paleomap from the Late Cretaceous at 94 million years ago. The map is referenced from the PALEOMAP Project by C. R. Scotese (www.scotese.com).

DISCUSSION

420 **Genome-wide rate acceleration in parasites contributes to long-branch attraction**
421 Despite meticulously curated genomic data, we did not obtain a single tidy placement for
422 Rafflesiaceae and Apodanthaceae within Malpighiales. The shift to such an extreme
423 parasitic lifestyle early in their evolution has undoubtedly reshaped the selective forces
424 acting on otherwise conserved autotrophic functions such as photosynthesis (Cai 2023). As
425 a result, relaxed selection and accelerated nucleotide substitution rates are often
426 characteristic of the plastid and nuclear genomes of these species (Bromham et al. 2013;
427 Wicke et al. 2016; Chen et al. 2023). This rate acceleration has contributed to exceptionally
428 long branches in Rafflesiaceae and Apodanthaceae (Fig. 1A). Previous genomic
429 investigations in Rafflesiaceae revealed that its synonymous substitution rate was 4.6 times
430 higher than its close relative *Manihot* (Euphorbiaceae) and 3.3 times higher than *Populus*
431 (Salicaceae) (Cai et al. 2021). After the divergence of Rafflesiaceae from its free-living
432 relatives, 35.1% of genes have experienced more than one synonymous substitution per
433 site (Cai et al. 2021) and this has led to extensive phylogenetic conflicts when trying to
434 place these parasites. These conflicts are reflected by the low gene and site concordance
435 factors where a remarkable 98.9% of the genes are uninformative for placing Rafflesiaceae
436 and Apodanthaceae (Fig. 1).
437
438 Compared to parsimony, likelihood methods are more robust to long-branch attractions
439 (Felsenstein 1978; Huelsenbeck and Hillis 1993), but overly simplistic substitution models
440 can still lead to long-branch attraction (Hendy and Penny 1989; Zharkikh and Li 1993;
441 Tateno et al. 1994; Bruno and Halpern 1999; Ho and Jermiin 2004; Bergsten 2005;
442 Brinkmann et al. 2005; Philippe et al. 2011). In Malpighiales, models without sufficient rate
443 heterogeneity accommodation often grouped Rafflesiaceae and Apodanthaceae with other
444 apparent "rogue" taxa with long branches such as Linaceae (Analysis 5, 11, and 17 in Table
445 1). For example, the concatenation Analysis 5 employed the GHOST heterotachy model but
446 our high-performance computing cluster can only accommodate four mixture classes due
447 to its high computational demand. As a result, despite designed to mitigate heterotachy, the
448 GHOST model with four mixture classes appeared to be inadequate for capturing cross-site
449 rate variation compared to partition models with 107 to 178 partitions, each with different
450 rate heterogeneity modeling. This overly simplified substitution model in our GHOST

451 analysis may have contributed to the grouping of Linaceae, Rafflesiaceae, and
452 Apodanthaceae, all of which have long root-to-tip branch lengths. In contrast, despite the
453 even longer terminal branch in Podostemaceae and its high GC content, the phylogenetic
454 affinity between Podostemaceae and Hypericaceae was maximally supported in all
455 analyses due to the longer internal branch supporting their shared ancestry (Fig. 2A; Fig. S3;
456 Data S2).
457
458 Our strongest evidence for the influence of long branch attraction stems from the *a*
459 *posteriori* selection of genes and models where Rafflesiaceae and Apodanthaceae are
460 monophyletic. Concatenation of the G446 dataset where the parasites were monophyletic
461 in gene trees placed Rafflesiaceae and Apodanthaceae sister to all other Malpighiales,
462 which is a classic signal of long branch attraction (Kapli et al. 2020). Coalescent analyses of
463 the same set of genes, on the other hand, recovered a more conventional topology (H5).
464 This result suggests that coalescent methods might be more robust to long branch
465 attraction compared to concatenation methods, but they are still not immune to these
466 effects. Among the five gene tree reconstruction methods, applying the GTR model with no
467 rate heterogeneity modeling yielded up to 54.3% more monophyletic Rafflesiaceae and
468 Apodanthaceae across genes (Table S3), but coalescent analysis of these gene trees using
469 ASTRAL and MP-EST consistently placed the parasites with Linaceae, similar to the
470 concatenation results above using the oversimplified GHOST model (Fig. 2C, F). Many of
471 these outcomes of long branch attraction were captured by our simulation. For example,
472 gene tree frequencies for the fifteen alternative topologies were largely consistent between
473 the true gene trees and gene trees inferred from the simulated alignments (Fig. 3C). Yet a
474 disproportionately high frequency of T15 placing Rafflesiaceae sister to all other
475 Malpighiales was recovered among ML-inferred gene trees compared to the true gene trees,
476 likely a result of long-branch attraction during the ML inference step. When aggregating
477 such erroneous signals into concatenated alignments, it contributed to the support of T15
478 in optimum species tree searches and rejection of the true species tree (Fig. 2B; Table S6).
479
480 **Navigating the phylogenetic "Danger Zone" of excessive incomplete lineage sorting**
481 **and long-branch attraction**

482 Like many other rapid radiations in mammals and birds (Liu et al. 2017; Stiller et al. 2024),
483 the explosive malpighialean speciation followed by parasitism-induced rate acceleration
484 created a sharp contrast in branch length before and after the origin of stem group
485 Rafflesiaceae+Apodanthaceae. The combination of high ILS and long-branch attraction
486 presents significant challenges for species reconstruction in both concatenation and
487 coalescent methods (Martyn and Steel 2012; Roch et al. 2019). We define the term "danger
488 zone" to represent an empirically identified species tree space under the extreme influence
489 of ILS and long-branch attraction such that phylogenetic reconstruction is excessively
490 challenging.
491
492 Selecting reliable genes with desired phylogenetic properties under these circumstances is
493 one of the ways to optimize signal-to-noise ratio (Koch 2021) and potentially deal with the
494 "danger zone" problem. Here, we identified that some commonly applied gene-filtering
495 criteria, including high gene tree–species tree congruence, high gene tree support, and low
496 missing taxa, strongly covary in our dataset (Fig. S1). Yet none of these global metrics were
497 useful when choosing informative genes to place Rafflesiaceae and Apodanthaceae. For
498 example, the branch support of Rafflesiaceae and Apodanthaceae was only strongly
499 influenced by the length of its short stem group branch rather than the long descending
500 branch leading to extant genera (Fig. S1). Therefore, the recalcitrant placement of
501 Rafflesiaceae and Apodanthaceae appears to be primarily driven by incomplete lineage
502 sorting. But long-branch attraction caused by fast evolving genes and sites further
503 exacerbated the problem and led to spurious relationships (Fig. 2). Our application of the
504 G446 dataset where Rafflesiaceae and Apodanthaceae were monophyletic was aimed to
505 select "informative" loci based on *a posteriori* criterion, but counterintuitively, phylogenetic
506 reconstruction using these genes demonstrated that they were more prone to long branch
507 attraction (Table 1). Thus, selecting a small set of highly informative genes is not a
508 promising approach to address the "danger zone" problem because an empirically
509 verifiable criterion to optimize signal-to-noise ratio is lacking in our system and many
510 others.
511

512 Alternatively, other efforts to resolve clades similarly suffering from "danger zone"
513 phenomena have turned to a big-data analytics approach. Examples include the application
514 of tens of thousands of loci from whole genome sequencing to resolve the bird Tree of Life
515 (Stiller et al. 2024). Yet our simulation concluded that current methods for modeling
516 substitution and coalescent processes are insufficient to fully resolve the placement of
517 Rafflesiaceae and Apodanthaceae with at least ten thousand loci. The empirical species tree
518 existed within a parameter space where both incomplete lineage sorting and long-branch
519 attraction were prevalent. Concordant gene trees appeared at a very low frequency, and
520 the most common gene trees consistently placed the parasite sister to all other
521 Malpighiales due to long-branch attraction (Fig. 3C). In addition, massive gene loss
522 demonstrated in Rafflesiaceae renders a greatly reduced pool of phylogenetic markers that
523 are potentially insufficient to achieve a definitive placement. Alternative methods
524 employing macro- and micro-scale gene synteny and rare genomic changes have been
525 increasingly implemented to untangle recalcitrant relationships in angiosperms and teleost
526 fishes (Zhao et al. 2021; Parey et al. 2023). As more genomic data is developed in
527 Malpighiales and beyond, the application of these new methods may provide novel insights
528 into this thorny phylogenetic problem.
529
530 **The Gondwana origin of Rafflesiaceae and Apodanthaceae**
531 Despite the phylogenetic uncertainty, coalescent analyses and concatenation analyses with
532 complex models consistently recover similar sister clades for Rafflesiaceae and
533 Apodanthaceae. This improvement in topological resolution, plus the mid-Cretaceous
534 divergence time of stem group Rafflesiaceae and Apodanthaceae, has led to novel insights
535 into the early evolutionary history of these enigmatic parasites.
536
537 Most coalescent and mixture-model-based concatenation analyses placed Rafflesiaceae and
538 Apodanthaceae with clades variously consisting of Putranjivaceae, Pandaceae,
539 Euphorbiaceae, and Peraceae (Table 1), all of which share a pantropical distribution.
540 Although the highly modified morphology of Rafflesiaceae and Apodanthaceae makes it
541 challenging to identify synapomorphies, several shared features support such phylogenetic
542 affinity. For example, the pentamerous dioecious flowers in Rafflesiaceae and

543 Apodanthaceae are commonly found in Pandaceae, Putranjivaceae, and its unsampled close
544 relative Lophopyxidaceae (Matthews and Endress 2013; Wurdack et al. 2004). In addition,
545 Putranjivaceae releases sulfur-rich, strong-smelling chemicals to attract pollinators like
546 Rafflesiaceae and Apodanthaceae. The foul odor of Rafflesiaceae largely comes from
547 dimethyl disulfide and dimethyl trisulfide (Jürgens et al. 2013; Wee et al. 2018), whereas
548 the strong smell in *Drypetes* (Putranjivaceae) comes from isothiocyanates of the
549 glucosinolates pathway (Johnson et al. 2009). Future studies focusing on the metabolic
550 pathway of these chemicals may elucidate whether they are derived from shared ancestry.
551
552 The parasitic common ancestor of Rafflesiaceae and Apodanthaceae diverged from their
553 free-living relatives in the mid-Cretaceous as part of the initial malpighialean radiation
554 (110.9 Ma; 95% CI = 109.2–111.9). This was then followed by the divergence between stem
555 group Rafflesiaceae and Apodanthaceae at 106.5 Ma (95% CI = 102.8–108.7). These Albian
556 age estimates not only suggested that extreme life history strategies such as
557 endoparasitism can evolve rapidly within less than 5 Ma but also pinpointed Rafflesiaceae
558 and Apodanthaceae as the oldest endoparasitic and holoparasitic clades known. Many
559 other holoparasitic lineages such as Hydnoraceae and Cynomoriaceae have comparable
560 stem group ages in the mid-Cretaceous (Naumann et al. 2013), but their much younger or
561 unknown crown group ages make it difficult to determine the exact date when
562 holoparasitism evolved. Our estimation of the crown group Apodanthaceae in Eocene (53.1
563 Ma, 95% CI = 35.9–66.4) aligns with those from Naumann et al. (2013) and Bellot and
564 Renner (2014), but we drastically push forward the age of crown group Rafflesiaceae to
565 Miocene (15.1 Ma; 95% CI = 9.2–28.1). This Miocene age is the youngest compared to
566 previous estimates from Bendiksby et al. (2010) at 81.7 Ma, Pelser et al. (2019) at 68 Ma,
567 and Naumann et al. (2013) at 65.3 Ma; but it aligns well with the relative age constraint
568 imposed by horizontal gene transfer (Davín et al. 2018)—the crown group age of the
569 recipient (Rafflesiaceae) should be younger than the stem group age of the donor
570 (*Ampelopsis* and *Tetrastigma*; 41.2-50.6 Ma) (Chen et al. 2011; Nie et al. 2012a; Cai et al.
571 2021). We anticipate that these age estimates will continue to be refined in future studies
572 as more comprehensive taxon sampling, fossil calibrations, and branch-rate modeling are
573 applied, such as those implemented in starBEAST (Ogilvie et al. 2017).

The mid-Cretaceous divergence and the distribution of extant Rafflesiaceae and Apodanthaceae in South America, Africa, Australia, and Northern India strongly suggest a Gondwanan origin (Fig. 4). Such a fragmented global distribution makes likelihood-based biogeographic reconstruction untraceable. Given the poor long-distance dispersal capacity and restricted current range of these endoparasites (Barkman et al. 2017; Malabrigo Jr. et al. 2025), we propose a southern Gondwana origin and an "out-of-India" dispersal route for Rafflesiaceae. Extant Rafflesiaceae are frequently found in southern Thailand to the Philippines, but the early-diverging *Sapria himalayana* are distributed in the seasonal forest south of the Himalayas in Assam India and Myanmar. This potentially points to an Indian origin of stem group Rafflesiaceae that rafted the continent until its collision with Eurasia during the Eocene. Particularly, we previously demonstrated a host shift in stem group Rafflesiaceae from *Ampelopsis* to *Tetrastigma* (Cai et al. 2021). The major geographical and climatic transitions associated with the India collision may well have served as a trigger for such shifts in life history strategies. The oldest fruit fossil of Vitaceae from the Late Cretaceous was also discovered in India (Manchester et al. 2013). Similar "out-of-India" dispersal has been demonstrated in *Dipterocarpus*, Crypteroniaceae, and multiple lineages of frogs (Bossuyt and Milinkovitch 2001; Conti et al. 2002; Bansal et al. 2022). A northern route via Laurasia is also possible during the Paleocene and Eocene, as has been proposed for the historical host clade *Ampelopsis* (Nie et al. 2012b). Future fossil discoveries, especially involving the distinctive seeds of Rafflesiaceae and Apodanthaceae, may help resolve this mystery.

On the other hand, the late Paleocene to early Eocene divergence between *Apodanthes* and *Pilostyles* in crown group Apodanthaceae was likely caused by ecological- and host-related shifts. This divergence time marks the warmest epoch on earth that permitted the dispersal of dry tropical *Pilostyles* through southern Patagonia, Antarctica to Australia, the corridor of which did not close until late Eocene (Amoo et al. 2022). However, our Eocene age estimates necessitate repeated long-distance dispersal of *Pilostyles* from the neotropics to tropical Africa and Persia. The patchy distribution of *Pilostyles* in present-day Persia that corresponds to the ancient Cimmerian continent could have been caused by large-scale

605 extinction events due to major volcanism events and drastic paleoclimate shifts during the
606 repositioning of this plate.
607
608 **Conclusion**
609 Our empirical investigation and simulation suggested that the combination of excessive
610 incomplete lineage sorting and long-branch attraction created a seemingly intractable
611 phylogenetic "danger zone" where the placement of Rafflesiaceae and Apodanthaceae may
612 never be resolved neatly. The unique life-history strategy of these endoparasites has led to
613 high substitution rates that contribute to long-branch attraction in both concatenation and
614 coalescent analyses. Potential sister clades of Rafflesiaceae and Apodanthaceae include
615 Putranjivaceae, Pandaceae, Euphorbiaceae, and Peraceae, all sharing a pantropical
616 distribution. Divergence time estimates suggest that Rafflesiaceae and Apodanthaceae
617 represent the oldest known parasitic plant lineage, with plant endoparasitism evolving in
618 the ecosystem as early as the mid-Cretaceous. The divergence between Rafflesiaceae and
619 Apodanthaceae and the early diversification of *Pilostyles* (Apodanthaceae) within the
620 neotropics and southern Australia may have been influenced by the Gondwana vicariance.
621 These novel phylogenetic and temporal insights have opened a series of questions
622 regarding the processes that shaped the distinct floral size, geographical distribution, and
623 host preference in these two families. In particular, what may serve as the ancestral host(s)
624 for Rafflesiaceae and Apodanthaceae given their highly specific present-day host
625 preference in Vitaceae, Salicaceae, and Fabaceae? Perhaps close examination of host-
626 derived gene transfers will help reveal the hidden history of these remarkable plants in
627 deep time.
628
629 **Materials and Methods**
630 **RNA sequencing and transcriptome assembly of Apodanthaceae**
631 Apodanthaceae was recently demonstrated to be a close relative of Rafflesiaceae (Alzate et
632 al. 2024) and we thus obtained the transcriptomes of both Apodanthaceae genera for
633 phylogenomic investigation. Sample collection, RNA extraction, sequencing, and
634 transcriptome assembly of *Pilostyles boyacensis* and *Apodanthes caseariae* were described
635 in González et al. (2020) and Alzate et al. (2024), respectively. Briefly, for each species,

636 three tissue samples spanning the full life cycle, including endophytic tissue, preanthetic
637 flowers, and fruits and seeds, were flash frozen in the field. The *Pilostyles* samples were at
638 xerophytic thickets around Villa de Leyva (Boyacá, Colombia) and voucher specimens were
639 deposited at COL (voucher FG 4518 and voucher FG 4519). The *Apodanthes* samples were
640 collected in wet tropical forests around Villavicencio (Department of Meta, Colombia) and
641 the voucher specimen was deposited at HUA (voucher NP 495).
642
643 For RNA extraction, we applied the TRIzol_TM Reagent (Invitrogen) Kit and the Spectrum
644 Plant Total RNA Kit (Sigma-Aldrich TM) for *Pilostyles* and *Apodanthes*, respectively. RNA
645 sequencing library was prepared using the TruSeq Stranded mRNA library construction kit
646 (Illumina) and sequenced on the Illumina Novaseq 6000 instrument in Macrogen (South
647 Korea). Read cleaning was performed with a quality threshold of Q30 and a minimum read
648 length of 70 bp, singletons were excluded. Contig was assembled using the Trinity V2.5.1
649 software (Grabherr et al. 2011) with TRIMMOMATIC adapter removal.
650
651 **Phylogenomic dataset assembly**
652 We combined the phylogenomic datasets from our previous studies (Cai et al. 2019; Cai et
653 al. 2021) and transcriptomes from Apodanthaceae to generate a 2,135-gene dataset (Data
654 S1). The combined dataset included 45 species representing all five extant genera of
655 Rafflesiaceae and Apodanthaceae, twenty-one families broadly sampled across
656 Malpighiales, and three outgroup species from Oxalidales and Celastrales (Table S1).
657 Briefly, the two datasets from Cai et al. (2019) and Cai et al. (2021) were merged based on
658 gene IDs of the three Malpighiales species sampled in both studies (*Manihot esculenta*,
659 *Populus trichocarpa*, and *Jatropha curcas*). Preliminary DNA alignments and gene trees
660 were inferred for each orthogroup using MAFFT v.7.299 (Katoh and Standley 2013) and
661 FastTree v2.1 (Price et al. 2010). These preliminarily filtered alignments were used to build
662 profile hidden Markov models to search for orthologous sequences in Apodanthaceae using
663 HMMER v3.4 (Mistry et al. 2013). For each ortholog, only the best HMMER hit with e-values
664 lower than 1e-70 was used for *Apodanthes* and *Pilostyles*. Finally, the tree-based ortholog
665 identification approach from Yang and Smith (2014) was applied iteratively to remove HGT
666 genes in Rafflesiaceae and Apodanthaceae as well as paralogs arising from gene

duplications. Any taxa with branches longer than 1.5 were removed. This cutoff was determined by the global distribution of branch length where 1.5 is a natural break (Fig. S9). The final phylogenomic dataset consisted of 2,135 protein-coding orthogroups satisfying the following three criteria: (1) included at least one vertically transmitted gene from Rafflesiaceae or Apodanthaceae; (2) included at least 10 species (23% species completeness); (3) all genes are orthologous to each other, and no paralogs generated by gene duplications prior to the divergence of Malpighiales were included. A detailed description of dataset assembly is provided in the supplementary note S1. All scripts used in this study are available on the GitHub repository (https://github.com/lmcai/Rafflesiaceae_phylogenomics).

Sequence alignment, cleaning, and gene tree reconstruction

Protein sequences of these 2,135 single-copy orthogroups were aligned using the E-INS-i algorithm implemented in MAFFT and then converted into the corresponding codon alignments using pal2nal (Suyama et al. 2006). Protein and DNA alignments were filtered using the profile hidden Markov models implemented in HmmCleaner v0.180750 (Di Franco et al. 2019). We applied a threshold of 50 and 10 to generate two sets of alignments with relaxed or stringent site filtering, respectively. In particular, a threshold of 50 can effectively remove nonhomologous regions while retaining fast-evolving sequences in Rafflesiaceae and Apodanthaceae. Consequently, 0.67% of the nucleotides and 0.94% of the amino acids from the raw alignments were flagged. In contrast, a threshold of 10 will remove these fast-evolving sequences and 2.4% of the nucleotides and 4.4% of the amino acids were flagged. We then used a custom Python script to mask the flagged regions and remove sites with >70% ambiguous characters (i.e., gaps and masked sites) while retaining codon positions for downstream analysis (available on the GitHub repository).

We inferred ML phylogenies for each locus based on five combinations of DNA partitioning strategies and substitution models using IQ-TREE v2.3.6 (Minh, Schmidt, et al. 2020). These ML inferences included (1) codon-based partition and forcing the GTR substitution model *without* rate heterogeneity across sites (i.e., no +I or +R or +G); (2) codon-based partition and forcing the GTR substitution model *with* rate heterogeneity across sites; (3) codon-

698 based partition and using the optimal substitution models determined by ModelFinder
699 (Kalyaanamoorthy et al. 2017); (4) no partition within loci and forcing the GTR substitution
700 model with rate heterogeneity across sites; (5) no partition within loci and using the
701 optimal substitution models determined by ModelFinder. For each ML analysis, branch
702 supports were assessed using ultrafast bootstrap approximation (UFBP) (Hoang et al. 2018)
703 with 1000 replicates along with the "-bnni" option to reduce the risk of overestimating
704 branch support resulting from model violations. We also used a non-parametric bootstrap
705 (BP) with 100 replications to assess branch support and generate BP gene trees for
706 coalescent analyses below. All DNA alignments, protein alignments, and individual gene
707 trees were deposited in Zenodo (10.5281/zenodo.11643025).
708
709 **Quantifying phylogenetic properties of individual genes**
710 Each gene was quantified for the following phylogenetic properties using IQ-TREE log files
711 and SortaDate (Smith et al. 2018): number of species, number of parsimony informative
712 sites, average gene tree support, total tree length, root-to-tip branch-length variance, and
713 gene tree–species tree congruence (Table S2). A species tree estimated by ASTRAL-IV
714 v1.19.4.5 (Zhang and Mirarab 2022) based on all 2,135 orthogroups (analysis 15 in Table1)
715 was used for assessing gene tree–species tree congruence, but we removed all
716 Rafflesiaceae and Apodanthaceae species to avoid the systematic bias stemming from
717 artificially fixing this unstable node. Four phylogenetic metrics specific to Rafflesiaceae
718 were extracted: the base compositional bias of Rafflesiaceae, minimum root-to-tip distance
719 among Rafflesiaceae species, the length of the short ascending parent branch of stem-group
720 Rafflesiaceae, and the branch support of Rafflesiaceae. We similarly defined and extracted
721 four Apodanthaceae-specific metrics. The base compositional heterogeneity of
722 Rafflesiaceae and Apodanthaceae was derived from the Chi-square test in IQ-TREE and
723 reported in its log file, whereas the other metrics were summarized using a custom Python
724 script from gene trees (available on GitHub). Finally, we modeled the pairwise Pearson
725 correlation among these phylogenetic metrics. The correlation heatmap was generated
726 using the function "heatmap" from the R package corrplot (Wei et al. 2017).
727
728 **Generation of phylogenetic hypotheses with data subsampling**

729 Twenty-two explorative phylogenomic analyses were conducted (Table1, Table S4). These
730 comprised concatenation (10 analyses) and coalescent methods (12 analyses) that used
731 either DNA (18 analyses) or protein alignments (4 analyses) for the complete dataset (17
732 analyses) or subsampled datasets (5 analyses). Two data subsets were created in an
733 attempt to select an *a posteriori* set of phylogenetic informative loci and to select clock-like
734 genes with less heterotachy to apply computationally intense mixture models. First, a 446–
735 locus dataset (referred to as G446) was selected from the complete dataset (referred to as
736 G2135) to include loci where Rafflesiaceae and Apodanthaceae formed a monophyletic
737 clade based on ML gene trees estimated with codon-based partition and optimal
738 substitution model determined by ModelFinder (Table S2; Data S1). Second, an 829–locus
739 dataset (referred to as G829) was created to select genes ranked in the top 50% by
740 SortaDate and included more than 32 (71%) species (Table S2; Data S1). A detailed
741 description of the command line used to perform the 22 analyses was provided on GitHub
742 (https://github.com/lmcai/Rafflesiaceae_phylogenomics/tree/master/4_phylogeny_hypo_
743 generation).
744
745 *Concatenation analysis*—We inferred the ML phylogeny based on the DNA and protein
746 alignments with various combinations of partition schemes and substitution models using
747 IQ-TREE. For all IQ-TREE inferences, branch support was evaluated using 1000 UFBP
748 replicates. Analyses 1–4 represented edge-linked partitioned DNA analyses with increasing
749 data filtering effort. The best partition model was determined by the greedy strategy from
750 ModelFinder implemented in IQ-TREE (Kalyaanamoorthy et al. 2017) where partitions
751 were iteratively merged until the model fit did not increase any further (-m MFP+MERGE).
752 Analyses 1 and 2 used DNA alignments filtered by HmmCleaner with a threshold of 50 and
753 a gene-based (Analysis 1) or codon-based (Analysis 2) initial partition. Analysis 3 used the
754 same DNA alignments and a codon-based initial partition, but with the third codon
755 removed. Analysis 4 used DNA alignments without the third codon, filtered by
756 HmmCleaner with a more stringent threshold of 10 and a codon-based partition. Analysis 5
757 implemented an edge-unlinked mixture GHOST model that can effectively handle
758 heterotachy by assigning sites into individual classes with unlinked model parameters and
759 edge lengths (Lopez et al. 2002; Crotty et al. 2020). We applied this GHOST model with four

760 mixture classes to the DNA alignment filtered by HmmCleaner with a threshold of 50 due to
761 the high RAM usage of this model. The last DNA concatenation (Analysis 6) used a
762 *posteriori* set of 446 loci (G446) where Rafflesiaceae and Apodanthaceae were
763 monophyletic in individual gene trees. The species tree was then inferred based on a gene-
764 based partition analysis in IQ-TREE.
765
766 Analysis 7-10 were protein-based analyses largely reflecting the data filtering strategies of
767 the DNA analyses. Analysis 7 and 8 applied gene-based partition schemes and protein
768 alignments filtered by HmmCleaner with a threshold of 50 and 10, respectively. The site-
769 specific profile mixture model implemented in Analysis 9 was applied to the subsampled
770 G829 dataset and 30 mixture components (LG+C30+F+G) due to the high RAM usage of this
771 model. This PMSF model approximates the empirical profile mixture model of Le et al.
772 (2008), which is a variant of the infinite mixture model CAT-GTR in PhyloBayes. It can
773 effectively ameliorate long-branch attraction artifacts and is computationally more efficient
774 (Wang et al. 2018). Finally, Analysis 10 applied the gene-based partition to the G446
775 dataset where Rafflesiaceae and Apodanthaceae were monophyletic.
776
777 *Coalescent analysis*—We used ASTRAL-IV and MP-EST v3.0 (Liu et al. 2010) to infer species
778 trees under the coalescent model. For both ASTRAL and MP-EST analyses, ML gene trees of
779 the G2135 dataset inferred from the five sets of partition schemes and substitution models
780 described above were used. Branch support was evaluated using the local posterior
781 probability in ASTRAL and 100 non-parametric bootstrapping in MP-EST. In addition to the
782 complete dataset, we also inferred species trees using the G446 dataset in both ASTRAL
783 and MP-EST.
784
785 **Evaluating gene and site phylogenetic conflicts**
786 We used gene and site concordance factors to quantify phylogenetic conflicts using IQ-
787 TREE. The gene concordance factor (gCF) is defined as the percentage of decisive gene
788 trees containing that branch (Minh, Hahn, et al. 2020). The site concordance factor (sCF)
789 was similarly defined but used a heuristic method to sample alignment sites while
790 minimizing the effects of homoplasy and taxon sampling (Mo et al. 2023). To calculate gCF

791 and sCF in IQ-TREE, we used the species tree, gene trees, and DNA alignments from the
792 coalescent-based Analysis 15 (Table 1). These gene trees were inferred under the optimal
793 substitution models determined by ModelFinder without codon-based partition. For sCF
794 calculation, 200 quartets were sampled to obtain stable site concordance factors (--scfl 200)
795 and we inferred sCF for both the concatenated DNA alignment and three codon positions
796 separately.
797
798 **Fit of gene trees and sites to proposed topologies**
799 The initial phylogenetic reconstructions revealed several possible placements of
800 Rafflesiaceae and Apodanthaceae associated with Euphorbiaceae+Peraceae, Putranjivaceae,
801 Pandaceae, Erythroxylaceae+Rhizophoraceae, Ixonanthaceae, and Linaceae. We tested the
802 fitness of individual genes and sites to the 8 alternative topologies commonly recovered
803 from our species inference analyses (Table 1) using the methods described in Shen et al.
804 (Shen et al. 2017).
805
806 Briefly, we conducted constrained phylogenetic searches enforcing the relationship
807 imposed by each of the 8 hypotheses but no constraints on the rest of the tree. Branch
808 length and the overall species tree topology were optimized using gene-based partition
809 (without merging) and substitution models determined by ModelFinder. Gene and site
810 likelihood scores for each tree were then obtained under the same partition and model
811 parameters from the previous step using the "-wpl" and "-wsl" flags in IQTREE, respectively.
812 The gene-specific rate was obtained from the IQTREE log file (under "Partition-specific
813 rates") and the site-specific rate was inferred using the "-wsr" option in IQTREE. These
814 gene and site likelihood scores and rates were reported in Data S3.
815
816 **Identify spurious relationships arising from fast-evolving genes**
817 To explore how fast-evolving genes may lead to spurious species tree reconstruction, we
818 first ordered the 2,135 loci based on the gene-specific rate for both DNA and protein
819 alignments (Data S3). We then examined the changes in cumulative difference of log
820 likelihoods (ΔLL) between the best-supported (H1 for DNA and H3 for protein) and
821 alternative topologies when more fast-evolving genes were added to the concatenated

822 matrix (Fig. 3). Results robust to evolutionary rates should demonstrate a steady increase
823 in ΔLL (e.g., Fig. 3C). In contrast, results sensitive to evolutionary rates are expected to have
824 ΔLL curves with less predictable trends and even sharp changes when the fastest-evolving
825 genes are added (e.g., Fig. 3B). To further evaluate the statistical significance of these
826 deviations, we generate 1000 bootstrap replicates where genes were randomly ordered
827 and plotted the same ΔLL curve. If the empirical curve falls out of the 95% confidence
828 interval, we consider a significant influence of rates to generate potentially spurious
829 relationships.
830
831 **Testing the anomaly zone**
832 The anomaly zone in a four-taxon species tree is defined by the length ratio of the two
833 internal branches (Degnan and Rosenberg 2006). To test if the focal clades fall within the
834 anomaly zone, we estimated the boundaries of the anomaly zone a(x) according to equation
835 4 from (Degnan and Rosenberg 2006). We used the phylogeny estimated by ASTRAL from
836 the complete dataset as the reference species tree (Analysis 15 in Table 1). Here, the three
837 internal nodes among the focal clades form three pairs of parent-descendent branches
838 whose topology aligns with the asymmetrical four-taxon phylogeny from (Degnan and
839 Rosenberg 2006). If the length of the descendant internal branch is smaller than a(x), then
840 the internode pair is in the anomaly zone.
841
842 **Coalescent simulation**
843 To further explore how ILS and homoplasy affect species tree reconstruction, we simulated
844 gene trees and DNA alignments based on the coalescent and mutational parameters from
845 the empirical dataset. The MP-EST species tree estimated from the complete dataset
846 (analysis 14 in Fig. 2) was used as the reference species tree and was pruned to include
847 only one species from each of the focal clades plus a non-Malpighiales outgroup,
848 *Crossopetalum*. We used the function "sim.coal.mpest" in the R package Phybase v2.0 (Liu
849 and Yu 2010) to simulate 10,000 gene trees using the reference species tree. We then
850 modified the branch lengths of the simulated gene trees to reflect the inter-species and
851 inter-locus rate variation in mutational units. To do so, we fit a Gaussian distribution to

852 each internal and terminal branch (e.g., terminal branches leading to *Sapria*, Rafflesiaceae)
853 from the empirical ML gene trees using function "fitdistr" in the R package MASS v7.3-54
854 (Ripley et al. 2013). We then assigned branch length in mutation units to the simulated
855 trees by randomly sampling the corresponding Gaussian distribution. Finally, we used the
856 parameters of the GTR+F+G4 model estimated by IQ-TREE based on the complete dataset
857 (analysis 1 in Fig. 2) to simulate DNA alignments in Seq-Gen v1.3.4 (Rambaut and Grassly
858 1997). For each gene, four alignments with different lengths (500, 1000, 1500, and 2000 bp)
859 were simulated to reflect high to low levels of gene tree estimation error (Mirarab et al.
860 2016; Cai et al. 2021b). We then inferred gene trees in IQ-TREE under the GTR+F+G4
861 model using 100 bootstrap replicates. Species trees were estimated using both
862 concatenation and coalescent methods. For concatenation analyses, we applied the
863 unpartitioned dataset in IQ-TREE under the GTR+F+G4 model. Branch support was
864 estimated by 1000 UFBP. For the coalescent method, we applied the ML gene trees in MP-
865 EST and ASTRAL. Branch support was estimated by 100 BP replicates. We also subsampled
866 the dataset to include 500, 1000, 2000, 3000, 4000, and 5000 loci with four alignment
867 lengths. Finally, we conducted AU tests in IQ-TREE and likelihood ratio tests in MP-EST to
868 statistically evaluate alternative topologies.
869
870 **Divergence time estimation**
871 To fully utilize our phylogenomic data for divergence time estimation, we applied both the
872 penalized likelihood method implemented in treePL v1.0 (Smith & O'Meara, 2012) and the
873 Bayesian method implemented in PhyloBayes v4.1 (Lartillot et al. 2009). For both analyses,
874 we fixed the topology using the species tree derived from the ASTRAL analysis 15 (Table 1).
875
876 For the treePL analysis, branch lengths of the species tree were inferred using the
877 concatenated DNA alignments, partition scheme, and substitution model from analysis 5
878 (Table 1) in IQ-TREE. This input alignment included all 2,135 loci, masked using
879 HmmCleaner with a threshold of 10, and excluded all third codon positions. We fixed the
880 root age at 118 Ma to reflect the crown group age of the Celastrales-Oxalidales-
881 Malpighiales clade from Ramírez-Barahona et al. (Ramirez-Barahona et al. 2020). In
882 addition, eight fossils in Malpighiales and one fossil from Oxalidales were used to constrain

883 the minimum age of the clade they each represented (Table S8). We then primed and cross-
884 validated our data to establish the best smoothing parameter (smooth = 0.01). Finally, we
885 conducted a thorough search to infer the best time tree.
886
887 For the PhyloBayes analysis, we similarly applied the concatenated DNA alignment without
888 the third codon position and masked with HmmCleaner using a threshold of 10. To ease
889 with convergence, we used the clock-like G829 loci. Divergence time estimation was
890 conducted under the log-normal autocorrelated relaxed clock model and the CAT Dirichlet
891 process mixture for among-site substitution heterogeneities. We constrained the root age
892 using a uniform distribution with a lower and upper bound of 89 and 125 Ma, respectively.
893 These boundaries were determined by the age of the oldest fossil *Palaeoclusia chevalieri*
894 (89 Ma) (Crepet and Nixon 1998) within Malpighiales and the oldest angiosperm
895 macrofossil *Archaefructus liaoningensis* at 125 Ma (Sun et al. 1998). For internal node age
896 calibrations, we used a uniform distribution for each fossil (Table S8) where the fossil's age
897 served as the minimum and 125 Ma as the maximum. We ran two Monte Carlo Markov
898 chains per model and kept every tenth sample until 30,000 samples were collected.
899 Convergence was checked by computing the discrepancies and effective sample sizes of the
900 posterior averages and visualizing the trend of the posterior. Consensus trees were
901 computed by removing 30% of the burn-in trees.
902
903 **Acknowledgements:**
904 We thank the two anonymous reviewers for their insightful comments on the initial version
905 of the manuscript. We thank Steven R Manchester, Peter Wilf, and members of the Davis lab
906 for feedback on early versions of this study.
907
908 **Funding**
909 This research was supported by startup funds from the University of Florida and the
910 Stengl-Wyer Postdoc Fellowship from the University of Texas at Austin to LC; National
911 Science Foundation (NSF) AToL grant (DEB-0622764), NSF grant DEB-1120243, and
912 startup funds from Harvard University to CCD.
913

914 **Data availability**

915 All DNA alignments, protein alignments, and individual gene trees were available on
916 Zenodo (10.5281/zenodo.11643025). All codes to conduct the analysis were available on
917 GitHub https://github.com/lmcai/Rafflesiaceae_phylogenomics.

918

919 **Reference**

920 Alzate JF, González FA, Pabón-Mora N. 2024. Back together: Over 1000 single-copy nuclear
921 loci and reproductive features support the holoendoparasitic Apodanthaceae and
922 Rafflesiaceae as sister lineages in the order Malpighiales. *Molecular Phylogenetics*
923 *and Evolution* 201:108217.

924 Amoo M, Salzmann U, Pound MJ, Thompson N, Bijl PK. 2022. Eocene to Oligocene
925 vegetation and climate in the Tasmanian Gateway region were controlled by
926 changes in ocean currents and pCO_2. *Climate of the Past* 18:525–546.

927 Baldwin BG. 1997. Adaptive radiation of the Hawaiian silversword alliance: congruence
928 and conflict of phylogenetic evidence from molecular and non-molecular
929 investigations. *Molecular evolution and adaptive radiation* 103:1840–1843.

930 Bansal M, Morley RJ, Nagaraju SK, Dutta S, Mishra AK, Selveraj J, Kumar S, Niyolia D, Harish
931 SM, Abdelrahim OB, et al. 2022. Southeast Asian Dipterocarp origin and
932 diversification driven by Africa-India floristic interchange. *Science* 375:455–460.

933 Barkman TJ, Klooster MR, Gaddis KD, Franzone B, Calhoun S, Manickam S, Vessabutr S,
934 Sasirat S, Davis CC. 2017. Reading between the vines: Hosts as islands for extreme
935 holoparasitic plants. *Am J Bot* 104:1382–1389.

936 Barkman TJ, Lim S-H, Salleh KM, Nais J. 2004. Mitochondrial DNA sequences reveal the
937 photosynthetic relatives of Rafflesia, the world's largest flower. *Proceedings of the*
938 *National Academy of Sciences* 101:787–792.

939 Bellot S, Renner S. 2014. The systematics of the worldwide endoparasite family
940 Apodanthaceae (Cucurbitales), with a key, a map, and color photos of most species.
941 *PhytoKeys* 36:41–57.

942 Bendiksby M, Schumacher T, Gussarova G, Nais J, Mat-Salleh K, Sofiyanti N, Madulid D,
943 Smith SA, Barkman T. 2010. Elucidating the evolutionary history of the Southeast
944 Asian, holoparasitic, giant-flowered Rafflesiaceae: Pliocene vicariance,
945 morphological convergence and character displacement. *Mol Phylogenet Evol*
946 57:620–633.

947 Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21:163–193.

Bossuyt F, Milinkovitch MC. 2001. Amphibians as Indicators of Early Tertiary "Out-of-India" Dispersal of Vertebrates. *Science* 292:93–95.

Brinkmann H, Van der Giezen M, Zhou Y, De Raucourt GP, Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic biology* 54:743–757.

Bromham L, Cowman PF, Lanfear R. 2013. Parasitic plants have increased rates of molecular evolution across all three genomes. *BMC Evol Biol* 13:1–11.

Bruno WJ, Halpern AL. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution* 16:564–566.

Cai L. 2023. Rethinking convergence in plant parasitism through the lens of molecular and population genetic processes. *American Journal of Botany* 110:e16174.

Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, Manickam S, Sasirat S, Nikolov LA, Mathews S. 2021. Deeply altered genome architecture in the endoparasitic flowering plant Sapria himalayana Griff.(Rafflesiaceae). *Current Biology* 31:1002-1011. e9.

Cai L, Xi Z, Amorim AM, Sugumaran M, Rest JS, Liu L, Davis CC. 2019. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytol* 221:565–576.

Cai L, Xi Z, Lemmon EM, Lemmon AR, Mast A, Buddenhagen CE, Liu L, Davis CC. 2020. 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.

Chen P, Wen J, Chen L. 2011. Spatial and temporal diversification of Tetrastigma (Vitaceae). *The Gardens' bulletin, Singapore* 63:307–327.

Chen X, Fang D, Xu Y, Duan K, Yoshida S, Yang S, Sahu SK, Fu H, Guang X, Liu M, et al. 2023. Balanophora genomes display massively convergent evolution with other extreme holoparasites and provide novel insights into parasite–host interactions. *Nat. Plants* 9:1627–1642.

Conti E, Eriksson T, Schönenberger J, Sytsma KJ, Baum DA. 2002. Early tertiary out-of-India dispersal of Crypteroniaceae: evidence from phylogeny and molecular dating. *Evolution* 56:1931–1942.

Crepet W, Nixon K. 1998. Fossil Clusiaceae from the Late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *Am J Bot* 85:1122–1122.

982 Crotty SM, Minh BQ, Bean NG, Holland BR, Tuke J, Jermiin LS, Haeseler AV. 2020. GHOST:
983 recovering historical signal from heterotachously evolved sequence alignments. *Syst*
984 *Biol* 69:249–264.

985 Davín AA, Tannier E, Williams TA, Boussau B, Daubin V, Szöllősi GJ. 2018. Gene transfers
986 can date the tree of life. *Nat Ecol Evol* 2:904.

987 Davis CC, Latvis M, Nickrent DL, Wurdack KJ, Baum DA. 2007. Floral gigantism in
988 Rafflesiaceae. *science* 315:1812–1812.

989 Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, Donoghue MJ. 2005. Explosive radiation of
990 Malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *Am*
991 *Nat* 165:E36–E65.

992 Davis CC, Wurdack KJ. 2004. Host-to-parasite gene transfer in flowering plants:
993 phylogenetic evidence from Malpighiales. *Science* 305:676–678.

994 Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene
995 trees. *PLOS Genet* 2:e68.

996 Di Franco A, Poujol R, Baurain D, Philippe H. 2019. Evaluating the usefulness of alignment
997 filtering methods to reduce the impact of errors on evolutionary inferences. *BMC*
998 *Evolutionary Biology* 19:21.

999 Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively
1000 misleading. *Syst Zool* 27:401–410.

1001 Filipowicz N, Renner SS. 2010. The worldwide holoparasitic Apodanthaceae confidently
1002 placed in the Cucurbitales by nuclear and mitochondrial gene trees. *BMC*
1003 *Evolutionary Biology* 10:219.

1004 Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-
1005 based phylogenetic reconstructions. *Journal of Molecular Evolution* 48:284–290.

1006 Givnish TJ, Millam KC, Mast AR, Paterson TB, Theim TJ, Hipp AL, Henss JM, Smith JF, Wood
1007 KR, Sytsma KJ. 2009. Origin, adaptive radiation and diversification of the Hawaiian
1008 lobeliads (Asterales: Campanulaceae). *Proceedings of the Royal Society B: Biological*
1009 *Sciences* 276:407–416.

1010 Glor RE. 2010. Phylogenetic Insights on Adaptive Radiation. *Annual Review of Ecology,*
1011 *Evolution, and Systematics* 41:251–270.

1012 González AD, Pabón-Mora N, Alzate JF, González F. 2020. Meristem genes in the highly
1013 reduced endoparasitic Pilostyles boyacensis (Apodanthaceae). *Frontiers in Ecology*
1014 *and Evolution* 8:209.

González F, Pabón-Mora N. 2017. Floral Development and Morphoanatomy in the Holoparasitic Pilostyles boyacensis (Apodanthaceae, Cucurbitales) Reveal Chimeric Half-Staminate and Half-Carpellate Flowers. *International Journal of Plant Sciences* 178:522–536.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* 29:644.

Grant PR, Grant BR. 2007. How and why species multiply: the radiation of Darwin's finches. Princeton University Press

Hendy MD, Penny D. 1989. A Framework for the Quantitative Study of Evolutionary Trees. *Systematic Biology* 38:297–309.

Ho SY, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. *Systematic Biology* 53:623–637.

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol* 35:518–522.

Huelsenbeck JP, Hillis DM. 1993. Success of Phylogenetic Methods in the Four-Taxon Case. *Systematic Biology* 42:247–264.

Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14:587.

Kapli P, Yang Z, Telford MJ. 2020. Phylogenetic tree building in the genomic age. *Nat Rev Genet* 21:428–444.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30:772–780.

Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol* 31:151–160.

Knight RD, Freeland SJ, Landweber LF. 2001. Rewiring the keyboard: evolvability of the genetic code. *Nature Reviews Genetics* 2:49–58.

Koch NM. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. *Molecular biology and evolution*.

Kuijt J. 1969. The biology of parasitic flowering plants. Berkeley, CA: University of California Press

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.

1048 Liu L, Yu L. 2010. Phybase: an R package for species tree analysis. *Bioinformatics* 26:962–
1049 963.

1050 Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating
1051 species trees under the coalescent model. *BMC Evol Biol* 10:302.

1052 Liu L, Zhang J, Rheindt FE, Lei F, Qu Y, Wang Y, Zhang Y, Sullivan C, Nie W, Wang J, et al.
1053 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by
1054 the KPg boundary. *Proceedings of the National Academy of Sciences* 114:E7282–
1055 E7290.

1056 Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein
1057 evolution. *Molecular biology and evolution* 19:1–7.

1058 Losos JB. 2011. Lizards in an evolutionary tree: ecology and adaptive radiation of anoles.
1059 Univ of California Press

1060 Losos JB, Mahler DL. 2010. Adaptive radiation: the interaction of ecological opportunity,
1061 adaptation, and speciation. In: Evolution since Darwin: The First 150 Years. Sinauer
1062 Associates. p. 381–420.

1063 Malabrigo Jr. P, Tobias AB, Witono J, Mursidawati S, Susatya A, Siti-Munirah MY, Wicaksono
1064 A, Raihandhany R, Edwards S, Thorogood CJ. 2025. Most of the world's largest
1065 flowers (genus Rafflesia) are now on the brink of extinction. *PLANTS, PEOPLE,*
1066 *PLANET* 7:331–346.

1067 Manchester SR, Kapgate DK, Wen J. 2013. Oldest fruits of the grape family (Vitaceae) from
1068 the Late Cretaceous Deccan cherts of India. *Am J Bot* 100:1849–1859.

1069 Martyn I, Steel M. 2012. The impact and interplay of long and short branches on
1070 phylogenetic information content. *Journal of Theoretical Biology* 314:157–163.

1071 Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for
1072 phylogenomic datasets. *Molecular biology and evolution* 37:2727–2733.

1073 Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear
1074 R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in
1075 the genomic era. *Mol Biol Evol* 37:1530–1534.

1076 Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search:
1077 HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*
1078 41:e121–e121.

1079 Mo YK, Lanfear R, Hahn MW, Minh BQ. 2023. Updated site concordance factors minimize
1080 effects of homoplasy and taxon sampling. *Bioinformatics* 39:btac741.

1081 Naumann J, Salomo K, Der JP, Wafula EK, Bolin JF, Maass E, Frenzke L, Samain M-S, Neinhuis
1082 C, dePamphilis CW, et al. 2013. Single-Copy Nuclear Genes Place Haustorial
1083 Hydnoraceae within Piperales and Reveal a Cretaceous Origin of Multiple Parasitic
1084 Angiosperm Lineages. *PLOS ONE* 8:e79204.

1085 Nickrent DL, Blarer A, Qiu Y-L, Vidal-Russell R, Anderson FE. 2004. Phylogenetic inference
1086 in Rafflesiales: the influence of rate heterogeneity and horizontal gene transfer. *BMC*
1087 *Evol Biol* 4:40.

1088 Nickrent DL, Duff RJ. 1996. Molecular studies of parasitic plants using ribosomal RNA.
1089 *Advances in parasitic plant research*:28–52.

1090 Nickrent DL, Starr EM. 1994. High rates of nucleotide substitution in nuclear small-subunit
1091 (18S) rDNA from holoparasitic flowering plants. *Journal of Molecular Evolution*
1092 39:62–70.

1093 Nie Z-L, Sun H, Manchester SR, Meng Y, Luke Q, Wen J. 2012a. Evolution of the
1094 intercontinental disjunctions in six continents in the Ampelopsis clade of the grape
1095 family (Vitaceae). *BMC evolutionary biology* 12:1–13.

1096 Nie Z-L, Sun H, Manchester SR, Meng Y, Luke Q, Wen J. 2012b. Evolution of the
1097 intercontinental disjunctions in six continents in the Ampelopsis clade of the grape
1098 family (Vitaceae). *BMC Evolutionary Biology* 12:17.

1099 Nikolov LA, Davis CC. 2017. The big, the bad, and the beautiful: Biology of the world's
1100 largest flowers. *J Syst Evol* 55:516–524.

1101 Nikolov LA, Endress PK, Sugumaran M, Sasirat S, Vessabutr S, Kramer EM, Davis CC. 2013.
1102 Developmental origins of the world's largest flowers, Rafflesiaceae. *Proceedings of*
1103 *the National Academy of Sciences* 110:18578–18583.

1104 Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree
1105 inference and accurate estimates of substitution rates. *Mol Biol Evol* 34:2101–2114.

1106 Parey E, Louis A, Montfort J, Bouchez O, Roques C, Iampietro C, Lluch J, Castinel A,
1107 Donnadieu C, Desvignes T. 2023. Genome structures resolve the early diversification
1108 of teleost fishes. *Science* 379:572–575.

1109 Pelser PB, Nickrent DL, van Ee BW, Barcelona JF. 2019. A phylogenetic and biogeographic
1110 study of Rafflesia (Rafflesiaceae) in the Philippines: limited dispersal and high island
1111 endemism. *Mol Phylogenet Evol* 139:106555.

1112 Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D.
1113 2011. Resolving difficult phylogenetic questions: why more sequences are not
1114 enough. *PLOS Biol* 9:e1000602.

1115 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees
1116 for large alignments. *PloS one* 5:e9490.

1117 Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of
1118 DNA sequence evolution along phylogenetic trees. *Computer applications in the*
1119 *biosciences: CABIOS* 13:235–238.

1120 Ramirez-Barahona S, Sauquet H, Magallon S. 2020. The delayed and geographically
1121 heterogeneous diversification of flowering plant families. *Nat Ecol Evol* 4:1232–
1122 1238.

1123 Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB. 2013. Package
1124 'mass.' *Cran r* 538:113–120.

1125 Roch S, Nute M, Warnow T. 2019. Long-branch attraction in species tree estimation:
1126 inconsistency of partitioned likelihood and topology-based summary methods. *Syst*
1127 *Biol* 68:281–297.

1128 Schluter D. 2000. The Ecology of Adaptive Radiation. Oxford, New York: Oxford University
1129 Press

1130 Schluter D, McPhail JD. 1992. Ecological character displacement and speciation in
1131 sticklebacks. *Am Nat* 140:85–108.

1132 Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies
1133 can be driven by a handful of genes. *Nat Ecol Evol* 1:0126.

1134 Smith SA, Brown JW, Hinchliff CE. 2013. Analyzing and synthesizing phylogenies using tree
1135 alignment graphs. *PLOS Comput Biol* 9:e1003223.

1136 Smith SA, Brown JW, Walker JF. 2018. So many genes, so little time: A practical approach to
1137 divergence-time estimation in the genomic era. *PLOS One* 13:e0197433.

1138 Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood
1139 for large phylogenies. *Bioinformatics* 28:2689–2690.

1140 Stiller J, Feng S, Chowdhury A-A, Rivas-González I, Duchêne DA, Fang Q, Deng Y, Kozlov A,
1141 Stamatakis A, Claramunt S, et al. 2024. Complexity of avian evolution revealed by
1142 family-level genomes. *Nature* 629:851–860.

1143 Sun G, Dilcher DL, Zheng S, Zhou Z. 1998. In search of the first flower: a Jurassic angiosperm,
1144 Archaefructus, from northeast China. *Science* 282:1692–1695.

1145 Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence
1146 alignments into the corresponding codon alignments. *Nucleic Acids Res* 34:W609–
1147 W612.

1148 Tateno Y, Takezaki N, Nei M. 1994. Relative efficiencies of the maximum-likelihood,
1149 neighbor-joining, and maximum-parsimony methods when substitution rate varies
1150 with site. *Molecular biology and evolution* 11:261–277.

1151 Teixeira-Costa L, Davis CC. 2021. Life history, diversity, and distribution in parasitic
1152 flowering plants. *Plant Physiol* 187:32–51.

1153 Thorogood CJ, Teixeira-Costa L, Ceccantini G, Davis C, Hiscock SJ. 2021. Endoparasitic
1154 plants and fungi show evolutionary convergence across phylogenetic divisions. *New*
1155 *Phytol* 232:1159–1167.

1156 Walker JF, Smith SA. 2023. Phylogenomic conflicts in plants. In: Species tree inference: A
1157 guide to methods and applications. Princeton, New Jersey: Princeton University
1158 Press. p. 149–160.

1159 Wang H, Singer GA, Hickey DA. 2004. Mutational bias affects protein evolution in flowering
1160 plants. *Molecular biology and evolution* 21:90–96.

1161 Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. 2017. Package 'corrplot.' *Statistician* 56:e24.

1162 Wicke S, Müller KF, Quandt D, Bellot S, Schneeweiss GM. 2016. Mechanistic model of
1163 evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants.
1164 *Proceedings of the National Academy of Sciences*:201607576.

1165 Wurdack KJ, Davis CC. 2009. Malpighiales phylogenetics: gaining ground on one of the most
1166 recalcitrant clades in the angiosperm tree of life. *Am J Bot* 96:1551–1570.

1167 Xi Z, Bradley RK, Wurdack KJ, Wong KM, Sugumaran M, Bomblies K, Rest JS, Davis CC. 2012.
1168 Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics*
1169 13:227.

1170 Xi Z, Wang Y, Bradley RK, Sugumaran M, Marx CJ, Rest JS, Davis CC. 2013. Massive
1171 mitochondrial gene transfer in a parasitic flowering plant clade. *PLOS Genet*
1172 9:e1003265.

1173 Xiong H, Wang D, Shao C, Yang X, Yang J, Ma T, Davis CC, Liu L, Xi Z. 2022. Species Tree
1174 Estimation and the Impact of Gene Loss Following Whole-Genome Duplication.
1175 *Systematic Biology* 71:1348–1361.

1176 Yang Y, Smith SA. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes
1177 and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for
1178 Phylogenomics. *Mol Biol Evol* 31:3081–3092.

1179 Young ND, dePamphilis CW. 2005. Rate variation in parasitic plants: correlated and
1180 uncorrelated patterns among plastid genes of different function. *BMC Evolutionary*
1181 *Biology* 5:16.

1182 Zhang C, Mirarab S. 2022. Weighting by gene tree uncertainty improves accuracy of
1183 quartet-based species trees. *Molecular Biology and Evolution* 39:msac215.

1184 Zhao T, Zwaenepoel A, Xue J-Y, Kao S-M, Li Z, Schranz ME, Van de Peer Y. 2021. Whole-
1185 genome microsynteny-based phylogeny of angiosperms. *Nat Commun* 12:3498.

1186 Zharkikh A, Li W-H. 1993. Inconsistency of the maximum-parsimony method: the case of
1187 five taxa with a molecular clock. *Systematic Biology* 42:113–125.

1188