The danger zone: the joint trap of incomplete lineage sorting and long-branch attraction in placing Rafflesiaceae

Liming Cai[1,2], Liang Liu[3], Charles C. Davis[2]

[1]Department of Integrative Biology, the University of Texas at Austin, Austin, TX 78712, USA

[2]Department of Organismal Biology, Harvard University Herbaria, Harvard University, Cambridge, MA 02138, USA

[3]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 anomaly zone—and erroneous gene tree estimation—commonly manifested by long-branch attraction in the Felsenstein zone. Seldom have these factors been analyzed when they occur simultaneously as in cases involving ancient adaptive radiations. Rafflesiaceae (Malpighiales) is an iconic holoparasitic plant clade growing 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 in Malpighiales. To complicate matters, parasitic plants like Rafflesiaceae are well-known to exhibit accelerated rates of nucleotide substitution. Here, using 2,141 genes we identify for the first time that accelerated substitution rates is a genome-wide phenomenon in holoparasites. We demonstrate two orders of magnitude increase in substitution rates in Rafflesiaceae compared to its free-living relatives, rendering nearly every gene prone to long-branch attraction. Our analyses rejected the monophyletic (Rafflesiaceae, Euphorbiaceae) clade proposed in previous studies but favored an earlier divergence of Rafflesiaceae in close affinity with Euphorbiaceae, Peraceae, Putranjivaceae, and Pandaceae. Rafflesiaceae is therefore prey to a joint trap of the anomaly and Felsenstein zones. Moreover, owing to massive gene loss

within Rafflesiaceae, our simulation suggests that this represents an intractable phylogenetic problem that will never achieve a single phylogenetic resolution. Our case study highlights a seemingly insurmountable confluence of factors—a danger zone—that hinders phylogenomic resolution and demonstrates the need to be flexible in cases where multiple phylogenomic placements are equally likely.

Keywords

Adaptive radiation, homoplasy, anomaly zone, composition bias, heterotachy, PhyloBayes, gene genealogy interrogation

Introduction

Adaptive radiation is one of the most fascinating problems 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. Well-known examples of adaptive radiations include Galapagos finches, Caribbean anoles, Hawaiian silverswords and lobeliads, and cloud forest *Viburnum* (Baldwin 1997; Grant and Grant 2007; Givnish et al. 2009; Losos 2011; Donoghue et al. 2022). Despite the intriguing nature of adaptive radiations, such clades often present challenges for phylogenetic inference. These clades often exhibit a combination of short and long branches resulting from extensive accumulation of mutations, which is a particularly vexing problem (Glor 2010). Species trees of clades that have undergone 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 descendent branches in these radiations can lead to various systematic errors, such as long-branch attraction and compositional bias (Kapli et al. 2020).

All these factors are further confounded when the radiation is particularly ancient, as in the case of Malpighiales—a Cretaceous explosive radiation that has produced plants with diverse morphologies that range from temperate deciduous trees (Salicaceae) and tropical mangroves (Rhizophoraceae) to submerged tropical aquatics (Podostemaceae; Davis et al.

2005). Among the strangest of all malpighialean members is the holoparasitic family Rafflesiaceae, which produces the world's largest flower (Nikolov and Davis 2017). Furthermore, the family is atypical even compared to most other parasitic plants. Species in Rafflesiaceae are endoparasites that live out their vegetative life hidden entirely inside their grapevine hosts; their flowers only emerge for sexual reproduction (*Tetrastigma*, Vitaceae; Kuijt 1969; Nikolov et al. 2013; Nikolov and Davis 2017). Endoparasitism has given rise to a series of distinct morphological modifications that have made their taxonomy controversial (Teixiera-Costa and Davis, 2021). For example, the vegetative organs of Rafflesiaceae are highly reduced and persist only as strands of cells embedded within their liana host *Tetrastigma* (Nikolov et al. 2014). Other plant endoparasites share similar modifications and thus it is not surprising that Cronquist grouped distantly related parasitic families such as Hydnoraceae and Mitrastemonaceae together with Rafflesiaceae (Barkman et al. 2004; Cronquist 1981; Davis and Wurdack 2004; Davis et al. 2007; Wurdack and Davis 2009).

Today, Rafflesiaceae sensu stricto includes only large-flowered endoparasites and are placed in Malpighiales (Barkman et al. 2004; Davis and Wurdack 2004; Davis et al. 2007; Wurdack and Davis 2009) as sister to Euphorbiaceae (Davis et al. 2007; Bendiksby et al. 2010), highlighting a prominent contrast in floral size between these two sister clades. Despite strong support for the placement of Rafflesiaceae as a close relative of Euphorbiaceae sensu lato, this placement was based only on five mitochondrial and nuclear genes. One recent attempt using the Angiosperm353 sequence capture probe set, on the other hand, did not confidently resolve its placement within Malpighiales (<0.5 posterior probability; Zuntini et al. 2024). The complete loss of the entire chloroplast genome in Rafflesiaceae also ruled out the potential of this widely used phylogenetic marker to resolve its placement (Molina et al. 2014; Cai et al. 2021a).

The recently published genome of *Sapria himalayana* Griff. (Cai et al. 2021a), together with the companion transcriptome data from the related genera *Rafflesia* and *Rhizanthes*, offered the hope of expanding the genomic toolkit for resolving the placement of Rafflesiaceae. However, these resources highlight four genomic aspects of the clade that

93 further confound their phylogenetic placement. First, nearly half of the conserved genes in
94 Rafflesiaceae are lost (Cai et al. 2021a), which greatly restricts the repertoire of gene
95 candidates for phylogenomic research. These losses are overrepresented in genes
96 corresponding to conserved functions such as photosynthesis and plastid organization,
97 which are common targets of phylogenomic markers. Second, members of Rafflesiaceae,
98 like numerous other parasitic plants, are known to acquire foreign genetic materials from
99 their hosts via horizontal gene transfer (HGT; Davis and Wurdack 2004; Xi et al. 2012; Xi et
100 al. 2013), which can confound species tree inference. Ancestral genes acquired by HGT that
101 are shared by all Rafflesiaceae are useful for resolving within-family relationships as
102 exemplified by the application of 16S in Barkman et al. (2008) and Bendiksby et al. (2010),
103 but such genes are inappropriate for placing Rafflesiaceae within Malpighiales. Third, the
104 well-characterized high substitution rate in Rafflesiaceae may lead to spurious species tree
105 reconstruction because of long-branch attraction (Nickrent and Starr 1994; Nickrent and
106 Duff 1996). This phenomenon was demonstrated in all three cellular genomes using
107 representative genes across multiple independently evolved parasitic lineages (Bromham
108 et al. 2013; Young and dePamphilis 2005), but to what extent this phenomenon spans the
109 entire nuclear genome has not previously been tested in Rafflesiaceae. Fourth, the unusual
110 AT-rich (GC% = 24%) genome of Rafflesiaceae likely contributes to compositional bias that
111 may incorrectly group distantly related taxa as close relatives (Foster and Hickey 1999;
112 Knight et al. 2001; Wang et al. 2004).
113
114 The ancient radiation of Malpighiales includes nine of the ten most difficult angiosperm
115 clades to date (Smith et al. 2013), easily placing its members as the most extraordinary and
116 difficult phylogenomic problem to resolve in angiosperms. For Rafflesiaceae, the potentially
117 greatly accelerated substitution rates combined with the aforementioned genomic
118 anomalies attributable to its parasitic modality likely make this an even more intractable
119 phylogenomic problem. Rafflesiaceae falls into a joint trap that we term 'the danger zone' of
120 ILS and long-branch attraction. The 'danger zone' represents a toxic combination for
121 achieving a tidy phylogenetic resolution. Here, using 2,141 genes including a broad
122 sampling of Malpighiales, we tackle the joint problem of ILS and long-branch attraction

using a combination of approaches that accommodate genealogical heterogeneity among loci and rate heterogeneity among sites.

RESULTS

Phylogenomic datasets

By combining data from two published genome and transcriptome studies within Malpighiales (Cai et al., 2019; Cai et al., 2021a), we generated a phylogenomic dataset of 2,141 loci, each containing at least 10 out of 43 species in Malpighiales (supplementary table S1, Supplementary Material online). A phylogenetic based approach from Yang and Smith (2014) was applied to remove paralogs and potential HGTs. These 43 species included all three genera of Rafflesiaceae (*Sapria, Rafflesia,* and *Rhizanthes*) and three outgroup species from Oxalidales and Celastrales. The concatenated DNA matrix consisted of 3,228,591 sites with a gene and character occupancy of 70.9% and 81.8%, respectively. Two species exhibiting exceptional levels of missing data were *Galearia maingayi* (88.1% missing) and *Ixonanthes reticulata* (79.5% missing). Each locus on average contained 31 species and the median length of DNA alignment was 1,275 bp after trimming. The mean branch support for maximum likelihood (ML) gene trees was 68.4 when using ultrafast bootstrap (UFBP) and 54.2 using traditional non-parametric bootstrap (BP), but the mean support for the placement of Rafflesiaceae was only 49.2 UFBP or 26.3 BP. We also created two additional datasets, namely G1553 and G647, with increasingly stringent filtering criteria for missing data. The G1553 dataset included 1,553 loci with less than 14 missing species (33%). The concatenated matrix of G1553 included 2,139,714 DNA sites and a mean branch support of 68.6 UFBP. The G647 dataset included 647 loci with less than 9 missing species (20%). The concatenated matrix of G647 included 846,640 DNA sites and a similar mean branch support of 68.2 UFBP. All curated alignments and gene trees are available on Zenodo (DOI: 10.5281/zenodo.11643026).

Identifying suitable genes for placing Rafflesiaceae

To identify suitable genes for phylogenetic reconstruction, especially regarding the placement of Rafflesiaceae, we used ten metrics to quantify various aspects of phylogenetic properties (supplementary table S2, Supplementary Material online). These metrics

included six tree-wise features such as number of species, alignment length, average nodal support, total tree length, root-to-tip branch length variance, and gene tree–species tree congruence; and four metrics specific to Rafflesiaceae, including its DNA compositional bias, branch lengths, and nodal support. These ten metrics were significantly correlated (*p*-value <0.05) in 24 out of the 45 comparisons (Fig. 1). For example, the mean nodal support, which is a widely applied metric for selecting high-quality phylogenetic markers, was significantly positively correlated with alignment length and gene tree–species tree congruence. Rafflesiaceae-specific features, however, largely formed a unique covarying cluster in the correlation heatmap and were relatively independent of the other tree-wise metrics (supplementary fig. S1, Supplementary Material online). For example, the nodal support of Rafflesiaceae's placement was strongly positively correlated with its own branch length (*p*-value < 2.2e-16, Pearson correlation coefficient = 0.48), but counterintuitively, negatively correlated with the number of species and gene tree–species tree congruence (Fig. 1).

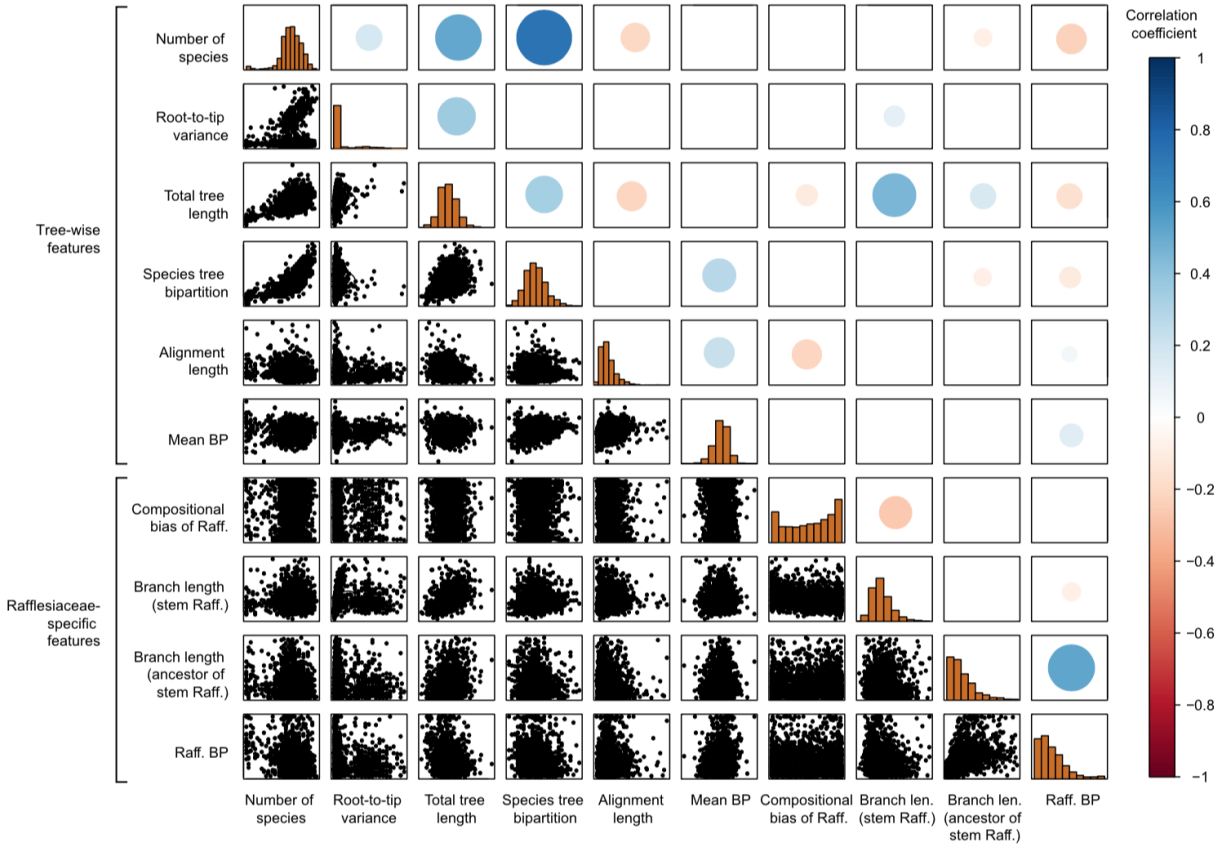

Fig. 1 Pairwise correlation between ten metrics of gene phylogenetic property. The scatter plots on the lower left depict the raw distribution between any pair of metrics; the circles on the upper right represent the significance and direction of Pearson correlation tests. The size of the circles is negatively scaled with the *p*-value in log scale, and the red-to-blue hue of the Pearson correlation coefficient represents the negative-to-positive correlation.

Initial phylogenetic reconstructions yielded weakly supported, conflicting species trees

Explorative phylogenetic reconstructions using concatenation and coalescent methods recovered multiple distinct placements of Rafflesiaceae (referred to as H1-H19 in Fig. 2B; supplementary table S3; supplementary data S1, Supplementary Material online). Despite the effort to reduce missing data, mitigate compositional and rate heterogeneity, and accommodate ILS, none of these results were well supported. Two topologies were frequently recovered, which either placed Rafflesiaceae as sister to Putranjivaceae alone (H2) or as sister to the most recent common ancestor (MRCA) of Putranjivaceae, Pandaceae, Rhizophoraceae, and Erythroxylaceae (H11). The monophyletic (Rafflesiaceae, Putranjivaceae) (H2) was supported by ML analysis of the concatenated DNA alignments with the least amount of missing data (G647) and the full coalescent analyses from starBEAST. For ML analyses, this topology received 67 BP support in RAxML and 87 UFBP support in IQ-TREE. The Bayesian inferences in starBEAST grouped these two families with 0.98 PP, but the relationship among (Rafflesiaceae, Putranjivaceae), (Peraceae, Euphorbiaceae), and (Erythroxylaceae, Rhizophoraceae, and Pandaceae) remained elusive (PP< 0.53; supplementary fig. S2, Supplementary Material online). On the other hand, an earlier divergence of Rafflesiaceae featured in H11 was supported by complex site models such as the PhyloBayes infinite profile mixture model (0.86 PP) and the empirical protein profile mixture model (69 BP), and most summary coalescent analyses. The other two topologies, H15 and H18, were also recovered by more than one analysis. H15 placed Rafflesiaceae sister to the MRCA of Euphorbiaceae, Peraceae, Rhizophoraceae, Erythroxylaceae, Pandaceae, and Putranjivaceae. It was recovered by the concatenated Dayhoff six-state recoded protein alignment (analysis 11) and two summary coalescent analyses with subsampled gene trees (analysis 16 and 17). H18 placed Rafflesiaceae with

200 the MRCA of Ixonanthaceae, Rhizophoraceae, and Erythroxylaceae, which was supported
201 by the DNA alignment of the G1553 dataset (UFBP = 32 in IQ-TREE and BP = 47 in RAxML).
202
203 Despite these conflicts, the initial analyses reveal four clades that are frequently associated
204 with Rafflesiaceae, including (Euphorbiaceae, Peraceae), Putranjivaceae, Pandaceae, and
205 (Erythroxylaceae, Rhizophoraceae). The divergence between these clades was
206 characterized by a series of short internal branches, leading to extensive phylogenetic
207 conflicts (Fig. 2A). For the four internal nodes among these clades plus Rafflesiaceae, the
208 concordant gene quartet frequencies, defined as the percentage of decisive gene trees
209 supporting that branch, ranged from 34.8% to 40.7% (Minh et al. 2020a). The quartet
210 frequencies for other early-diverging lineages in Malpighiales were similarly low and
211 ranged from 34.8% to 41.0%, reflecting the well-recognized recalcitrant relationship in
212 Malpighiales. Similarly, site concordance factor summarized from DNA and protein
213 alignments also suggested strongly conflicting signals. Among the five focal clades, the site
214 concordance factors ranged from 30.6 to 40.8% for DNA and from 30.6 to 38.8% for
215 protein (Fig. 2A). The GC content of Rafflesiaceae was comparable to other free-living
216 Malpighiales and did not cause the low concordance factors observed above
217 (Supplementary fig. S3, Supplementary Material online).
218

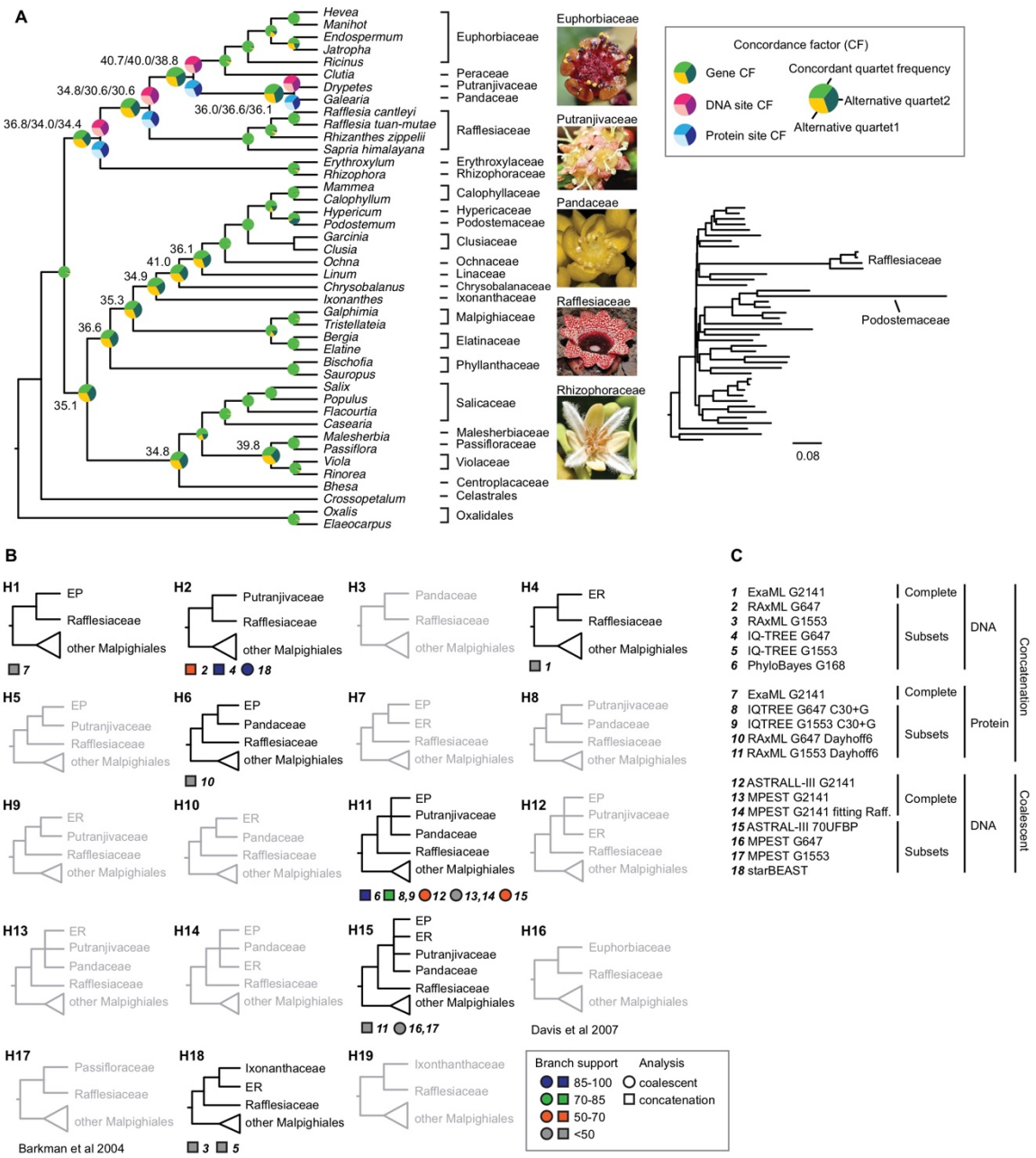

Fig. 2 Extensive phylogenetic conflicts lead to weakly supported conflicting placements of Rafflesiaceae. **A**. Conflicting phylogenetic signals evidenced by gene and site concordance factors. The reference species tree is inferred by MP-EST from the complete dataset (analysis 14 in **C**). The numbers on each node illustrate the proportions of concordant versus discordant genes. These concordance factors (CFs) are also visualized as pie charts for genes, DNA sites, and protein sites. **B**. Alternative placement of Rafflesiaceae across various data subsampling and species tree inference schemes. The 19 hypothesized placements of Rafflesiaceae are variously supported by coalescent or concatenation analyses and their branch support values are color coded. **C**. Method summary for 18 explorative phylogenetic analyses.

230 **Modeling is more effective than masking in mitigating site heterogeneity**

231 Given the exceptionally long branch leading to stem group Rafflesiaceae (Fig. 2), we
232 employed two divergent strategies to mitigate the effects of compositional and rate
233 heterogeneity. We first applied the CAT-GTR model in PhyloBayes v3.1.1 to a smaller
234 dataset of 168 genes, which provides infinite mixture models of site-specific evolutionary
235 profiles (Lartillot et al. 2013). Two out of four Markov chains converged with a maximum
236 discrepancy of 0.15 after running for 41,253 generations, which took more than one month
237 and 32 nodes on the high-performance computing cluster. This phylogeny placed
238 Rafflesiaceae as sister to the MRCA of Euphorbiaceae, Peraceae, Pandaceae, and
239 Putranjivaceae (H11 in Fig. 2B) with moderate support (0.86 PP). However, because of the
240 low support of its immediate descendant branch (0.58 PP), these three clades—including
241 Rafflesiaceae, (Pandaceae, Putranjivaceae), and (Euphorbiaceae, Peraceae)—formed a
242 polytomy. Moreover, the branch length of stem group Rafflesiaceae was inferred to be
243 0.543 and its ancestral branch to be 0.0008. This result suggested a 679-fold change in
244 branch length before and after the divergence of Rafflesiaceae. Despite the even longer
245 terminal branch leading to Podostemaceae and high GC content, the phylogenetic affinity
246 between Podostemaceae and Hypericaceae was maximumly supported in all analyses due
247 to the longer internal branch supporting their shared ancestry (Fig. 2A; supplementary fig.
248 S3; supplementary data S1, Supplementary Material online).
249
250 Second, we applied the six-state Dayhoff amino acid recoding (Dayhoff et al. 1978) to mask
251 mutational noises. This amino acid recoding did not effectively resolve the placement of
252 Rafflesiaceae, nor did it improve the overall resolution of the species tree. The Dayhoff6
253 recoding allowed us to use more loci compared to the mixture models in PhyloBayes.
254 Species tree inferred from the recoded G647 dataset placed Rafflesiaceae with 12 BP as
255 sister to the MRCA of Euphorbiaceae, Peraceae, and Pandaceae (H6 in Fig. 2B). Results from
256 the recoded G1553 dataset placed Rafflesiaceae with 19 BP as sister to the MRCA of
257 Euphorbiaceae, Peraceae, Putranjivaceae, Pandaceae, Erythroxylaceae, and Rhizophoraceae
258 (H15 in Fig. 2B).
259
260 **Topology tests supported conflicting hypotheses**

261 We tested the goodness-of-fit of individual genes and concatenated alignments to proposed
262 species trees under the concatenation and coalescent models.
263
264 First, we applied the Gene Genealogy Interrogation (GGI) implemented in (Arcila et al.
265 2017) to assess the fit of individual genes to the proposed topologies. Only three out of the
266 2,141 genes significantly supported one of the hypotheses in the AU test (*p*-value >0.95).
267 The monophyly of Rafflesiaceae and Putranjivaceae (H2) was slightly favored over other
268 hypotheses (supplementary fig. S4, Supplementary Material online). Results from the KH
269 test and SH test were similar; very few genes contained sufficient information to
270 significantly support any of the hypotheses (supplementary fig. S4, Supplementary Material
271 online).
272
273 Second, we assessed topologies using concatenated DNA and protein alignments. The
274 results were sensitive to the type of input data (DNA vs. protein), the amount of missing
275 data (G647 vs. G1553), and even the phylogenetic inference program (RAxML vs. IQ-TREE).
276 Ten hypotheses (H1, H5, H7–10, H12–14, and H16; supplementary table S4, Supplementary
277 Material online) were universally rejected by all tests, including the monophyly of
278 (Rafflesiaceae, Euphorbiaceae). Four hypotheses (H2, H4, H6, and H11) were supported by
279 at least two topology tests (AU test *p*-value >0.05, RELL bootstrap >0.05). The monophyly
280 of Rafflesiaceae and Putranjivaceae (H2) was supported by DNA alignments under the GTR
281 model, but strongly rejected using protein alignments under the C30 profile mixture model.
282 The close relationship between Rafflesiaceae and (Erythroxylaceae, Rhizophoraceae) (H4)
283 was similarly supported by DNA alignments but not protein data. We also obtained
284 inconsistent results when using RAxML or IQ-TREE, despite using the same input data and
285 substitution model. For example, the AU test in IQ-TREE accepted H4 (*p*-value = 0.544)
286 when applying the G1553 DNA alignment. But this hypothesis was rejected with a *p*-value
287 of 0.03 when using the likelihood calculated by RAxML (supplementary table S4,
288 Supplementary Material online).
289
290 Finally, we evaluated alternative species trees under the multiple species coalescent model.
291 The optimum tree placed Rafflesiaceae as sister to the MRCA of all four focal clades (H15)

292 with a pseudolikelihood of -7433872.434 in MP-EST. The second-best tree was H11,
293 followed by H12 as the third-best tree. Using parametric bootstrapping, we found that the
294 difference in pseudolikelihoods was not statistically significant among the top three (*p*-
295 value ≥ 0.05) but were all significantly better than the fourth tree. Therefore, our likelihood
296 ratio test rejected all but three hypotheses (H11, H12, H15) under the coalescent model
297 (supplementary table S4, Supplementary Material online).
298
299 **Long-branch attraction contributes to erroneous species tree reconstruction**
300 To further explore how fast-evolving genes prone to long-branch attraction may contribute
301 to the conflicting placements of Rafflesiaceae, we partitioned the data into eight subsets
302 with increasing synonymous substitution rates (d_S) in stem group Rafflesiaceae. The
303 median d_S value of stem group Rafflesiaceae ranged from 0.47 to 0.85 in these datasets and
304 35.1% of genes have experienced more than one synonymous substitution per site (d_S >1;
305 Fig. 3B). These substitution rates were 4.6 times higher than in *Manihot* (Euphorbiaceae)
306 and 3.3 times higher than in *Populus* (Salicaceae; supplementary fig. S5, Supplementary
307 Material online). Topology tests on these data subsets recaptured the results above that
308 DNA and protein support different hypotheses (Fig. 3C-E). Five hypotheses were favored
309 by all DNA datasets under the GTR model (AU test *p*-value >0.05), which variously placed
310 Rafflesiaceae with Putranjivaceae (H2), (Euphorbiaceae, Peraceae) (H1), and
311 Ixonanthaceae (H19), or with clades containing these families (H5, H18). Among these five
312 hypotheses, the phylogenetic affinity between Rafflesiaceae, Putranjivaceae (H2) decreased
313 when rapidly evolving genes were added to the analysis (Fig. 4C). In contrast, support for
314 the affinity between Rafflesiaceae, Ixonanthaceae, and (Euphorbiaceae, Peraceae) (H4, H18
315 and H19) increased instead. The protein alignment revealed a similar trend under the JTT
316 model. The monophyletic (Rafflesiaceae, Putranjivaceae) (H2) was generally favored
317 except when genes with the highest d_S values were added. However, the C30 protein
318 mixture model exclusively placed Rafflesiaceae with the MRCA of Euphorbiaceae, Peraceae,
319 Pandaceae, and Putranjivaceae (H11), which was strongly rejected under the JTT model
320 (Fig. 3E).
321

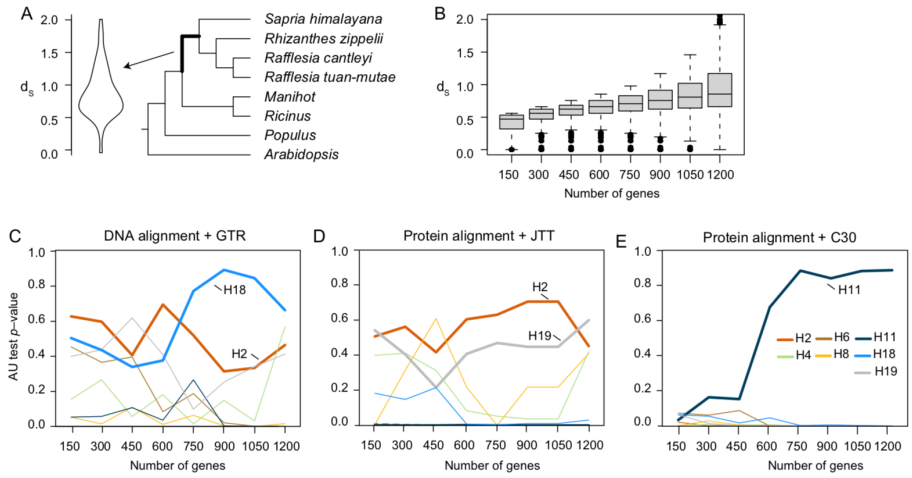

Fig. 3 Changes in topology test results when adding more rapidly evolving genes. **A.** Distribution of synonymous substitution rate (d_S) in the focal branch of stem group Rafflesiaceae. **B.** Distribution of d_S and number of genes in eight data subsets by gradual addition of genes with higher d_S. **C–E.** Dynamics in AU test *p*-values for alternative hypotheses across eight datasets under the GTR substitution model of DNA alignment (**C**), the JTT substitution model of protein alignment (**D**), and the C30 profile mixture model of protein alignment (**E**).

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 Rafflesiaceae and the other four focal clades ranged from 0.031 to 0.064 in coalescent units (Fig. 4A). The estimated minimum branch length to escape the anomaly zone was 0.536, which is ten times longer than the actual branch lengths (Fig. 4B). This result suggests that these rapid radiations involving Rafflesiaceae are clearly victims of the anomaly zone. Our gene tree simulation also corroborated this conclusion; the most common gene tree was discordant from the true species tree (Fig. 4D), which ranked 11th in frequency. The highly consistent frequencies between the empirical and simulated gene trees suggested that most gene tree variations can be explained by ILS. However, a significant difference was observed in topology H15, in which Rafflesiaceae was placed at the base of Malpighiales—a classic signal of long-branch attraction in empirical data (Kapli et al. 2020). This topology was

present in only 11% of the true gene trees in simulation but increased to 17% when adding in the ML inference step. This result captured the biases generated during the ML inference step, likely stemming from long-branch attraction, and aligned well with the empirical distribution of gene tree frequencies (Fig. 4D).

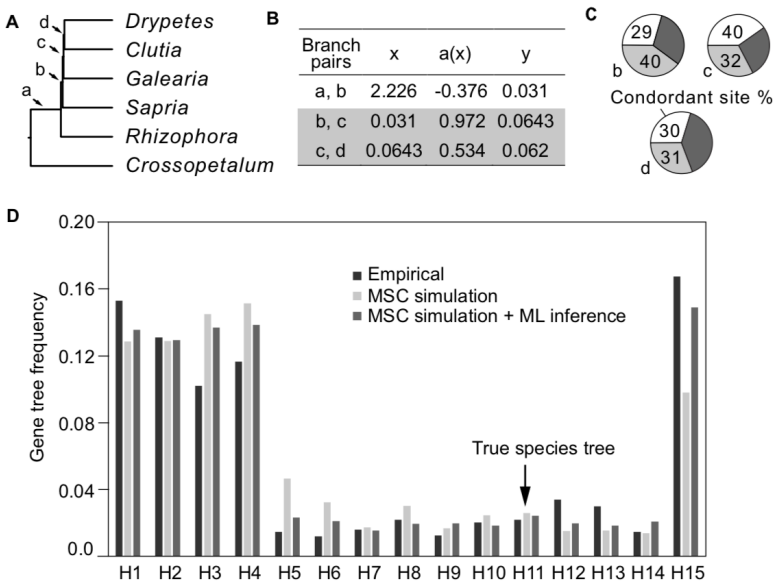

Fig 4. The anomaly zone and coalescent simulation. **A.** Reference species tree for coalescent simulation with labeled branches. **B**. For each internal branch pair in a four-taxon tree, boundaries of the anomaly zone a(x) calculated from (Degnan and Rosenberg 2006) are compared to the empirical branch length (y). Two out of three branch pairs (b-c and c-d) fall within the anomaly zone as y< a(x). **C.** Site concordance factors for branches b, c, and d in the simulated DNA sequences. The concordant sites are less than 33% (random distribution) in branches b and c. **D.** Gene tree frequency for the 15 alternative placements of Rafflesiaceae (H1–H15) in the empirical gene trees (dark grey), simulated true gene trees under the multispecies coalescent model (MSC; medium grey), and inferred maximum likelihood (ML) gene trees based on the simulated sequences (light grey). In all three cases, gene trees consistent with the true species tree topology have very low frequency.

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

364 the true topology was strongly rejected, suggesting statistical inconsistency
365 (supplementary table S5, Supplementary Material online). Instead, the incorrect species
366 topologies were variously supported by AU tests. H5 and H15 were most frequently
367 favored, which placed Rafflesiaceae with (Putranjivaceae, Euphorbiaceae) or the MRCA of
368 Rhizophoraceae and Euphorbiaceae, respectively. In contrast, summary coalescent
369 methods consistently recovered the true species when at least 4,000 loci were used
370 (supplementary table S6, Supplementary Material online). The gene tree estimation error
371 was as high as 78.1% for 500-bp-alignments and 70.0% for 2000-bp-alignments in terms of
372 the placement of Rafflesiaceae. Such high estimation error led to incorrect species tree
373 estimation when less than 3,000 genes were used. But even when enough loci were used
374 and the true species tree was recovered, the BP support for Rafflesiaceae was still
375 moderate to low (35–84 BP) and did not increase with the number of genes used.
376 Moreover, alternative topologies could not be rejected based on the likelihood ratio test.
377 For example, the difference in pseudolikelihoods between the true species tree H11 and an
378 alternative topology H5 was 0.71 when using 1,000 genes (pseudolikelihood = -10968.5)
379 and 20.3 when using 10,000 genes (pseudolikelihood = -109763.3), which were not
380 statistically significant (*p*–value >0.51, supplementary table S6, Supplementary Material
381 online).
382
383 **DISCUSSION**
384 **Challenges in identifying genes with optimum phylogenetic signal**
385 Selecting reliable genes with desired phylogenetic properties has been a routine practice to
386 remove biased signals and reduce computational burden. These phylogenetic properties
387 often exhibit various levels of covariance (Koch 2021). In our dataset, we found that some
388 commonly applied criteria for selecting phylogenetically informative loci, including high
389 gene tree–species tree congruence, high gene tree support, and low missing taxa, strongly
390 covary (Fig. 2). We also discovered that genes exhibiting clock-like features such as lower
391 root-to-tip variance do not necessarily have higher gene tree resolution (*p*-value 0.34),
392 which also corroborated with previous studies (Koch 2021). Regarding the placement of
393 Rafflesiaceae, the genes ideal for improving global phylogenetic resolution may not be
394 suitable for resolving this particular clade because Rafflesiaceae-specific phylogenetic

395 metrics do not covary with most tree-wise statistics. Instead, identified correlation patterns
396 were counter-intuitive—the nodal support of Rafflesiaceae was positively correlated with
397 missing data (*p*-value < 2.2e-16, Pearson correlation coefficient = -0.23) and negatively
398 correlated with global gene tree–species tree congruence (*p*-value = 9.1e-08, Pearson
399 correlation coefficient = -0.12; Fig. 2). As a result, canonical practices to select genes based
400 on tree-wise phylogenetic properties could confound the placement of Rafflesiaceae.
401 Focusing on Rafflesiaceae-specific metrics, the nodal support of Rafflesiaceae was strongly
402 influenced only by its own branch length, rather than the long descending branch or
403 compositional bias (Fig. 2). Therefore, its recalcitrant placement appears to be primarily
404 driven by the ancestral rapid radiation and less by the mutational saturation accumulated
405 in the stem group. The low branch support of Rafflesiaceae (mean BP = 23.6), on the other
406 hand, makes it less reliable as a locus selection criterion. For example, further investigation
407 of genes with high support values of Rafflesiaceae suggested that they were artifacts of
408 incorrect placement—genes with higher than 85 BP almost exclusively placed Rafflesiaceae
409 at the base of Malpighiales (supplementary table S2, Supplementary Material online).
410 These genes are in fact more likely to suffer from long-branch attraction. Therefore, neither
411 tree-wise properties nor Rafflesiaceae-specific metrics are reliable criteria for selecting
412 genes to place Rafflesiaceae. This is potentially a result of the altered selective pressure in
413 Rafflesiaceae stemming from parasitism (Cai et al. 2021a), making otherwise slowly
414 evolving phylogenetic markers behave differently. Another evidence of the differentiated
415 selection is our unsuccessful attempt to amplify angiosperm probe sets in Rafflesiaceae,
416 which are highly informative universal phylogenetic markers, but are largely lost or greatly
417 modified in Rafflesiaceae (Davis, unpublished data). When such species are the focus of
418 phylogenetic investigation, additional efforts should be made to test whether commonly
419 applied tree-wise metrics for phylogenetic informativeness are consistent at lineage-
420 specific levels.
421
422 **Genome-wide rate acceleration leads to extreme homoplasy**
423 The shift to a parasitic lifestyle has fundamentally reshaped the selective forces acting on
424 autotrophic functions such as photosynthesis (Cai 2023). As a result, relaxed selection and
425 accelerated nucleotide substitution rates are often characterized in the plastid and nuclear

426 genomes (Bromham et al. 2013; Wicke et al. 2016; Chen et al. 2023), but less so in the
427 mitochondria (Petersen et al. 2020). In Rafflesiaceae, this rate acceleration and the
428 ancestral rapid radiation that gave rise to stem group Rafflesiaceae created a sharp
429 contrast in branch lengths before and after the origin of Rafflesiaceae. The largest length
430 difference between these two branches was observed in the PhyloBayes analysis under the
431 CAT-GTR mixture model, which implied a 679-fold change. Estimates based on the
432 homogeneous site models in ML methods consistently implied a smaller change (68–82
433 fold). It is possible that homoplasy sites were incorrectly counted as ancestral mutations in
434 simplified site models, thus underestimating derived changes. Regardless, at least two
435 orders of magnitude of change were seen in these branch lengths, leading to extensive
436 phylogenetic conflicts. These conflicts are reflected by the low gene and site concordance
437 factors. Among the five focal clades shown in Fig. 2A, the lowest concordant factors were
438 seen in Rafflesiaceae. These low values remained consistent when using alternative
439 topologies, suggesting that incorrect phylogenetic placement is not the cause. Instead, it
440 originated from the excessive homoplasy in stem group Rafflesiaceae. This is corroborated
441 by our coalescent simulation in which only 29.41% of the simulated sites were concordant
442 with the true species tree (Fig. 4C).
443
444 The combination of high ILS and homoplasy present significant challenges for species
445 reconstruction using either concatenation or coalescent methods (Martyn and Steel 2012;
446 Roch et al. 2019). Our species tree falls well within the anomaly zone (Fig. 4B). As a result,
447 the concatenation method is statistically inconsistent because in our simulation analysis,
448 the true species tree was strongly rejected when more data were added. In the empirical
449 data, we also observed increased support for spurious relationships when rapidly evolving
450 genes were added (e.g., H4, H18, H19 in Fig. 3C). Homoplasy and long-branch attraction
451 also can confound coalescent-based inferences (Liu et al. 2023). In our simulation, gene
452 tree frequencies for the fifteen alternative topologies were largely consistent between the
453 simulated true gene trees and gene trees inferred from the simulated alignments (Fig. 4D).
454 Yet, a disproportionately high frequency of H15 was recovered among estimated gene trees
455 when they were inferred from sequences likely suffering from long-branch attraction.
456 When aggregating such erroneous signals into concatenated alignments, it may contribute

457 to the support of H15 in optimum species tree searches and in likelihood ratio tests (Fig.
458 2B; supplementary table S4, Supplementary Material online).
459
460 **Insights to the phylogenetic placement of Rafflesiaceae**
461 Our simulation concluded that current methods for modeling substitution and coalescent
462 processes are insufficient to fully resolve the origin of Rafflesiaceae within Malpighiales.
463 Massive gene loss in this parasitic plant clade renders a greatly reduced pool of
464 phylogenetic markers that are far from enough to achieve a definitive placement.
465 Meanwhile, alternative methods employing macro- and micro-scale gene synteny have
466 been increasingly implemented to untangle recalcitrant relationships in angiosperms and
467 teleost fishes (Zhao et al. 2021; Parey et al. 2023). As more genomic data to be developed in
468 Malpighiales and beyond, the application of these new methods may provide some novel
469 insights into this thorny phylogenetic problem.
470
471 Despite these challenges, two hypotheses were consistently recovered across multiple
472 analyses, placing Rafflesiaceae with Putranjivaceae (H2), or with the MRCA of
473 Putranjivaceae, Pandaceae, Euphorbiaceae, and Peraceae (H11). First, the monophyletic
474 (Rafflesiaceae, Putranjivaceae) relationship (H2) was supported by the 'democratic vote'
475 approach of GGI, concatenation methods with homogeneous site models, and full
476 coalescent models. In the GGI test, this topology was slightly favored by most genes, yet the
477 difference is not statistically significant for nearly all genes. Such a result is not surprising
478 because individual loci may not contain enough information to reject alternative
479 hypotheses (Arcila et al. 2017; Puttick et al. 2018). When aggregating signals into
480 concatenated alignments, support for the monophyletic (Rafflesiaceae, Putranjivaceae)
481 became statistically significant (supplementary table S4, Supplementary Material online).
482 All concatenation methods with homogeneous site models (i.e., GTR and JTT) supported
483 this hypothesis with consistently high confidence across datasets with different levels of
484 missing data or evolutionary rates. This included the full coalescent model from starBEAST
485 at high confidence (0.98 PP; supplementary fig. S2, Supplementary Material online).
486 However, this hypothesis was strongly rejected when mixture models were used. Both the
487 CAT-GTR and C30 mixture models favored an earlier divergence of Rafflesiaceae as

488 suggested in H11 (supplementary table S4, Supplementary Material online). Moreover,
489 ASTRAL and MP-EST also recovered H11 as the optimum species tree and the support is
490 statistically significant under the likelihood ratio test.
491
492 Between these two hypotheses, H11 was supported by more complex and realistic models
493 such as the site mixture model and the multispecies coalescent model. In addition, it is also
494 important to recognize that the GGI and AU test of concatenated alignments can be
495 positively misleading when ILS is prevalent. The drawbacks of concatenation under high
496 ILS have been discussed extensively (Song et al. 2012; Xi et al. 2014; Edwards et al. 2016).
497 Similarly, the GGI approach can also suffer from ILS especially in the anomaly zone where
498 the most common gene tree differs from the species tree (Degnan and Rosenberg 2006).
499 Indeed, our simulation demonstrates that the concordant gene trees occur at very low
500 frequency, which will lead to the rejection of the true species tree using the GGI approach
501 (Fig. 4).
502
503 The four tropical families suggested to be close relatives of Rafflesiaceae include
504 Putranjivaceae, Pandaceae, Euphorbiaceae, and Peraceae, all of which have a pantropical
505 distribution. Pentamerous dioecious flowers, such as those in Rafflesiaceae, are commonly
506 found in Putranjivaceae (and the unsampled close relative Lophopyxidaceae) and
507 Pandaceae (Matthews and Endress 2013; Wurdack et al. 2004). But all these free-living
508 lineages produce only minute flowers, corroborating the previous claim that the enormous
509 flowers of Rafflesiaceae evolved from ancestors with tiny flowers (Davis et al. 2007).
510 Although the highly modified morphology of Rafflesiaceae makes it challenging to identify
511 synapomorphies, several shared features support its close relationship with the proposed
512 sister families. For example, like Rafflesiaceae, Putranjivaceae also releases sulfur-rich,
513 strong-smelling chemicals to attract pollinators. The foul odor of Rafflesiaceae flowers
514 largely comes from dimethyl disulfide and dimethyl trisulfide (Jürgens et al. 2013; Wee et
515 al. 2018). In *Drypetes* (Putranjivaceae), such a strong smell comes from isothiocyanates
516 from the glucosinolates pathway (Johnson et al. 2009). Future studies focusing on the
517 metabolic pathway of these chemicals will elucidate whether they are derived from a
518 shared ancestor.

519

520 **Conclusion**

521 Both incomplete lineage sorting and long-branch attraction have played an important role
522 in shaping the recalcitrant placement of Rafflesiaceae. The unique life-history strategy of
523 this endoparasitic family of plants has led to unique evolutionary patterns of genes that
524 have different patterns of "phylogenetic informativeness" compared to their free-living
525 relatives. We have demonstrated for the first time that accelerated substitution rate is a
526 genome-wide phenomenon in this parasitic plant lineage and contributes to long-branch
527 attraction. We also identified two plausible phylogenetic scenarios for the placement of
528 Rafflesiaceae, either with Putranjivaceae or the common ancestor of Putranjivaceae,
529 Pandaceae, Euphorbiaceae, and Peraceae. Both placements support a tropical origin of the
530 family and the striking contrast in flower size between Rafflesiaceae and its ancestors.
531 However, our simulation also suggested that the combination of incomplete lineage sorting,
532 long-branch attraction, and massive gene loss has created a vexing phylogenetic scenario
533 that is impossible to resolve with currently available methods for phylogenetic inference.

534

535 **Materials and Methods**

536 ***Phylogenomic dataset assembly, sequence alignment, and gene tree reconstruction***

537 We combined the two phylogenomic datasets from our previous studies (Cai et al. 2019;
538 Cai et al. 2021a) to generate a 2,141-gene dataset (supplementary table S1, Supplementary
539 Material online). The combined dataset included 43 species representing all three genera of
540 Rafflesiaceae, twenty-one families broadly sampled across Malpighiales, and three
541 outgroup species from Oxalidales and Celastrales. Briefly, the two datasets were merged
542 based on gene IDs of the three Malpighiales species sampled in both studies (*Manihot*
543 *esculenta*, *Populus trichocarpa*, and *Jatropha curcas*). Then, the tree-based ortholog
544 identification approach from Yang and Smith (2014) was applied iteratively to remove HGT
545 genes in Rafflesiaceae or paralogs arising from gene duplications. The final phylogenomic
546 dataset consisted of 2,141 protein-coding orthogroups satisfying the following three
547 criteria: (1) included at least one vertically transmitted gene from Rafflesiaceae; (2)
548 included at least 10 species (23% species completeness); (3) all genes are orthologous to
549 each other, and no paralogs generated by gene duplications prior to the divergence of

550 Malpighiales were included. No new molecular data was generated for this study and a
551 detailed description of how we assembled the dataset is provided in supplementary note
552 S1, Supplementary Material online.
553
554 Protein and CDS sequences of these 2,141 single-copy orthogroups were aligned using
555 MAFFT v.7.299 (Katoh and Standley 2013) and then converted into the corresponding
556 codon alignments using pal2nal (Suyama et al. 2006). Alignments were trimmed by trimAL
557 v1.2 (Capella-Gutiérrez et al. 2009) to remove sites containing more than 50% missing data
558 in the DNA and protein alignments.
559
560 We inferred a ML phylogeny for each locus based on DNA alignments using IQ-TREE v2.3.2
561 (Minh et al. 2020b). Optimal substitution models for each alignment were determined by
562 ModelFinder (Kalyaanamoorthy et al. 2017) using the Bayesian Information Criterion (BIC)
563 within IQ-TREE. Branch supports were assessed using ultrafast bootstrap approximation
564 (UFBP; Hoang et al. 2018) with 1000 replicates along with the '-bnni' option to reduce the
565 risk of overestimating branch support resulting from model violations. We also used a non-
566 parametric bootstrap (BP) with 100 replications to assess branch support for each gene
567 tree. All DNA alignments, protein alignments, and individual gene trees were deposited in
568 Zenodo (DOI: 10.5281/zenodo.11643026). All scripts used for orthogroup selection and
569 gene tree reconstruction are available on the GitHub repository
570 (https://github.com/lmcai/Rafflesiaceae_phylogenomics).
571
572 **Quantifying phylogenetic properties of individual genes**
573 Each gene was quantified for the following phylogenetic properties using SortaDate (Smith
574 et al. 2018): number of species, alignment length, average gene tree support, total tree
575 length, root-to-tip branch-length variance, and gene tree–species tree congruence
576 (supplementary table S2, Supplementary Material online). A species tree estimated by
577 ASTRAL-III v5.7.3 (Zhang et al. 2017) based on all 2,141 orthogroups (analysis 12 in Fig.2;
578 see the following section for species tree estimation) was used for assessing gene tree–
579 species tree congruence, but we removed all Rafflesiaceae species to avoid the systematic
580 bias stemming from artificially fixing this unstable node. Four phylogenetic metrics specific

581 to Rafflesiaceae were extracted: the base compositional bias of Rafflesiaceae, the branch
582 length of stem-group Rafflesiaceae and its ascending parent branch, and the nodal support
583 of Rafflesiaceae. The base compositional heterogeneity was derived from the Chi-square
584 test in IQ-TREE, whereas the other metrics were summarized using a custom Python script
585 from gene trees (available on GitHub). Finally, we modeled the pairwise Pearson
586 correlation among these phylogenetic metrics. The correlation heatmap was generated
587 using the function 'heatmap' from the R package corrplot (Wei et al. 2017).
588
589 **Generation of phylogenetic hypotheses with data subsampling**
590 Eighteen explorative phylogenomic analyses were conducted (Figure 1, supplementary
591 table S3, Supplementary Material online). These comprised concatenation (11 analyses)
592 and coalescent methods (7 analyses) that used either DNA (13 analyses) or protein
593 alignments (5 analyses) for the complete dataset (5 analyses) or subsampled datasets (7
594 analyses). Two data subsets were created in an attempt to minimize the artifacts from
595 missing data, especially in the four focal clades that were closely related to Rafflesiaceae.
596 These four clades included (Euphorbiaceae, Peraceae), Putranjivaceae, Pandaceae, and
597 (Erythroxylaceae, Rhizophoraceae). First, a 1,553–loci dataset (referred to as G1553) was
598 selected from the complete dataset (referred to as G2141) to include genes that contained
599 at least 29 species (67% species completeness) and at least three of the four focal clades.
600 Second, a 647–loci dataset (referred to as G647) was created to select genes containing at
601 least 34 species (80% species completeness) and three of the four focal clades.
602
603 *Concatenation analysis*—We inferred the ML phylogeny based on DNA alignments (analysis
604 1–5 in Fig. 2) using ExaML v3.0.18 (Kozlov et al. 2015), RAxML v8.2.11 (Stamatakis 2014),
605 and IQ-TREE. For both complete and subsampled datasets, the best partition schemes were
606 determined using PartitionFinder v2.1.1 (Lanfear et al. 2017), the corrected Akaike
607 Information Criterion, and the heuristic rcluster searching algorithm (--rcluster-percent
608 10). Species trees were subsequently inferred under the GTR + G model. Branch support
609 was evaluated using 100 BP replicates in ExaML and RAxML; and using 1000 UFBP
610 replicates in IQ-TREE. For protein analyses (analysis 7–9 in Fig. 2), we applied the GAMMA
611 model to the unpartitioned complete dataset (analysis 7) in ExaML and a site-specific

612 posterior mean site frequency (PMSF) model (Wang et al. 2018) to the subsampled
613 datasets in IQ-TREE (analysis 8 and 9). The PMSF model approximates the empirical profile
614 mixture model of Le et al. (2008), which is a variant of the infinite mixture model CAT-GTR
615 in PhyloBayes. It can effectively ameliorate long-branch attraction artifacts and is
616 computationally more efficient (Wang et al. 2018). Because of the memory limit of our
617 computing cluster, we only applied the PMSF model to subsampled datasets (G647 and
618 G1553) and used 30 mixture components for protein profile (LG+C30+F+G).
619
620 *Summary coalescent analysis*—We used ASTRAL-III and MP-EST v2.1 (Liu et al. 2010) to
621 infer species trees under the coalescent model. For ASTRAL analyses (analysis 12, 15 in Fig.
622 2C; supplementary table S3, Supplementary Material online), ML gene trees inferred from
623 DNA alignments of the complete dataset G2141 were used and branch support was
624 evaluated using the local posterior probability. To reduce the impact of gene-tree
625 estimation error, we also collapsed spurious branches with less than 70 UFBP using
626 Newick Utilities v1.6 (Junier and Zdobnov 2010). The collapsed gene trees were then used
627 to infer the species tree in ASTRAL (analysis 15). For MP-EST analyses, we used both the
628 complete dataset (analysis 13, 14) and the subsampled datasets (analysis 16, 17) for
629 species tree estimation (supplementary table S3, Supplementary Material online). Branch
630 support was evaluated using 100 replicates from BP gene trees inferred from the DNA
631 sequences.
632
633 *Full coalescent analysis*—We inferred a species tree using the full coalescent model and the
634 uncorrelated lognormal clock (UCLN) model implemented in the starBEAST2 package of
635 BEAST v.2.4.5 (analysis 18 in Fig. 2; Ogilvie et al., 2017; Bouckaert et al., 2014). The high
636 computational demand of starBEAST2 required us to further subsample the dataset by
637 species and loci. We therefore included 13 representative species from focal clades and
638 filtered the DNA alignments to generate an 81-exon alignment. The selection of exon
639 instead of full length CDS sequence is because the exceptionally long intron in Rafflesiaceae
640 (up to 97.8 kb; Cai et al. 2021a) may violate the recombination assumption of the
641 coalescent model. We ran starBEAST2 for 500 million generations sampling every 50,000
642 generations. Each locus was assigned to its own relaxed clock and a HKY model of

643 molecular evolution. For more details regarding starBEAST, see supplementary note 2,
644 Supplementary Material online. The sequence alignment and xml file for running the
645 analysis is provided in the GitHub repository.
646
647 **Application of complex compositional and rate heterogeneity models**
648 We employed two divergent strategies to mitigate the effects of compositional and rate
649 heterogeneity. We first used the CAT-GTR model in Phylobayes v3.1.1, which provides
650 infinite mixture models of site-specific evolutionary profiles (Lartillot et al. 2013; analysis 6
651 in Fig. 2). To ease convergence and reduce computational burden, we used a subsampled
652 dataset of 168 genes (209,782 DNA sites) that are present in at least 75% of the species
653 and all focal clades. We ran the CAT-GTR model in four independent chains until
654 convergence was reached between at least two chains. We assessed convergence based on
655 effective sample size of parameters (target > 200) and the maximum discrepancy observed
656 across all bipartitions (target < 0.15) using bpcomp and tracecomp in PhyloBayes.
657
658 In addition to modeling the complex site-specific process, we also applied the six-state
659 Dayhoff amino acid recoding (Dayhoff et al. 1978) in an attempt to mask mutational noises,
660 especially in stem group Rafflesiaceae (analysis 10 and 11 in Fig. 2). Dayhoff recoding
661 groups the twenty amino acids into six bins, each containing a set of biochemically similar
662 amino acids. We recoded the protein alignment from the G647 and G1553 datasets using
663 Phylogears v2.2 (Tanabe 2012). ML trees were subsequently inferred for each of the two
664 concatenated datasets under the MULTIGAMMA model in RAxML. Branch support was
665 assessed with 100 rapid BP replicates followed by a thorough ML search.
666
667 **Evaluating gene and site phylogenetic conflicts**
668 We used gene tree quartet frequency and site concordance factors to quantify phylogenetic
669 conflicts using ASTRAL and IQ-TREE, respectively. Gene tree quartet frequency
670 summarizes the percentages of consistent quartets among decisive gene trees based on a
671 reference species tree (Sayyari and Mirarab 2016). The site concordance factor was
672 similarly defined but used a heuristic method to sample alignment sites (Minh et al. 2020a).
673 The gene tree quartet frequency was calculated by ASTRAL when invoking the full branch

674 annotation option (-t 2). We used the filtered gene tree branches from the complete dataset
675 to infer the quartet frequency (branch support >70 UFBP; analysis 15 in Fig. 2). In doing so,
676 we aimed to summarize quartet frequency from only informative branches. To calculate
677 site concordance factor in IQ-TREE, we used the same ASTRAL tree as the reference species
678 tree and the concatenated DNA and protein alignments from the G1553 dataset as input.
679 For each internal branch, 200 quartets were sampled to obtain stable site concordance
680 factors (--scf 200).
681
682 **Fit of individual gene trees to proposed topologies**
683 The initial phylogenetic reconstructions revealed four focal clades that are closely related
684 to Rafflesiaceae (Fig. 2). To more finely place Rafflesiaceae, we proposed a total of 19
685 topologies with all possible placements of Rafflesiaceae relative to these four clades (H1–
686 H15 in Fig. 2). We also added two topologies proposed by Barkman et al (2004) and Davis
687 et al (2007), which placed Rafflesiaceae with Passifloraceae (H17) and Euphorbiaceae
688 (H16), respectively. Finally, Ixonanthaceae or (Ixonanthaceae, (Erythroxylaceae,
689 Rhizophoraceae)) was also inferred as a close relative of Rafflesiaceae in some
690 concatenation analyses. Thus, we also added two hypotheses that placed Rafflesiaceae with
691 (Ixonanthaceae, (Erythroxylaceae, Rhizophoraceae)) (H18) or Ixonanthaceae alone (H19).
692
693 We tested the fitness of individual gene trees to the 19 alternative topologies using Gene
694 Genealogy Interrogation (GGI) implemented in Arcila et al. (2017). Briefly, for each of the
695 2,141 genes, we conducted constrained phylogenetic searches enforcing the relationship
696 imposed by each of the 19 hypotheses. Only genes containing at least two of the four focal
697 clades were used for GGI analysis. Site likelihood scores for each tree were obtained under
698 the GTR+G model based on DNA alignments in RAxML. Topology tests were conducted by
699 statistically comparing the site likelihood scores among all proposed topologies via three
700 tests implemented in CONSEL v0.20 (Shimodaira and Hasegawa 2001)—the AU test
701 (Shimodaira 2002), the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa 1999),
702 and the Kishino-Hasegawa (KH) test (Kishino and Hasegawa 1989). Among these three
703 approaches, the AU test uses multi-scale bootstrapping techniques and is most suitable for

704 simultaneous comparisons of multiple trees. Finally, trees were ranked according to the *p*-
705 values supporting each hypothesis and visualized using R.
706
707 **Fit of concatenated alignments to proposed topologies**
708 We evaluated the support for 19 *a priori* hypotheses using the concatenated alignments of
709 the G647 and G1553 datasets. For DNA alignments, site likelihood scores for each
710 constraint topology were obtained under the GTR + G model in RAxML. We then conducted
711 AU, SH, and KH tests in CONSEL following the same method described above. We also
712 conducted AU, SH, and KH tests in IQ-TREE using RELL approximation and bootstrapping.
713 Topology tests for protein alignments were conducted under the LG+C30+F+G profile
714 mixture model in IQ-TREE.
715
716 **Fit of proposed topologies under the multispecies coalescent model**
717 To compare the goodness-of-fit of species trees to the empirical gene trees under the
718 coalescent model, we conducted a pairwise likelihood ratio test between the 19 proposed
719 topologies (Liu et al. 2019). We first conducted constrained species tree searches in
720 ASTRAL-III using the empirical gene trees from the G2141 dataset. We then optimized the
721 branch length of these proposed species trees and calculated the likelihood in MP-EST. We
722 used the test2.sptree function from the R package phybase v2.0 (Liu and Yu 2010) to
723 conduct the likelihood ratio test described in Liu et al. (Liu et al. 2019). The likelihood ratio
724 test statistic is given by $t = 2(L_{\mathrm{Tree2}} - L_{\mathrm{Tree1}})$, in which L_{Tree1} and L_{Tree2} are the log-likelihoods
725 of the null and alternative hypotheses. The statistical significance (*p*-value) of the
726 comparison was evaluated by generating a null distribution of *t* via bootstrap sampling
727 gene trees 100 times (Liu et al. 2019).
728
729 **Identify spurious relationships arising from mutational noises**
730 To explore the sensitivity of these topology tests to saturation and model selection, we
731 generated eight data subsets with increasing mutation rates in Rafflesiaceae and then
732 evaluated the 19 hypotheses using the AU test and RELL bootstrapping. First, we calculated
733 the d_S value of stem group Rafflesiaceae using data from Cai et al (2021a; Supplementary
734 Table S2). Briefly Cai et al (2021a) used codon alignments from nine species—including all

735 four Rafflesiaceae sampled in this study, plus *Manihot esculenta* (Euphorbiaceae), *Jatropha*
736 *curcas* (Euphorbiaceae), *Populus trichopoda* (Salicaceae), *Arabidopsis thaliana*
737 (Brassicaceae), and *Glycine max* (Fabaceae)—to infer substitution rates under a free ratio
738 model in PAML v4.8 (Yang 2007). This allowed for the estimation of synonymous (d_S) and
739 nonsynonymous (d_N) rates for each branch on the species tree. Detailed description for
740 PAML analysis can be found in Cai et al. (2021a) and on GitHub
741 (https://github.com/lmcai/Sapria_genomics/tree/master/5_selection).
742
743 We then ordered the G1553 dataset based on the d_S of stem group Rafflesiaceae. We
744 excluded loci with larger than 3.0 d_S in stem group Rafflesiaceae to avoid saturation. A total
745 of 1,182 loci were retained for further investigation. These loci were then ordered based on
746 d_S values from low to high. Eight data subsets were generated by including 150, 300, 450,
747 600, 750, 900, 1,050, and 1,182 loci with increasing d_S values. Finally, we conducted
748 topology tests on concatenated DNA and protein alignments using the AU test and RELL
749 bootstrapping following the methods described above. For DNA analyses, the GTR+G model
750 was applied to the unpartitioned alignments. For protein analyses, both JTT+F+R4 and
751 LG+C30+F+G model were applied to the unpartitioned alignments.
752
753 **Test of the anomaly zone**
754 The anomaly zone in a four-taxon species tree is defined by the length ratio of the two
755 internal branches (Degnan and Rosenberg 2006). To test if the focal clades fall within the
756 anomaly zone, we estimated the boundaries of the anomaly zone a(x) according to equation
757 4 from (Degnan and Rosenberg 2006). We used the phylogeny estimated by MP-EST from
758 the complete dataset as the reference species tree (analysis 14 in Fig. 2). Here, the three
759 internal nodes among the focal clades form three pairs of parent-descendent branches
760 whose topology matches the asymmetrical four-taxon phylogeny from (Degnan and
761 Rosenberg 2006). If the length of the descendant internal branch is smaller than a(x), then
762 the internode pair is in the anomaly zone.
763
764 **Coalescent simulation**

765 To further explore how ILS and homoplasy affect species tree reconstruction, we simulated
766 gene trees and DNA alignments based on the coalescent and mutational parameters from
767 the empirical dataset. The MP-EST species tree estimated from the complete dataset
768 (analysis 14 in Fig. 2) was used as the reference species tree and was pruned to include
769 only one species from each of the five focal clades plus an outgroup, *Crossopetalum*. We
770 used the function 'sim.coal.mpest' in the R package Phybase v2.0 (Liu and Yu 2010) to
771 simulate 10,000 gene trees using the reference species tree. We then modified the branch
772 lengths of the simulated gene trees to reflect the inter-species and inter-locus rate variation
773 in mutational units. To do so, we fit a Gaussian distribution to each internal and terminal
774 branches (e.g., terminal branches leading to *Sapria*, Rafflesiaceae) from the empirical ML
775 gene trees using function 'fitdistr' in the R package MASS v7.3-54 (Ripley et al. 2013). We
776 then assigned branch length in mutation units to the simulated trees by randomly sampling
777 the corresponding Gaussian distribution. Finally, we used the parameters of the GTR+F+G4
778 model estimated by ExaML based on the complete dataset (analysis 1 in Fig. 2) to simulate
779 DNA alignments in Seq-Gen v1.3.4 (Rambaut and Grassly 1997). For each gene, four
780 alignments with different lengths (500, 1000, 1500, and 2000 bp) were simulated to reflect
781 high to low levels of gene tree estimation error (Mirarab et al. 2016; Cai et al. 2021b). We
782 then inferred gene trees in IQ-TREE under the GTR+F+G4 model using 100 bootstrap
783 replicates. Species trees were estimated using both concatenation and coalescent methods.
784 For concatenation analyses, we applied the unpartitioned dataset in IQ-TREE under the
785 GTR+F+G4 model. Branch support was estimated by 1000 UFBP. For the coalescent
786 method, we applied the ML gene trees in MP-EST and ASTRAL. Branch support was
787 estimated by 100 BP replicates. We also subsampled the dataset to include 500, 1000,
788 2000, 3000, 4000, and 5000 loci with four alignment lengths. Finally, we conducted AU
789 tests in IQ-TREE and likelihood ratio tests in MP-EST to statistically evaluate alternative
790 topologies.
791
792 **Acknowledgements:**
793 We thank the two anonymous reviewers for insightful comments on the initial version of
794 the manuscript. We thank members of the Davis lab for feedback on early versions of this
795 study.

Funding

This research was supported by National Science Foundation (NSF) AToL grant (DEB-0622764), NSF grant DEB-1120243, and startup funds from Harvard University to CCD. LC was supported by the Stengl-Wyer Postdoc Fellowship from the University of Texas at Austin.

Data availability

No new sequencing data was generated. All DNA alignments, protein alignments, and individual gene trees were available on Zenodo (DOI: 10.5281/zenodo.11643026). All codes to conduct the analysis were available on GitHub https://github.com/lmcai/Rafflesiaceae_phylogenomics.

Reference

Arcila, D., Ortí, G., Vari, R., Armbruster, J.W., Stiassny, M.L., Ko, K.D., Sabaj, M.H., Lundberg, J., Revell, L.J., Betancur-R, R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat Ecol Evol, 1:0020.

Baldwin, B.G. 1997. Adaptive radiation of the Hawaiian silversword alliance: congruence and conflict of phylogenetic evidence from molecular and non-molecular investigations. Molecular evolution and adaptive radiation, 103:1840-1843.

Barkman, T.J., Bendiksby, M., Lim, S.-H., Salleh, K.M., Nais, J., Madulid, D., Schumacher, T. 2008. Accelerated rates of floral evolution at the upper size limit for flowers. Current Biology, 18:1508-1513.

Barkman, T.J., Lim, S.-H., Salleh, K.M., Nais, J. 2004. Mitochondrial DNA sequences reveal the photosynthetic relatives of Rafflesia, the world's largest flower. Proceedings of the National Academy of Sciences, 101:787-792.

Bendiksby, M., Schumacher, T., Gussarova, G., Nais, J., Mat-Salleh, K., Sofiyanti, N., Madulid, D., Smith, S.A., Barkman, T. 2010. Elucidating the evolutionary history of the Southeast Asian, holoparasitic, giant-flowered Rafflesiaceae: Pliocene vicariance, morphological convergence and character displacement. Molecular Phylogenetics and Evolution, 57:620-633.

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

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., and Drummond, A.J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10, e1003537.

Cai L. 2023. Rethinking convergence in plant parasitism through the lens of molecular and population genetic processes. Am. J. Bot. 110:e16174.

835 Cai, L., Arnold, B.J., Xi, Z., Khost, D.E., Patel, N., Hartmann, C.B., Manickam, S., Sasirat, S.,
836 Nikolov, L.A., Mathews, S., *et al.* 2021a. Deeply Altered Genome Architecture in the
837 Endoparasitic Flowering Plant Sapria himalayana Griff.(Rafflesiaceae). Current
838 Biology.
839 Cai, L., Xi, Z., Amorim, A.M., Sugumaran, M., Rest, J.S., Liu, L., Davis, C.C. 2019. Widespread
840 ancient whole-genome duplications in Malpighiales coincide with Eocene global
841 climatic upheaval. New Phytologist, 221:565-576.
842 Cai, L., Xi, Z., Lemmon, E.M., Lemmon, A.R., Mast, A., Buddenhagen, C.E., Liu, L., Davis, C.C.
843 2021b. The Perfect Storm: Gene Tree Estimation Error, Incomplete Lineage Sorting,
844 and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade,
845 Malpighiales. Systematic Biology, 70:491-507.
846 Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. 2009. trimAl: a tool for automated
847 alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25:1972-
848 1973.
849 Chen X., Fang D., Xu Y., Duan K., Yoshida S., Yang S., Sahu S.K., Fu H., Guang X., Liu M., Wu C.,
850 Liu Y., Mu W., Chen Y., Fan Y., Wang F., Peng S., Shi D., Wang Y., Yu R., Zhang W., Bai
851 Y., Liu Z.-J., Yan Q., Liu X., Xu X., Yang H., Wu J., Graham S.W., Liu H. 2023.
852 Balanophora genomes display massively convergent evolution with other extreme
853 holoparasites and provide novel insights into parasite–host interactions. Nat. Plants.
854 9:1627–1642.
855 Cronquist, A. 1981. An integrated system of classification of flowering plants. Columbia
856 university press.
857 Davis, C.C., Latvis, M., Nickrent, D.L., Wurdack, K.J., Baum, D.A. 2007. Floral gigantism in
858 Rafflesiaceae. Science, 315:1812-1812.
859 Davis, C.C., Webb, C.O., Wurdack, K.J., Jaramillo, C.A., Donoghue, M.J. 2005. Explosive
860 radiation of Malpighiales supports a mid-Cretaceous origin of modern tropical rain
861 forests. The American Naturalist, 165:E36-E65.
862 Davis, C.C., Wurdack, K.J. 2004. Host-to-parasite gene transfer in flowering plants:
863 phylogenetic evidence from Malpighiales. Science, 305:676-678.
864 Dayhoff, M., Schwartz, R., Orcutt, B. 1978. 22 a model of evolutionary change in proteins.
865 Atlas of protein sequence and structure, 5:345-352.
866 Degnan, J.H., Rosenberg, N.A. 2006. Discordance of species trees with their most likely gene
867 trees. PLOS Genet, 2:e68.
868 Degnan, J.H., Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the
869 multispecies coalescent. Trends in ecology & evolution, 24:332-340.
870 Donoghue, M.J., Eaton, D.A., Maya-Lastra, C.A., Landis, M.J., Sweeney, P.W., Olson, M.E.,
871 Cacho, N.I., Moeglein, M.K., Gardner, J.R., Heaphy, N.M. 2022. Replicated radiation of
872 a plant clade along a cloud forest archipelago. Nature Ecology & Evolution, 6:1318-
873 1329.
874 Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., Zhong, B., Wu, S.,
875 Lemmon, E.M., Lemmon, A.R. 2016. Implementing and testing the multispecies
876 coalescent model: a valuable paradigm for phylogenomics. Molecular phylogenetics
877 and evolution, 94:447-462.
878 Foster, P.G., Hickey, D.A. 1999. Compositional bias may affect both DNA-based and protein-
879 based phylogenetic reconstructions. Journal of molecular evolution, 48:284-290.

880 Givnish, T.J., Millam, K.C., Mast, A.R., Paterson, T.B., Theim, T.J., Hipp, A.L., Henss, J.M., Smith,
881 J.F., Wood, K.R., Sytsma, K.J. 2009. Origin, adaptive radiation and diversification of
882 the Hawaiian lobeliads (Asterales: Campanulaceae). Proceedings of the Royal
883 Society B: Biological Sciences, 276:407-416.
884 Glor R.E. 2010. Phylogenetic Insights on Adaptive Radiation. Annu. Rev. Ecol. Evol. Syst.
885 41:251–270.
886 Grant, P.R., Grant, B.R. 2007. How and why species multiply: the radiation of Darwin's
887 finches. Princeton University Press.
888 Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S. 2018. UFBoot2:
889 improving the ultrafast bootstrap approximation. Mol Biol Evol, 35:518-522.
890 Jin, D.-M., Wicke, S., Gan, L., Yang, J.-B., Jin, J.-J., Yi, T.-S. 2020. The Loss of the Inverted
891 Repeat in the Putranjivoid Clade of Malpighiales. Frontiers in plant science, 11:942.
892 Johnson, S.D., Griffiths, M.E., Peter, C.I., Lawes, M.J. 2009. Pollinators,"mustard oil" volatiles,
893 and fruit production in flowers of the dioecious tree Drypetes natalensis
894 (Putranjivaceae). American Journal of Botany, 96:2080-2086.
895 Junier, T., Zdobnov, E.M. 2010. The Newick utilities: high-throughput phylogenetic tree
896 processing in the UNIX shell. Bioinformatics, 26:1669-1670.
897 Jürgens, A., Wee, S.L., Shuttleworth, A., Johnson, S.D. 2013. Chemical mimicry of insect
898 oviposition sites: a global analysis of convergence in angiosperms. Ecology Letters,
899 16:1157-1167.
900 Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., von Haeseler, A., Jermiin, L.S. 2017.
901 ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods,
902 14:587.
903 Kapli, P., Yang, Z., Telford, M.J. 2020. Phylogenetic tree building in the genomic age. Nat Rev
904 Genet, 21:428-444.
905 Katoh, K., Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7:
906 improvements in performance and usability. Mol Biol Evol, 30:772-780.
907 Kishino, H., Hasegawa, M. 1989. Evaluation of the maximum likelihood estimate of the
908 evolutionary tree topologies from DNA sequence data, and the branching order in
909 Hominoidea. J Mol Evol, 29:170-179.
910 Knight, R.D., Freeland, S.J., Landweber, L.F. 2001. Rewiring the keyboard: evolvability of the
911 genetic code. Nature Reviews Genetics, 2:49-58.
912 Koch, N.M. 2021. Phylogenomic subsampling and the search for phylogenetically reliable
913 loci. Molecular Biology and Evolution.
914 Kozlov, A.M., Aberer, A.J., Stamatakis, A. 2015. ExaML version 3: a tool for phylogenomic
915 analyses on supercomputers. Bioinformatics, 31:2577-2579.
916 Kuijt, J. 1969. The biology of parasitic flowering plants. Berkeley, Los Angeles: Univ. Calif.
917 Press.
918 Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., Calcott, B. 2017. PartitionFinder 2: new
919 methods for selecting partitioned models of evolution for molecular and
920 morphological phylogenetic analyses. Molecular biology and evolution, 34:772-773.
921 Lartillot, N., Rodrigue, N., Stubbs, D., Richer, J. 2013. PhyloBayes MPI: phylogenetic
922 reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol,
923 62:611-615.
924 Le, S.Q., Gascuel, O., Lartillot, N. 2008. Empirical profile mixture models for phylogenetic
925 reconstruction. Bioinformatics, 24:2317-2323.

Liu, L., Anderson, C., Pearl, D., Edwards, S.V. 2019. Modern phylogenomics: building phylogenetic trees using the multispecies coalescent model. Evolutionary Genomics, Springer, p. 211-239.

Liu, L., Yu, L. 2010. Phybase: an R package for species tree analysis. Bioinformatics, 26:962-963.

Liu, L., Yu, L., Edwards, S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC evolutionary biology, 10:1-18.

Liu L, Yu L, Wu S, Arnold J, Whalen C, Davis C, Edwards S. 2023. Short branch attraction in phylogenomic inference under the multispecies coalescent. Front. Ecol. Evol. 11: 1134764.

Losos, J.B. 2011. Lizards in an evolutionary tree: ecology and adaptive radiation of anoles. Univ of California Press.

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

Matthews M.L., Endress P.K. 2013. Comparative floral structure and systematics of the clade of Lophopyxidaceae and Putranjivaceae (Malpighiales). Bot. J. Linn. Soc. 172:404–448.

Minh, B.Q., Hahn, M.W., Lanfear, R. 2020a. New methods to calculate concordance factors for phylogenomic datasets. Molecular biology and evolution, 37:2727-2733.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R. 2020b. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol, 37:1530-1534.

Mirarab, S., Bayzid, M.S., Warnow, T. 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Systematic Biology, 65:366-380.

Molina, J., Hazzouri, K.M., Nickrent, D., Geisler, M., Meyer, R.S., Pentony, M.M., Flowers, J.M., Pelser, P., Barcelona, J., Inovejas, S.A. 2014. Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae). Molecular Biology and Evolution, 31:793-803.

Nickrent, D., Duff, R.J. 1996. Molecular studies of parasitic plants using ribosomal RNA. Advances in parasitic plant research:28-52.

Nickrent, D., Starr, E. 1994. High rates of nucleotide substitution in nuclear small-subunit (18S) rDNA from holoparasitic flowering plants. Journal of Molecular Evolution, 39:62-70.

Nikolov, L.A., Davis, C.C. 2017. The big, the bad, and the beautiful: Biology of the world's largest flowers. Journal of Systematics and Evolution, 55:516-524.

Nikolov, L.A., Endress, P.K., Sugumaran, M., Sasirat, S., Vessabutr, S., Kramer, E.M., Davis, C.C. 2013. Developmental origins of the world’s largest flowers, Rafflesiaceae. Proceedings of the National Academy of Sciences, 110:18578-18583.

Nikolov, L.A., Tomlinson, P., Manickam, S., Endress, P.K., Kramer, E.M., Davis, C.C. 2014. Holoparasitic Rafflesiaceae possess the most reduced endophytes and yet give rise to the world's largest flowers. Annals of botany, 114:233-242.

Ogilvie, H.A., Bouckaert, R.R., and Drummond, A.J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular biology and evolution 34, 2101-2114.

971 Parey, E., Louis, A., Montfort, J., Bouchez, O., Roques, C., Iampietro, C., Lluch, J., Castinel, A.,
972 Donnadieu, C., Desvignes, T. 2023. Genome structures resolve the early
973 diversification of teleost fishes. Science, 379:572-575.
974 Petersen G., Anderson B., Braun H.-P., Meyer E.H., Møller I.M. 2020. Mitochondria in
975 parasitic plants. Mitochondrion. 52:173–182.
976 Roch S, Nute M, Warnow T. 2019. Long-branch attraction in species tree estimation:
977 inconsistency of partitioned likelihood and topology-based summary methods. Syst
978 Biol 68:281–297.
979 Puttick, M.N., Morris, J.L., Williams, T.A., Cox, C.J., Edwards, D., Kenrick, P., Pressel, S.,
980 Wellman, C.H., Schneider, H., Pisani, D. 2018. The interrelationships of land plants
981 and the nature of the ancestral embryophyte. Current Biology, 28:733-745. e732.
982 Rambaut, A., Grassly, N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of
983 DNA sequence evolution along phylogenetic trees. Bioinformatics, 13:235-238.
984 Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., Ripley, M.B. 2013.
985 Package 'mass'. Cran r, 538:113-120.
986 Sayyari, E., Mirarab, S. 2016. Fast coalescent-based computation of local branch support
987 from quartet frequencies. Molecular biology and evolution, 33:1654-1668.
988 Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst
989 Biol, 51:492-508.
990 Shimodaira, H., Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with
991 applications to phylogenetic inference. Mol Biol Evol, 16:1114-1114.
992 Shimodaira, H., Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic
993 tree selection. Bioinformatics, 17:1246-1247.
994 Si Quang, L., Gascuel, O., Lartillot, N. 2008. Empirical profile mixture models for
995 phylogenetic reconstruction. Bioinformatics, 24:2317-2323.
996 Smith S.A., Brown J.W., Hinchliff C.E. 2013. Analyzing and synthesizing phylogenies using
997 tree alignment graphs. PLOS Comput. Biol. 9:e1003223.
998 Smith, S.A., Brown, J.W., Walker, J.F. 2018. So many genes, so little time: A practical
999 approach to divergence-time estimation in the genomic era. PloS one, 13:e0197433.
1000 Song, S., Liu, L., Edwards, S.V., Wu, S. 2012. Resolving conflict in eutherian mammal
1001 phylogeny using phylogenomics and the multispecies coalescent model. Proceedings
1002 of the National Academy of Sciences, 109:14942-14947.
1003 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1004 large phylogenies. Bioinformatics, 30:1312-1313.
1005 Suyama, M., Torrents, D., Bork, P. 2006. PAL2NAL: robust conversion of protein sequence
1006 alignments into the corresponding codon alignments. Nucleic Acids Res, 34:W609-
1007 W612.
1008 Tanabe, A. 2012. Phylogears version 2.2. 2012.02. 13.
1009 Teixeira-Costa, L., Davis, C.C. 2021. Life history, diversity, and distribution in parasitic
1010 flowering plants. Plant Physiology, 187:32-51.
1011 Wang, H.-C., Minh, B.Q., Susko, E., Roger, A.J. 2018. Modeling site heterogeneity with
1012 posterior mean site frequency profiles accelerates accurate phylogenomic
1013 estimation. Systematic biology, 67:216-235.
1014 Wang, H.-c., Singer, G.A., Hickey, D.A. 2004. Mutational bias affects protein evolution in
1015 flowering plants. Molecular biology and evolution, 21:90-96.

1016 Wee, S.L., Tan, S.B., Jürgens, A. 2018. Pollinator specialization in the enigmatic Rafflesia
1017 cantleyi: a true carrion flower with species-specific and sex-biased blow fly
1018 pollinators. Phytochemistry, 153:120-128.
1019 Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J. 2017. Package 'corrplot'. Statistician,
1020 56:e24.
1021 Wurdack, K.J., Davis, C.C. 2009. Malpighiales phylogenetics: gaining ground on one of the
1022 most recalcitrant clades in the angiosperm tree of life. American Journal of Botany,
1023 96:1551-1570.
1024 Xi, Z., Bradley, R.K., Wurdack, K.J., Wong, K., Sugumaran, M., Bomblies, K., Rest, J.S., Davis,
1025 C.C. 2012. Horizontal transfer of expressed genes in a parasitic flowering plant. BMC
1026 genomics, 13:1-8.
1027 Xi, Z., Liu, L., Rest, J.S., Davis, C.C. 2014. Coalescent versus concatenation methods and the
1028 placement of Amborella as sister to water lilies. Systematic biology, 63:919-932.
1029 Xi, Z., Wang, Y., Bradley, R.K., Sugumaran, M., Marx, C.J., Rest, J.S., Davis, C.C. 2013. Massive
1030 mitochondrial gene transfer in a parasitic flowering plant clade. PLoS Genet,
1031 9:e1003265.
1032 Yang, Y., Smith, S.A. 2014. Orthology inference in nonmodel organisms using
1033 transcriptomes and low-coverage genomes: improving accuracy and matrix
1034 occupancy for phylogenomics. Molecular biology and evolution, 31:3081-3092.
1035 Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol,
1036 24:1586-1591.
1037 Young N.D., dePamphilis C.W. 2005. Rate variation in parasitic plants: correlated and
1038 uncorrelated patterns among plastid genes of different function. BMC Evol. Biol.
1039 5:16.
1040 Zhang, C., Sayyari, E., Mirarab, S. 2017. ASTRAL-III: increased scalability and impacts of
1041 contracting low support branches. RECOMB international workshop on comparative
1042 genomics, Springer, p. 53-75.
1043 Zhao, T., Zwaenepoel, A., Xue, J.-Y., Kao, S.-M., Li, Z., Schranz, M.E., Van de Peer, Y. 2021.
1044 Whole-genome microsynteny-based phylogeny of angiosperms. Nat Commun,
1045 12:3498.
1046 Wicke S., Müller K.F., Quandt D., Bellot S., Schneeweiss G.M. 2016. Mechanistic model of
1047 evolutionary rate variation en route to a nonphotosynthetic lifestyle in plants. Proc.
1048 Natl. Acad. Sci.:201607576.
1049 Wurdack K.J., Davis C.C. 2009. Malpighiales phylogenetics: gaining ground on one of the
1050 most recalcitrant clades in the angiosperm tree of life. Am. J. Bot. 96:1551–1570.
1051 Wurdack K.J., Hoffmann P., Samuel R., de Bruijn A., van der Bank M., Chase M.W. 2004.
1052 Molecular phylogenetic analysis of Phyllanthaceae (Phyllanthoideae pro parte,
1053 Euphorbiaceae sensu lato) using plastid RBCL DNA sequences. Am. J. Bot. 91:1882–
1054 1900.
1055 Zuntini A.R., Carruthers T., Maurin O., Bailey P.C., Leempoel K., Brewer G.E., Epitawalage N.,
1056 Françoso E., Gallego-Paramo B., McGinnie C., Negrão R., Roy S.R., Simpson L., Toledo
1057 Romero E., Barber V.M.A., Botigué L., Clarkson J.J., Cowan R.S., Dodsworth S., Johnson
1058 M.G., Kim J.T., Pokorny L., Wickett N.J., Antar G.M., DeBolt L., Gutierrez K., Hendriks
1059 K.P., Hoewener A., Hu A.-Q., Joyce E.M., Kikuchi I.A.B.S., Larridon I., Larson D.A., de
1060 Lírio E.J., Liu J.-X., Malakasi P., Przelomska N.A.S., Shah T., Viruel J., Allnutt T.R.,
1061 Ameka G.K., Andrew R.L., Appelhans M.S., Arista M., Ariza M.J., Arroyo J., Arthan W.,

Bachelier J.B., Bailey C.D., Barnes H.F., Barrett M.D., Barrett R.L., Bayer R.J., Bayly M.J., Biffin E., Biggs N., Birch J.L., Bogarín D., Borosova R., Bowles A.M.C., Boyce P.C., Bramley G.L.C., Briggs M., Broadhurst L., Brown G.K., Bruhl J.J., Bruneau A., Buerki S., Burns E., Byrne M., Cable S., Calladine A., Callmander M.W., Cano Á., Cantrill D.J., Cardinal-McTeague W.M., Carlsen M.M., Carruthers A.J.A., de Castro Mateo A., Chase M.W., Chatrou L.W., Cheek M., Chen S., Christenhusz M.J.M., Christin P.-A., Clements M.A., Coffey S.C., Conran J.G., Cornejo X., Couvreur T.L.P., Cowie I.D., Csiba L., Darbyshire I., Davidse G., Davies N.M.J., Davis A.P., van Dijk K., Downie S.R., Duretto M.F., Duvall M.R., Edwards S.L., Eggli U., Erkens R.H.J., Escudero M., de la Estrella M., Fabriani F., Fay M.F., Ferreira P. de L., Ficinski S.Z., Fowler R.M., Frisby S., Fu L., Fulcher T., Galbany-Casals M., Gardner E.M., German D.A., Giaretta A., Gibernau M., Gillespie L.J., González C.C., Goyder D.J., Graham S.W., Grall A., Green L., Gunn B.F., Gutiérrez D.G., Hackel J., Haevermans T., Haigh A., Hall J.C., Hall T., Harrison M.J., Hatt S.A., Hidalgo O., Hodkinson T.R., Holmes G.D., Hopkins H.C.F., Jackson C.J., James S.A., Jobson R.W., Kadereit G., Kahandawala I.M., Kainulainen K., Kato M., Kellogg E.A., King G.J., Klejevskaja B., Klitgaard B.B., Klopper R.R., Knapp S., Koch M.A., Leebens-Mack J.H., Lens F., Leon C.J., Léveillé-Bourret É., Lewis G.P., Li D.-Z., Li L., Liede-Schumann S., Livshultz T., Lorence D., Lu M., Lu-Irving P., Luber J., Lucas E.J., Luján M., Lum M., Macfarlane T.D., Magdalena C., Mansano V.F., Masters L.E., Mayo S.J., McColl K., McDonnell A.J., McDougall A.E., McLay T.G.B., McPherson H., Meneses R.I., Merckx V.S.F.T., Michelangeli F.A., Mitchell J.D., Monro A.K., Moore M.J., Mueller T.L., Mummenhoff K., Munzinger J., Muriel P., Murphy D.J., Nargar K., Nauheimer L., Nge F.J., Nyffeler R., Orejuela A., Ortiz E.M., Palazzesi L., Peixoto A.L., Pell S.K., Pellicer J., Penneys D.S., Perez-Escobar O.A., Persson C., Pignal M., Pillon Y., Pirani J.R., Plunkett G.M., Powell R.F., Prance G.T., Puglisi C., Qin M., Rabeler R.K., Rees P.E.J., Renner M., Roalson E.H., Rodda M., Rogers Z.S., Rokni S., Rutishauser R., de Salas M.F., Schaefer H., Schley R.J., Schmidt-Lebuhn A., Shapcott A., Al-Shehbaz I., Shepherd K.A., Simmons M.P., Simões A.O., Simões A.R.G., Siros M., Smidt E.C., Smith J.F., Snow N., Soltis D.E., Soltis P.S., Soreng R.J., Sothers C.A., Starr J.R., Stevens P.F., Straub S.C.K., Struwe L., Taylor J.M., Telford I.R.H., Thornhill A.H., Tooth I., Trias-Blasi A., Udovicic F., Utteridge T.M.A., Del Valle J.C., Verboom G.A., Vonow H.P., Vorontsova M.S., de Vos J.M., Al-Wattar N., Waycott M., Welker C.A.D., White A.J., Wieringa J.J., Williamson L.T., Wilson T.C., Wong S.Y., Woods L.A., Woods R., Worboys S., Xanthos M., Yang Y., Zhang Y.-X., Zhou M.-Y., Zmarzty S., Zuloaga F.O., Antonelli A., Bellot S., Crayn D.M., Grace O.M., Kersey P.J., Leitch I.J., Sauquet H., Smith S.A., Eiserhardt W.L., Forest F., Baker W.J. 2024. Phylogenomics and the rise of the angiosperms. Nature.:1–8.

Supplementary Information

Supplementary figures

Fig S1 Correlation heatmap of ten phylogenetic metrics. The intensity of the color indicates the strength of the correlation and the dendrogram shows hierarchical clustering results. The four Rafflesiaceae-specific metrics are highlighted in red. Definitions of these phylogenetic metrics are provided in the Materials and Methods.

1107 **Fig S2** Optimum placement and divergence time estimation of Rafflesiaceae. The time tree
1108 is inferred using 81 exons under the full coalescent model implemented in starBEAST. The
1109 blue bars on the node indicate 95% confidence intervals of divergence time (million years,
1110 Ma).
1111
1112 **Fig S3** DNA composition of Malpighiales species summarized from 2141 genes. Species are
1113 ranked based on AT% from high to low and Rafflesiaceae species are highlighted in red.
1114
1115 **Fig S4** Gene genealogy interrogation (GGI) to test alternative placements of Rafflesiaceae.
1116 Colored lines represent the cumulative number of genes (x-axes) supporting each
1117 hypothesis with the highest probability (rank 1) and their associated *p*-values (y-axes)
1118 according to the approximately unbiased (AU) topology test (A), the Kishino-Hasegawa
1119 (KH) topology test (B), and the Shimodaira–Hasegawa (SH) topology test (C). Values above
1120 the dashed line indicate all rank 1 hypotheses that are significantly better than the
1121 alternatives (p-value < 0.05), whereas those below the dashed line are not statistically
1122 significant.
1123
1124 **Fig S5** Distribution of synonymous substitutions per synonymous site (d_s) in free-living
1125 and parasitic Malpighiales. d_s values for *Manihot* (Euphorbiaceae), *Populus* (Salicaceae), and
1126 *Sapria* (Rafflesiaceae) are inferred using the free ratio model implemented in PAML.
1127
1128
1129 **Supplementary tables**
1130 **Table S1** Taxon sampling and source of sequence data.
1131
1132 **Table S2** Phylogenetic properties of 2141 genes. These metrics include six tree-wise
1133 features for each gene including the number of species, alignment length, average nodal
1134 support measured by ultrafast bootstrap values (UFBP), total tree length, root-to-tip
1135 branch length variance, and gene tree–species tree congruence. Four metrics specific to
1136 Rafflesiaceae are also calculated, including DNA compositional bias of Rafflesiaceae (chi-
1137 square test *p*-values), lengths of the branches prior to and immediately after the divergence
1138 of Rafflesiaceae, and nodal support of Rafflesiaceae measured by UFBP and non-parametric
1139 bootstrapping (BP).
1140
1141 **Table S3** Phylogenetic analysis schemes for 18 concatenation and multispecies coalescent
1142 model (MSC) based methods. Columns show the input data type, number of genes, and
1143 evolutionary models for each analysis.
1144
1145 **Table S4** Topology tests for alternative placements of Rafflesiaceae under the
1146 concatenation and coalescent models. Hypotheses 1 to 19 (H1-H19) are defined in Fig. 1
1147 and statistically significant support is marked by "**". For concatenation analyses, the
1148 number of genes, input data type, species tree inference program, and evolutionary model
1149 are provided at the top; the maximum likelihoods, AU test *p*-values, and RELL bootstrap
1150 values (bp-RELL) are calculated for each hypothesis. For coalescent analysis, the pseudo-
1151 likelihoods and likelihood ratio test *p*-values are calculated for the top three topologies.
1152

1153 **Table S5** Topology test on simulated data using concatenated DNA alignments. Hypotheses
1154 1 to 15 (H1-H15) are defined in Fig 1 and H11 is used as the true species tree for
1155 simulation. AU-test *p*-values are calculated using concatenated DNA matrices of different
1156 sizes (a-d). The lengths of simulated genes vary from 500 bp (a), 1000 bp (b), 1500 bp (c),
1157 to 2000 bp (d).
1158
1159 **Table S6** Topology test on simulated data using the coalescent based likelihood ratio test.
1160 Hypotheses 1 to 15 (H1-H15) are defined in Fig 1 and H11 is used as the true species tree
1161 for simulation. Differences in pseudo-likelihoods between the inferred best tree and the
1162 true species tree (H11) are calculated for datasets with different gene lengths and numbers
1163 (a-d). For the biggest dataset (gene length =2000 bp; d), we calculate the likelihood ratio
1164 test *p*-values for a randomly selected alternative tree H5, which cannot be rejected in any
1165 test.
1166
1167 **Supplementary Data**
1168 **Data S1** Optimum species trees inferred from the 18 explorative phylogenetic analyses in
1169 newick format. Node labels indicate branch support. See Table S3 for input data and
1170 program parameters.