A consensus phylogenomic approach highlights the ancient rapid radiation of Ericales

Drew A. Larson[1,4], Joseph F. Walker[2], Oscar M. Vargas[3] and Stephen A. Smith[1]

[1]Department of Ecology & Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

[2]Sainsbury Laboratory (SLCU), University of Cambridge, Cambridge, CB2 1LR, UK

[3]Department of Ecology & Evolutionary Biology, University of California, Santa Cruz, CA, 95064, USA

[4]Author for correspondence: D. A. Larson (larsonda@umich.edu)

Manuscript received _____; revision accepted _____.

Running Head: A consensus approach to resolving the Ericales phylogeny

ABSTRACT

Premise of study: Large genomic datasets offer the promise of resolving historically recalcitrant phylogenetic problems. However, different methodologies can yield conflicting results, especially when diversification occurs rapidly. Here, we used an array of dataset filtering strategies and species tree methods to infer a consensus topology of Ericales and explored sources of uncertainty associated with an ancient radiation.

Methods: We used a hierarchical clustering approach, along with tree-based homology and orthology detection, to generate multiple phylogenomic datasets. Support for species relationships was inferred from multiple lines of evidence including shared gene duplications, gene tree conflict, gene-wise edge-based analyses, concatenation, and coalescent-based species tree methods and summarize these results in a consensus framework.

Key Results: Our consensus approach supported a topology largely concordant with the current theorized relationships, but suggests that the data are not capable of resolving several early relationships due to lack of informative characters, sensitivity to methodology, and extensive gene tree conflict. We find evidence of one or more paleopolyploidy events before the radiation of ericalean families that likely contributes to the high levels of gene tree conflict observed.

Conclusions: Our approach provides a novel hypothesis regarding the history of Ericales and confidently resolves several nodes. However, we also demonstrate that some ancient divergences are unresolvable with our current data. Whether this is because of rapid and unresolvable ancient speciation or lack of data needs to be addressed with additional data collection efforts. The role that paleopolyploidy plays in generating gene tree conflict warrants further investigation.

Keywords: consensus topology; Ericales; gene duplication; gene tree conflict; multifurcation, phylogenetic uncertainty; phylogenomics; polytomy; rapid radiation; whole genome duplication

INTRODUCTION

The flowering plant clade Ericales contains several ecologically important lineages that shape the structure and function of ecosystems including tropical rainforests (e.g. Lecythidaceae, Sapotaceae, Ebenaceae), heathlands (e.g. Ericaceae), and open habitats (e.g. Primulaceae) around the globe (ter Steege et al., 2006; Hedwall et al., 2013; He et al., 2014; Memiaghe et al., 2016; Moquet et al., 2017). With 22 families comprising ca. 12,000 species (Chase et al., 2016; Stevens, 2001 onward), Ericales are a diverse and disparate clade with an array of economically and culturally important plants. These include agricultural crops such as blueberries (Ericaceae), kiwifruits (Actinidiaceae), sapotas (Sapotaceae), Brazil nuts (Lecythidaceae), and tea (Theaceae) as well as ornamental plants such as cyclamens and primroses (Primulaceae), rhododendrons (Ericaceae), and phloxes (Polemoniaceae). The group has also given rise to multiple parasitic lineages and the carnivorous American pitcher plants (Sarraceniaceae). Although Ericales has been a well-recognized clade throughout the literature (Chase et al., 1993; Anderberg et al., 2002; Schönenberger et al., 2005; Rose et al., 2018), the evolutionary relationships among major clades within Ericales remain contentious.

One of the first molecular studies investigating these deep relationships used three plastid and two mitochondrial loci; the authors concluded that the dataset was unable to resolve major familial relationships (Anderber et al., 2002). These relationships were revisited with 11 loci (two nuclear, two mitochondrial, and seven chloroplast), where maximum parsimony and Bayesian analyses provided support for the resolution of several early diverging lineages (Schönenberger et al., 2005). Recent work employing three nuclear, nine mitochondrial, and 13 chloroplast loci in a concatenated supermatrix consisting of 49,435 aligned sites and including 4,531 ericalean species but with 87.6% missing data, provided the most statistically robust

70 hypothesis to date (Rose et al., 2018). Despite the broad sampling utilized in Rose et al. (2018),
71 many relationships remain poorly supported, including several deep divergences that seem to be
72 the result of an ancient, rapid radiation. Moreover, these and other studies have recovered several
73 conflicting interfamilial relationships, highlighting the need to investigate possible biological and
74 methodological explanations for this topological incongruence (Anderber et al., 2002; Bremer et
75 al., 2002; Schönenberger et al., 2005).

76 Large genomic datasets can be used to help understand conflicting species tree results
77 and as these data have become more affordable, it is possible to conduct thorough investigations
78 of elusive relationships across the Tree of Life. Yet, despite the increasing availability of
79 genome-scale datasets, many relationships remain controversial, as research groups recover
80 different answers for the same evolutionary questions, often with seemingly strong support (Shen
81 et al., 2017). One benefit of genome-scale data for phylogenetics (i.e. phylogenomics) is the
82 ability to examine conflicting signal within and among datasets and a key finding in the
83 phylogenomics literature has been the high prevalence of gene tree conflict at contentious nodes
84 (Brown et al. 2017b, Reddy et al., 2017; Vargas et al., 2017; Walker et al., 2018a). The
85 differences seen across previous analyses may in fact be the result of real biological conflict
86 among the molecular data.. However, biological sources of conflict (e.g. hybridization,
87 incomplete lineage sorting, horizontal gene transfer) can also provide valuable information
88 regarding the evolutionary history of lineages. By identifying this conflict, it becomes possible to
89 locate where focused analyses are warranted and future sampling efforts might prove useful. One
90 prominent method for examining genomic conflict is analyzing the relationships among gene
91 trees, which may be obtained through a variety of methods including from transcriptome data
92 (Yang and Smith, 2014). Transcriptomes provide data on hundreds to thousands of coding

93 sequences per sample, and have elucidated many of the most contentious relationships in the
94 green plant phylogeny (e.g. Simon et al., 2012; Wickett et al., 2014; Walker et al., 2017).
95 Transcriptomes also provide the necessary information to identify gene and genome duplications.
96 Gene duplications may be associated with important molecular evolutionary events but also
97 represent a potentially informative event in a lineage's history and shared duplications can
98 therefore provide additional evidence for how species are related. These multiple lines of
99 evidence regarding evolutionary history can be applied as metrics of support to generate robust
100 phylogenetic hypotheses.

101 In this study, we sought to understand the evolutionary history of Ericales and asked
102 whether the available genome-scale data robustly supported the resolution of deep relationships
103 in the clade. We leveraged coding sequences from hundreds/thousands of genes to infer major
104 evolutionary events, such as ancient gene/genome duplication, rates of evolution, and conflicting
105 signal among genes. In applying a consensus approach that considers several methodological
106 alternatives in an effort to resolve the phylogeny, we allow the possibility that methods might
107 disagree and that there may not be the data available to support a single, bifurcating species tree
108 may. While more data may be available, in cases where whole genomes and transcriptomes have
109 been sequenced, a polytomy may represent a more accurate representation of the evolutionary
110 history of a clade than a fully bifurcating species tree. Our investigation of the evolution of the
111 Ericales with genomic data is particularly well-suited to instantiate a conversation about when
112 polytomies may be considered and to what extent biological polytomies occur across the Tree of
113 Life.
114
115

116 **MATERIALS AND METHODS**

117 ***Taxon sampling and de novo assembly procedures*—**We assembled a dataset consisting of
118 coding sequence data from 97 genomes and transcriptomes from the 1KP project and NCBI
119 databases (Appendix S1; see Supplemental Data with this article). Samples assembled from raw
120 reads were done so with Trinity version 2.5.1 using default parameters and the "--trimmomatic"
121 option (Kodama et al. 2011; Grabherr et al., 2011; Matasci et al., 2014). Assembled reads were
122 translated using TransDecoder v5.0.2 (Haas et al., 2013) with the option to retain blast hits to a
123 custom protein database consisting of *Beta vulgaris*, *Arabidopsis thaliana*, and *Daucus carota*
124 obtained from Ensembl plants (https://plants.ensembl.org/). The translated amino acid sequences
125 from each assembly were reduced by clustering with CD-HIT version 4.6 with settings -c 0.995 -
126 n 5 (Fu et al., 2012). As a quality control step, the chloroplast genes *rbcL* and *matK* were
127 extracted from each assembly using a custom script (available on Dryad [TBD]) and then queried
128 using BLAST against the NCBI online database (Altschul et al., 1990). In cases where the top
129 hits were to a species not closely related to that of the query, additional sequences were
130 investigated and if contamination, misidentification, or other problems seemed likely, the
131 transcriptome was not included in further analyses or homolog clustering. The final
132 transcriptome sampling included 86 ingroup taxa spanning 17 of the 22 families within Ericales
133 as recognized in APG VI (Appendix S1; Chase et al., 2016).

134 ***Homology inference*—**We used the hierarchical clustering procedure from Walker et al. (2018b)
135 for homology identification. In short, the method involves performing all-by-all blast procedure
136 on user-defined clades; homolog clusters identified within each clade are then combined
137 recursively with clusters of a sister clade based on sequence similarity until clusters from all
138 clades have been combined. To distribute taxa among groups for clustering we estimated a

139 species tree with RAxML v8.2.12 (Stamatakis, 2014) using a concatenated matrix of *rpoC2*,
140 *rbcL*, *nhdF*, and *matK*, identified in the quality control step. The resulting tree was used to
141 manually assign taxa to one of eight clades for initial homolog clustering (Appendix S2).
142 Homolog clustering within each group was performed following the methods of Yang and Smith
143 (2014).

144 Nucleotide sequences for the inferred homologs within each group were aligned with
145 MAFFT v7.271 (Katoh et al., 2002; Katoh and Standley, 2013) and trimmed with the phyx
146 command pxclsq (Brown et al., 2017a) such that columns with less than 10% occupancy were
147 removed. Homolog trees were estimated with RAxML, unless the homolog had >500 tips, in
148 which case FastTree v2.1.8 was used (Price et al., 2010). Tips with branch lengths longer than
149 1.5 substitutions/site were trimmed because of their potential to reflect misidentified homology
150 (Yang and Smith, 2014). When a clade is formed of sequences from a single taxon, it likely
151 represents either in-paralogs or alternative splice sites and since neither of these provide
152 phylogenetic information, we retained only the tip with the most character rich sequence data.
153 This procedure was repeated with refined clusters twice using the same settings. Homologs
154 between each of the eight groups were then combined (https://github.com/jfwalker/Clustering)
155 and homolog trees were again estimated and refined with the same settings except that internal
156 branches longer than 1.5 subs/site were also cut—again to reduce the possibility of misidentified
157 homology. Trees were re-estimated and refined by cutting terminal branches longer than 0.8
158 substitutions/site and internal branches longer than 1.0 substitutions/site. The final homolog set
159 contained 9469 clusters.

160 ***Initial ortholog identification and species tree estimation*—**Putative orthologs were extracted
161 from homolog trees using the rooted tree (RT) method which allows for robust orthology

162 detection even after genome duplications (Yang and Smith, 2014). Orthologs were required to
163 have at least 50 taxa and seven cornalean taxa as well as *Arabidopsis thaliana* and *Beta vulgaris*
164 were used as outgroups (Appendix S1). Orthologs were not extracted from homolog groups with
165 more than 5000 tips because of the uncertainty in reconstructing very large homolog trees
166 (Walker et al., 2018b). Here, we use the term ortholog to describe clusters of sequences that have
167 been inferred to be monophyletic based on their position within inferred homolog trees after
168 accounting for gene duplication. This usage reflects our interest in addressing phylogenetic
169 questions, rather than those of gene functionality. The tree-aware ortholog identification
170 employed here provides the best defense available against misidentified orthology that could
171 mislead phylogenetic analyses (Eisen, 1998; Gabaldón, 2008; Brown and Thomson, 2016).
172 The resulting ortholog trees were then filtered to require at least one euasterid taxon,
173 *Helianthus annuus* or *Solanum lycopersicum*, for use as outgroups, and if both were present the
174 two were required to be monophyletic. This was done to ensure that orthologs could be rooted in
175 subsequent analyses. Terminal branches longer than 0.8 were again trimmed, resulting in a
176 refined dataset containing 387 orthologs. Final nucleotide alignments were estimated with
177 PRANK v.150803 (Löytynoja, 2014) and cleaned for a minimum occupancy of 30% occupancy
178 using the pxclsq function in phyx. Alignments for the 387 orthologs were concatenated using
179 pxcat in phyx and a maximum likelihood (ML) species tree and rapid bootstrap support (200
180 replicates) with was inferred using RAxML v8.2.4 with the command raxmlHPC-PTHREADS
181 and a separate GTRCAT model of evolution estimated for each partition; this resulted in a
182 topology we refer to as the maximum likelihood topology (MLT). In separate runs, 200 regular
183 bootstraps with and without the MLT as a starting tree were conducted using the -b option. Trees
184 for each ortholog were also estimated individually using RAxML, with the GTRCAT model of

185 evolution and 200 rapid bootstraps. A coalescent-based maximum quartet support species tree
186 (MQSST) was estimated using ASTRAL v5.6.2 (Zhang et al., 2018) with the resulting 387
187 ortholog trees. In order to investigate the possible effect of ML tree search algorithm on
188 phylogenetic inference, an ML tree for the 387 ortholog supermatrix was estimated using
189 RAxML v8.2.11 and the command raxmlHPC-PTHREADS-AVX and with IQ-TREE v1.6.1
190 with -m GTR+Γ -n 0 and model partitions for each ortholog specified with the -q option.
191 ***Gene-wise log-likelihood comparisons***—Gene-wise likelihood support for the MLT and an
192 alternative backbone topology recovered by the consensus all 200 rapid bootstraps (i.e. the rapid
193 bootstrap topology, RBT) were investigated using a two-topology analysis (Shen et al., 2017;
194 Walker et al., 2018a). A three-topology comparison was also conducted by calculating the gene-
195 wise log-likelihoods of a third topology where the MLT was modified such that the inter-familial
196 backbone relationships were those recovered in the ASTRAL topology (AT). This constructed
197 tree was used in lieu of the actual ASTRAL topology to ensure that the within-family topological
198 differences did not affect the likelihood calculations. In both tests, the ML score for each gene is
199 calculated while constraining the topology under multiple alternatives. Branch lengths are
200 optimized for each input topology and GTR+Γ is optimized for each supermatrix partition for
201 each topology (Shen et al., 2017; Walker et al., 2018a). Results from both comparisons were
202 visualized using a custom R script.
203 An edge-based analysis was performed to compare the likelihoods of competing
204 topologies while allowing gene tree conflict to exist outside the relationship of interest (Walker
205 et al., 2018a). Our protocol was similar to that of Walker et al. (2018a), except that instead of a
206 defined "TREE SET", we used constraint trees with clades defining key bipartitions
207 corresponding to each of three competing topologies using the program EdgeTest.py (Walker et

208 al., 2019). The likelihood for each gene was calculated using RAxML-ng (Kozlov et al., 2019)
209 with the GTR+Γ model of evolution, using the brlopt nr_safe option and an epsilon value of
210 0.000001 while a given relationship was constrained. The log-likelihood of each gene was then
211 summed to give a likelihood score for that relationship. An additional edge-based analysis was
212 conducted to investigate the placement of Ebenaceae using the program Phyckle (Smith et al.,
213 2018), which uses a supermatrix and a set of constraint trees specifying conflicting relationships
214 and reports, for each gene, the ML as well as the difference between the best and second-best
215 topology. This allows quantification of gene-wise support at a single edge as well as a metric of
216 how strongly the relationships is supported in terms of likelihood.
217 ***Gene duplication comparative analysis*—**Homolog clusters were filtered such that sequences
218 shorter than half the median length of their cluster and clusters with more than 4000 sequences
219 were removed. Homolog trees were then re-estimated with IQ-TREE with the GTR+Γ model and
220 SH-aLRT support followed by cutting of internal branches longer than 1.0 and terminal branches
221 longer than 0.8 inferred substitutions per base pair. Rooted ingroup clades were extracted with
222 the procedure from Yang et al. (2015) with all non-ericalean taxa as outgroups and gene
223 duplications were inferred with the program phyparts, requiring at least 50% SH-aLRT support
224 in order to avoid the inclusion of very poorly supported would-be duplications (Smith et al.,
225 2015). By mapping gene duplications in this way to competing topological hypotheses (i.e. the
226 MLT, RBT, and AT), as well as several hypothetical topologies not recovered, we determined
227 the number of duplications uniquely shared among several key ericalean clades.
228 Owing to how duplications are inferred using tree-based methods, the location to which
229 duplications are mapped depends on the topology of the phylogeny; a gene duplication can
230 appear along a branch other than that which the homolog trees actually contains evidence

231 because duplications are mapped to the bipartition in the species tree that contains all the taxa
232 that share a given duplication in a homolog tree. For example, if in a homolog tree, taxon A and
233 taxon B share a gene duplication, but in the species tree taxon C is sister to taxon A and taxon B
234 sister to those, then the duplication will be mapped to the branch that includes all three taxa,
235 because the duplication is mapped to the smallest clade that includes both taxon A and taxon B.
236 A differently specified topology can allow the relevant duplication to instead be mapped to the
237 correct, more exclusive branch that specifies the clade for which there is evidence of that
238 duplication in the homolog tree (i.e. only taxa A and B in the previous example). Once it has
239 been determined how many duplications are actually present in the homolog trees, comparisons
240 between competing relationships can be made. It is important to note that incomplete taxon
241 sampling and other biases should be considered when applying such a comparative test of gene
242 duplication number. Assuming there are duplicated genes present in the taxa of interest, clades
243 with a greater number of sampled taxa or more complete transcriptomes will likely share more
244 duplications simply due to the fact that more will have been sampled in the dataset. Furthermore,
245 the location to which a duplication is mapped will depend on which assemblies contain that
246 transcript, whether or not that gene is present in the genome of any particular organism.
247 Therefore, we used gene duplications shared amongst clades as an additional, relative metric of
248 topological support capable of corroborating other results (i.e. that if clades share many gene
249 duplications unique to them, they are more likely to be closely related), while recognizing that
250 the absolute number of duplications shared by various clades are impacted by imperfect
251 sampling.
252 ***Synthesizing support for competing topologies*—**Support from each of the main tree hypotheses
253 (MLT, RBT and AT) was reviewed and the most commonly supported backbone of familial

254 relationships was determined. Because the comparative gene duplication analysis and constraint
255 tree analyses both supported the RBT over other candidates, this corroborated topology was the
256 focus of additional measures of support. Quartet support was assessed on the RBT using the
257 program Quartet Sampling with 1000 replicates (Pease et al., 2018). We used this procedure to
258 measure quartet concordance (QC), quartet differential (QD), quartet informativeness (QI), and
259 quartet fidelity (QF). Briefly, QC measures how often the concordant relationship is recovered
260 with respect to other possible relationships, QD helps identify if a relationship has a dominant
261 alternative, and QI correspond to the ability of the data to resolve a relationship of interest, where
262 a quartet that is at least 2.0 log-likelihood (LL) better than the alternatives is considered
263 informative. Finally, QF aids in identifying rogue taxa (Pease et al., 2018). Gene conflict with
264 the corroborated topology was assessed using the bipartition method as implemented in phyparts
265 using gene trees for the 387 orthologs, which were first rooted on the outgroups *Helianthus*
266 *annuus* and *Solanum lycopersicum* using a ranked approach with the phyx command pxrr. Gene
267 conflict was assessed both requiring node support (BS≥70) and without any support
268 requirement; the support requirement should help reduce noise in the analysis, but we also ran
269 the analysis without a support requirement to ensure that potentially credible (but only partially-
270 supported) bipartitions were not overlooked. Results for both were visualized using the program
271 phypartspiecharts (https://github.com/mossmatters/phyloscripts/).
272 ***Expanded ortholog datasets*—**In order to explore the impact of dataset construction, orthologs
273 were inferred from the homolog trees as for the 387 ortholog set but with modifications to taxon
274 requirements and refinement procedures described below. For each dataset, sequences were
275 aligned separately with both MAFFT and PRANK and cleaned for 30% occupancy. A
276 supermatrix was constructed with each and an ML tree was estimated with IQ-TREE with and

277 without a separate GTR+Γ model partition for each ortholog. Individual ortholog trees were
278 estimated with RAxML and used to construct an MQSST with ASTRAL.
279 *2045 ortholog set*—Orthologs were filtered such that there was no minimum number of taxa and
280 at least two tips from each of the following five groups: 1) Primulaceae; 2) Polemoniaceae and
281 Fouquieriaceae; 3) Lecythidaceae; 4) outgroups including *Solanum* and *Helianthus* as well as
282 Marcgraviaceae and Balsaminaceae (i.e. the earliest diverging ericalean clade in all previous
283 analyses); and 5) all other taxa. This resulted in a dataset with 2045 orthologs. Ortholog tree
284 support for conflicting placements of Ebenaceae was assessed for the PRANK-aligned orthologs
285 using Phyckle.
286 *4682 ortholog set*—Orthologs were not filtered for any taxon requirements. Sequences were
287 aligned with MAFFT, ortholog trees were estimated with RAxML and terminal branches longer
288 than 0.8 were trimmed. Sequences were then realigned separately with MAFFT and PRANK and
289 cleaned as before.
290 *1899, 661, and 449 ortholog sets*—To assess the effect of requiring *Helianthus* and *Solanum*
291 outgroups in the 387 ortholog set and to further explore the effect of taxon requirements on the
292 inferred topology, each homolog tree that produced an ortholog in that dataset was re-estimated
293 in IQ-TREE with SH-aLRT support. Calculating this support allowed visual assessment of
294 whether uncertain homolog tree construction was affecting the ortholog identification process.
295 This did not appear to be a major issue as orthologs (i.e. clades of ingroup tips subtended by
296 outgroup tips) within homolog trees typically received strong support as monophyletic.
297 Following homolog tree re-estimation, orthologs were identified as described above with no
298 minimum taxa requirement. Sequences were aligned with MAFFT and any taxon with >75%
299 missing data was removed. Filtered alignments were then realigned separately with both MAFFT

300 and PRANK, and cleaned for 30% column occupancy. This resulted in a dataset with 1899
301 orthologs. Two subsets of this 1899 ortholog set were generated by requiring a minimum of 30
302 and 50 taxa, resulting in 661 and 449 orthologs respectively. Gene tree conflict was assessed for
303 the 449 MAFFT-aligned dataset by rooting all ortholog trees on all taxa in Balsaminaceae and
304 Marcgraviaceae (i.e. the balsaminoid Ericales) since all previous analyses showed this clade to
305 be sister to the rest of Ericales; phyparts was then used to map ortholog tree bipartitions to the
306 ML tree for this dataset and the results were visualized with phypartspiecharts.
307 ***Estimating substitutions supporting contentious clades*—**In order to gain additional insight into
308 the amount of signal present in ortholog alignments informing various relationships, we
309 developed a custom script that identifies clades of interest within an ortholog tree and uses the
310 estimated branch length leading to that clade, multiplied by the length of the corresponding
311 sequence alignment, to estimate the number of substitutions implied by that branch. Applying
312 this approach to the trees from the 449 ortholog MAFFT-aligned dataset with appropriate taxon
313 sampling to allow rooting on a member of the balsaminoid Ericales, we calculated the
314 approximate number of substitutions that informed the branch leading to the most recent
315 common ancestor (MRCA) of two or more of the following: Primulaceae,
316 Polemoniaceae+Fouquieriaceae, Lecythidaceae, Ebenaceae, and Sapotaceae. In addition, we
317 assessed substitution support for several non-controversial relationships, namely that each of the
318 following was monophyletic: Polemoniaceae+Fouquieriaceae, Lecythidaceae, and the non-
319 balsaminoid Ericales. Mean and median values for substitution support were calculated and a
320 distribution of these values was plotted using custom R scripts.
321 ***Synthesis of uncertainty, consensus topology, and genome duplication inference*—**We took
322 into consideration all previous results, including those of the expanded datasets, gene tree

323 conflict, and substitution support, to determine which relationships were generally well-
324 supported by the results of this study, and which were not. In cases where an interfamilial or
325 intrafamilial relationship remained irresolvable when considering the preponderance of the
326 evidence, that relationship was not included in the consensus topology. Gene duplications were
327 mapped to the consensus topology using the methods of the comparative duplication analysis
328 described above. Internodes on inferred species trees with notably high numbers of gene
329 duplications were used as one line of evidence for assessing putative whole genome duplications.
330 Further investigation into possible genome duplications was conducted by plotting the number of
331 synonymous substitutions (K_s) between paralogs in K_s plots according to the methods of Yang et
332 al. (2015) after removing sequences shorter than half the median length of their cluster.
333 Multispecies K_s values for selected combinations of taxa were generated according to the
334 methods of Wang et al. (2018). The effect of evolutionary rate-heterogeneity among species was
335 investigated by conducting a multispecies K_s analysis of each non-balsaminoid ericalean taxon
336 against *Impatiens balsamifera*, since all evidence suggests each of these pairs have the same
337 MCRA. Because of this, the ortholog peak resulting from each of these pairs represents the same
338 speciation event, and therefore differences in the location of this peak in the K_s plots should be
339 the result of differences in evolutionary rate among species. Plots where the location of the
340 ortholog peak was ambiguous were not considered for this rate-heterogeneity analysis. All single
341 and multispecies K_s plots were generated using custom R scripts.

342 **RESULTS**

343 ***Initial 387 ortholog dataset***—The concatenated supermatrix for the 387 ortholog dataset
344 contained 441,819 aligned sites with 76.1% ortholog occupancy and 57.8% matrix occupancy.
345 The MLT recovered by RAxML v8.2.4 is shown in Figure 1. A consensus of the rapid

346 bootstrapping trees unanimously supported a backbone topology (i.e. the RBT) that differed from
347 that of the MLT by one relationship. The MLT contained the clade
348 Polemoniaceae+Fouquieriaceae as sister to a clade consisting of Ebenaceae, Sapotaceae, and
349 what is referred to here as the "Core" Ericales: the clade that includes the Actinidiaceae,
350 Roridulaceae, Sarraceniaceae, Theaceae, Diapensiaceae, Styracaceae, Symplocaceae,
351 Pentaphylacaceae, and Ericaceae. The RBT instead placed Lecythidaceae sister to Ebenaceae,
352 Sapotaceae, and the Core (a hypothetical clade herein referred to as ESC). A gene-wise log-
353 likelihood analyses comparing these two topologies is shown in Appendix S3. The cumulative
354 log-likelihood difference between the MLT and RBT was approximately 3.57; there were 27
355 orthologs that supported the MLT over the RBT by a score larger than this and the exclusion of
356 any of these from the supermatrix could cause the RBT to become the topology with the best
357 likelihood. Both of these topologies were considered as candidates in a search for a corroborated
358 interfamilial backbone. Full bootstrapping resulted in 6.5% support for
359 Polemoniaceae+Fouquieriaceae sister to ESC unless the ML tree was given as a starting tree,
360 under which conditions the that topology received 100% bootstrap support. Unguided, full
361 bootstrapping instead provided strong support for a clade consisting of Lecythidaceae and ESC
362 (90%, Appendix S4). The ASTRAL topology (AT) for the 387 ortholog set is available in
363 Appendix S5. Interfamilial relationships in the AT were largely congruent with those of the MLT
364 except that Primulaceae was recovered as sister to Lecythidaceae+ESC but with only 54% local
365 posterior probability. A total of seven intrafamilial relationships differed between the AT and
366 MLT. The AT was considered as a third candidate for an interfamilial backbone.
367 There was an effect of the ML tree search algorithm that may be important to note for
368 future phylogenomic studies. Re-conducting an ML tree search for the 387 ortholog supermatrix

369 with RAxML v8.2.11 using PTHREADS-AVX architecture and -m GTRCAT or with IQ-TREE
370 and GTR+Γ resulted in a species tree with a different topology than under any other conditions
371 for the 387 ortholog dataset. The final log-likelihood of the original ML tree was -5948732.774.
372 Using the PTHREADS-AVX architecture returned a final likelihood of -5948752.263, while the
373 best log-likelihood reported by IQ-TREE was -5949077.382 (19.489 and 344.607 log-likelihood
374 points worse than the original results, respectively).
375 ***Gene-wise log-likelihood comparative analysis across three candidate topologies***—The results
376 from the gene-wise comparisons of likelihood contributions showed that there were 155
377 orthologs in the dataset that most strongly supported the MLT, 90 supported the RBT, and 142
378 supported the AT (Appendix S6). The AT had a cumulative log-likelihood score that was >300
379 points worse, even though more individual genes support the AT over the RBT (Appendix S6).
380 ***Edge-based comparative analyses across three candidate topologies***—Of the three candidate
381 topologies investigated by constraining key edges, Lecythidaceae sister to ESC received the best
382 score, while Primulaceae sister to ESC received the worst (Appendix S7). Similarly, the clade
383 Lecythidaceae+Polemoniaceae+Fouquieriaceae+ESC received a better likelihood score than the
384 clade Lecythidaceae+Primulaceae+ESC. Regarding the placement of Ebenaceae for this dataset,
385 Ebenaceae+Sapotaceae was supported by the highest number of orthologs whether or not two
386 log-likelihood support difference was required (Appendix S8). However, a number of orthologs
387 support each of the investigated placements for Ebenaceae and less than half of genes supported
388 any placement over the next best alternative by at least two log-likelihood points.
389 ***Gene duplication comparative analysis***—Gene duplications mapped to each candidate backbone
390 topology and the five additional hypothetical topologies revealed differing numbers of shared
391 duplications that can be used as a metric of support among candidate topologies (Fig. 2).

392 Regarding which clade is better supported as sister to ESC, Lecythidaceae uniquely shares 433
393 duplications with ESC, more than twice as many as either alternative.
394 Polemoniaceae+Fouquieriaceae shares 1300 unique duplications with Lecythidaceae+ESC, more
395 than Primulaceae which shares 626 unique duplications (Fig. 2).
396 ***A corroborated topology for the 387 ortholog set***—The above comparative analyses supported
397 one topology among the three candidates identified for the 387 ortholog dataset, namely the
398 RBT. In this tree, all taxonomic families are recovered as monophyletic. Marcgraviaceae and
399 Balsaminaceae (i.e. the balsaminoid Ericales) are found to be sister, and form a clade that is
400 sister to the rest of Ericales (i.e. the non-balsaminoid Ericales). Ebenaceae and Sapotaceae form
401 a clade that is sister to the Core. The monogeneric Fouquieriaceae is sister to Polemoniaceae. A
402 clade containing Symplocaceae, Diapensiaceae, and Styracaceae is sister to Theaceae.
403 Roridulaceae is sister to Actinidiaceae and these together form a clade that is sister to Ericaceae.
404 A grade containing Primulaceae, Polemoniaceae and Fouquieriaceae, and Lecythidaceae leading
405 to ESC is supported by the rapid bootstrapping in RAxML, duplication, and constraint-tree
406 analyses.
407 ***Quartet sampling***—In our results and discussion we consider a QC score of (≥0.5) to be strong
408 support as this signifies strong concordance among quartets (Pease et al., 2018). We found
409 varying levels of support for several key relationships in the RBT (Appendix S9). The
410 monophyly for all families received strong support (QC ≥ 0.90). There was strong support
411 (QC=0.54) for the node placing Lecythidaceae sister to ESC, while equivocal support
412 (QC=0.035) for Polemoniaceae+Fouquieriaceae sister to Lecythidaceae and ESC. Within ESC,
413 there was moderate support (QC=0.28) for Ebenaceae sister to Sapotaceae, but poor (QC=-0.13)
414 support for this clade as sister to the Core. There was strong support (QC=0.85) for the clade

415 including Symplocaceae, Diapensiaceae, and Styracaceae and moderate support (QC=0.26) for
416 this clade as sister to Theaceae. Roridulaceae was very strongly supported (QC=0.99) as sister to
417 Actinidiaceae but there was no support for this clade as sister to Ericaceae (QC=-0.23).
418 However, the monophyly of the clade that includes Roridulaceae, Actinidiaceae, Ericaceae, and
419 Sarraceniaceae received very strong support (QC=0.90).
420 The QD scores for several contentious relationships indicate that discordant quartets
421 tended to be highly skewed towards one conflicting topology as indicated by scores below 0.3
422 (Pease et al., 2018). However, the QD score for the relationship placing Lecythidaceae sister to
423 ESC in the RBT was 0.53, indicating relative equality in occurrence frequency of discordant
424 topologies. Similarly, the relationship placing Polemoniaceae+Fouquieriaceae sister to
425 Lecythidaceae+ESC received a QD score of 0.83, indicating that among alternative topologies
426 (e.g. Primulaceae sister to the clade Lecythidaceae+ESC), there was no clear alternative to the
427 RBT recovered through quartet sampling for this dataset. The QI scores for all nodes defining
428 interfamilial relationships were above 0.9, indicating that in the vast majority of sampled quartets
429 there was a tree that was at least two log-likelihoods better than the alternatives. The QF scores
430 for all but one taxon were above 0.70 and the majority were above 0.85, which shows that rogue
431 individual taxa were not a major issue (Pease et al., 2018).
432 ***Conflict analyses***—Assessing ortholog tree concordance and conflict for the 387 ortholog set
433 mapped to the RBT showed that backbone nodes that differed between candidate topologies
434 were poorly supported with the majority of orthologs failing to achieve ≥70% bootstrap support
435 (Appendix S10). Among informative orthologs, the majority of trees conflict with any candidate
436 topology at these nodes and there is no dominant alternative to the RBT. There were 77 ortholog
437 trees with appropriate taxon sampling that placed Primulaceae sister to the rest of the non-

438 balsaminoid Ericales and 37 did so with at least 70% bootstrap support. The clade
439 Lecythidaceae+Primulaceae+ESC was recovered in 47 ortholog trees, and in 17 with at least
440 70% bootstrap support. Of the 33 ortholog trees that contained the clade
441 Primulaceae+Ebenaceae, nine did so with at least 70% bootstrap support.
442 ***Expanded ortholog sets***—Seven combinations of relationships along the backbone were
443 recovered in analyses of the expanded ortholog sets (Fig. 3). Full output for each of these
444 analyses are available on Dryad [TBD]. Among these, the ML tree estimated from the 2045-
445 ortholog PRANK-aligned, partitioned supermatrix placed Lecythidaceae and Ebenaceae in a
446 clade sister to Sapotaceae and the Core, with the other interfamilial relationships recapitulating
447 those of the RBT; the topology recovered by ASTRAL for these orthologs placed
448 Polemoniaceae+Fouquieriaceae sister to Sapotaceae and the Core, with the clade
449 Primulaceae+Ebenaceae sister to that and Lecythidaceae sister to a clade including all of those.
450 In regard to the placement of Ebenaceae for the 2045 PRANK-aligned set, the edge-based
451 Phyckle analysis showed that 432 of orthologs in this dataset with appropriate taxon sampling
452 supported Ebenaceae sister to Primulaceae, while 202 did so by at least two log-likelihood over
453 any alternative (Appendix S11). However, 416 orthologs supported Ebenaceae sister to
454 Sapotaceae (158 with ≥2LL), and 316 supported Ebenaceae sister to Lecythidaceae (140 with
455 ≥2LL). When the 2045 ortholog set was aligned with MAFFT to produce a supermatrix, the
456 resulting ML topology was such that the clade Primulaceae+Ebenaceae were sister to
457 Sapotaceae, with that clade sister to the Core and Polemoniaceae+Fouquieriaceae sister to all of
458 those. When the same orthologs were aligned with MAFFT, the ASTRAL topology recovered
459 placed Polemoniaceae+Fouqieriaceae sister to the Core, with the clade Ebenaceae+Primulaceae
460 sister to that, and Lecythidaceae sister to all of those.

461 The backbone topology resulting from the 4682 ortholog PRANK-aligned, partitioned
462 supermatix was the same as that recovered from the 2045 ortholog, partitioned, PRANK-aligned
463 ML tree. The ASTRAL topology for this dataset placed Lecythidaceae sister to the Core, the
464 clade Primulaceae+Ebenaceae sister to that, and Polemoniaceae+Fouquieriaceae sister to all of
465 those. When the 449, 661, or 1899 ortholog set were aligned with MAFFT to produce a
466 supermatrix, the resulting ML backbone topology was the same as that of the MAFFT-aligned
467 2045 and 4682 ortholog sets, except when the 449 ortholog supermatrix was used without
468 partitioning. When ortholog alignments were instead produced with PRANK, the ML backbone
469 topology recovered from the 449 and ortholog set was the same as that of the 2045 PRANK-
470 aligned ML tree (i.e Ebenaceae sister to Lecythidaceae and that clade sister to the Core). The
471 backbone of the ML trees produced from the 1899 and 661 ortholog sets recapitulated that of the
472 2045 ortholog ASTRAL tree, namely that Polemoniaceae+Fouquieriaceae was sister to
473 Sapotaceae and the Core, with the clade of Primulaceae+Ebenaceae sister to that and
474 Lecythidaceae sister to all of those.
475 ***Synthesis of uncertainty and determination of an overall consensus***—In the species trees
476 generated with the 387, 449, 661, 1899, 2045, and 4682 ortholog datasets, the placement of
477 several taxonomic families were resolved in conflict with one another. Many of these
478 relationships also conflict with the thoroughly investigated RBT, which was shown to be very
479 well-corroborated by the 387 ortholog set. Nine out of ten ML trees generated with MAFFT-
480 aligned supermatrices recovered Primulaceae and Ebenaceae sister to one another and forming a
481 clade with Sapotaceae, however this pattern was not recovered under any other circumstances
482 (Fig. 3). In all, Primulaceae was recovered as sister to Ebenaceae in 17 of the 33 species trees
483 generated in this study (51.5%), but these families were sister in only eight of the 24 trees

484 (33.3%) where they did not form a clade with Sapotaceae. Sapotaceae was recovered as sister to
485 the Core in 21 of the 33 species trees (63.6%). Thus, there is no majority consensus that
486 reconciles these conflicting relationships; Primulaceae is only sister to Ebenaceae in a majority
487 of trees if those that contain the clade Sapotaceae+Ebenaceae+Primulaceae are considered and
488 that topology is in direct conflict with Sapotaceae+Core, a relationship that is recovered in a
489 majority of trees. In addition, edge-wise support for the position of Ebenaceae was dataset
490 dependent and showed equivocal gene-wise support. The positions of Lecythidaceae or the clade
491 containing Polemoniaceae and Fouquieriaceae that is not agreed by a majority of species trees.
492 Given this, the relationships among Primulaceae, Polemoniaceae and Fouqieriaceae,
493 Lecythidaceae, Ebenaceae, and Sapotaceae are not resolved in the consensus topology. All other
494 interfamilial relationships were unanimously supported by all species trees except for the
495 relationships between the clade Symplocaceae, Diapensiaceae, and Styracaceae sister to
496 Theaceae, which was recovered in 29 of the 33 trees (87.9%).
497 ***Genome duplication inference***—Gene duplications mapped to the consensus topology shows
498 that the single largest number of duplications occurs on the branch leading the non-Balsaminoid
499 Ericales (Fig. 4). The other largest numbers of duplications appear along branches leading to
500 *Actinidia* (Actinidiaceae), *Camellia* (Theaceae), several members of Primulaceae, *Rhododendron*
501 (Ericaceae), *Impatiens* (Balsaminaceae), and Polemoniaceae (i.e. *Phlox* and *Saltugilia*). If gene
502 duplications are instead mapped to the species tree resulting from the 449 ortholog, MAFFT-
503 aligned supermatrix with partitioning, 3485 duplications occur along the branch leading to the
504 non-balsaminoid Ericales (Appendix S12). The K_s plots suggest support for several of the recent
505 lineage-specific duplications, indicated by K_s peaks near zero (Appendix S13). The K_s plots for
506 most, but not all, taxa also contain a peak that appears between 0.4 and 0.8, which is consistent

507 with a genome duplication shared by all Ericales. The multispecies K_s plots for some ericalean
508 taxa paired with a member of Cornales show an ortholog peak with a higher K_s value (i.e. farther
509 to the right) than the paralog peaks near 1.0, which would be indicative of two separate
510 duplication events occurring after the divergence of these species. However, other combinations
511 of ingroup and outgroup taxa result in an ortholog peak to the left of the paralogs peaks, which
512 would suggest that the implied duplication event occurred in a shared ancestor of the two orders
513 (Appendix S14). When comparing the locations of unambiguous ortholog peaks of all non-
514 balsaminoid ericalean taxa to *Impatiens balsamifera*, the location of the peak varied between
515 values of 1.01 to 1.57 in *Schima superba* and *Primula poissonii* respectively, indicating the
516 presence of rate heterogeneity in the accumulation of synonymous substitutions in Ericales
517 (Appendix S15).
518 ***Estimating substitutions supporting contentious clades*—**The estimated number of substitutions
519 informing clades containing at least two backbone families whose placement is contentious
520 tended to be much less than for relationships that garnered widespread support. These backbone
521 clades had a median value of 8.65 estimated substitutions informing them, compared to 30.08,
522 75.09, and 144.00 informing the monophyly of Polemoniaceae and Fouqieriaceae,
523 Lecythidaceae, and the non-Balsaminoid Ericales, respectively, in the MAFFT-aligned ortholog
524 trees (Fig. 4). In trees for the same orthologs with alignments generated instead with PRANK,
525 the corresponding median values were 8.96, 30.73, 74.16 and 145.98 respectively.
526 **DISCUSSION**
527 Our focal dataset, consisting of 387 orthologs, supports an evolutionary history of
528 Ericales that is largely consistent with previous work on the clade for many, but not all,
529 relationships (Appendix S9). The balsaminoid clade, which includes the families Balsaminaceae,

530 Tetrameristaceae (not sampled in this study), and Marcgraviaceae was confidently recovered as
531 sister to the rest of the order as has been shown previously (e.g. Geuten et al., 2004; Rose et al.,
532 2018; Gitzendanner et al., 2018). Similarly, the monogeneric family Fouquieriaceae is sister to
533 Polemoniaceae, the para-carnivorous Roridulaceae are sister to Actinidiaceae and the
534 circumscription of Primulaceae sensu lato is monophyletic (Rose et al., 2018). The majority of
535 analyses for this dataset recovered a sister relationship between Sapotaceae and Ebenaceae, with
536 that clade sister to the Core Ericales and Lecythidaceae sister to those. Notably, the topology
537 supported suggests that Primulaceae diverged earlier than has been recovered in most previous
538 phylogenetic studies and does not form a clade with Sapotaceae and Ebenaceae as has been
539 suggested by others, including Rose et al. (2018). The topology recovered by Gitzendanner et al.
540 (2018) using coding sequences from the chloroplast placed Primulaceae sister to Ebenaceae, with
541 those sister to the rest of the non-basaminoid Ericales, though that study did not include sampling
542 from Lecythidaceae. In addition to the traditional phylogenetic reconstruction methods, by
543 applying the available data on gene duplications as a metric of support and leveraging methods
544 that make use of additional phylogenetic information present in the supermatrix, we were able to
545 more holistically summarize the evidence present in a 387 ortholog dataset in an effort to resolve
546 the Ericales phylogeny (Fig. 1; Appendix S3-S10).
547 It has been shown repeatedly that large phylogenetic datasets have a tendency to resolve
548 relationships with strong support, even if the inferred topology is incorrect (Seo, 2008).
549 However, our results suggest extreme sensitivity to tree-building methods. For example, the
550 initial ML analysis resulted in an ML topology (MLT) with zero rapid bootstrap support for the
551 placement of Lecythidaceae (Fig. 1). The rapid bootstrap consensus instead unanimously
552 supported Lecythidaceae sister to a clade including Ebenaceae, Sapotaceae and the Core Ericales

553 (ESC), a hypothesis we refer to as the rapid bootstrap topology (RBT). Gene-wise investigation
554 of likelihood contribution confirmed that these two topologies have very similar likelihoods but
555 did not identify outlier genes that seemed to have an outsized effect on ML calculation
556 (Appendix S3). Instead, the cumulative likelihood influence of the 387 genes in the supermatrix
557 provides nearly equal support for the two topologies, while ASTRAL resulted in a third, and full
558 bootstrapping recovered an even more diverse set of topologies (Appendix S4). These results
559 suggest that there are several topologies with relatively close likelihood scores for this dataset
560 and despite the fact that the additional comparative analyses applied to the 387 ortholog dataset
561 supported a single alternative among those investigated (i.e. the RBT), the recovery of multiple
562 topologies by various tree-building and bootstrapping methods suggested that the criteria used to
563 generate and filter orthologs could have marked potential to influence the outcome of our efforts
564 to resolve relationships amongst the families of Ericales, possibly due to the evolutionary history
565 of the group and the amount of data available to inform these relationships.

566 In addition to sensitivities associated with tree building methods, variation in dataset
567 assembly parameters and methods can result in different topologies. We investigated additional
568 datasets constructed with a variety of methods and filtering parameters to shed further light on
569 the nature of the problem of resolving the ericalean phylogeny. While it is clear from
570 investigation of gene tree topologies in the 387 ortholog dataset that phylogenetic conflict is the
571 rule rather than the exception for those data, the expanded datasets show that this is true for a
572 variety of approaches to dataset construction and not simply an artifact of one approach. While
573 the monophyly of most major clades and the relationships discussed above were recovered
574 across these datasets, we also demonstrate that many combinations of contentious relationships

575 can be recovered depending on the methods used in dataset construction, alignment, and analysis
576 (Fig 2).
577 ***An unresolved consensus topology***—Based on the data available, we suggest that while the
578 relationships recovered in the Core Ericales and within most families are robust to
579 methodological alternatives, there is insufficient evidence to resolve several early-diverging
580 relationships along the ericalean backbone. We, therefore, suggest that the appropriate
581 representation is as a polytomy (Fig. 4). Whether this is biological or the result of data
582 limitations remains to be determined. A biological polytomy (i.e. hard polytomy) can be the
583 result of three or more populations diverging rapidly without significant time for the
584 accumulation of substitutions to distinguish lineage divergence. The major clades of Ericales
585 may or may not have diverged simultaneously; however, if divergence occurred rapidly enough
586 as to preclude the evolution of genomic synapomorphies, then a polytomy may be a reasonable
587 representation of such historical events rather than signifying a failure of methods or lack of data.
588 Most of the inferred orthologs contain little information useful for inferring relationships along
589 the backbone of the phylogeny; we investigated this explicitly by estimating the number of
590 nucleotide substitutions that inform these backbone relationships and find that branches that
591 would resolve the polytomy were based on a fraction of the number of substitutions that
592 informed better-supported clades (Fig. 5). The information that is present results in extensive
593 gene tree conflict, albeit mostly with low support (Appendix S11). Due to the timing of these
594 divergence events (80–100 million years ago; Magallón et al., 2015; Rose et al., 2018), coding
595 DNA sequences may be particularly well suited for inference of these deep relationships due to
596 their slow rate of evolution relative to noncoding regions, while allowing the inclusion of
597 potentially informative silent substitutions invisible in amino acid alignments. Our use of a tree-

598 based orthology inference method provides the best protection available against misidentified
599 orthology that would introduce systematic error into the analysis. Additionally, the stringent
600 filtering criteria applied during the initial 387 ortholog supermatix construction, requirement of
601 an outgroup taxon, and repeated trimming of long branches meant that any sequences that were
602 included in this dataset were required to be relatively slowly evolving sequences that were
603 robustly inferred to be orthologous (Yang et al., 2014).

604 The high levels of gene tree conflict and lack of a clear consensus among datasets for a
605 resolved topology is likely to have multiple causes. Among these is the fact that this series of
606 divergence events seems to have happened relatively rapidly over the course of around 10
607 million years (Rose et al., 2018). We also find evidence that a whole genome duplication is
608 likely to have occurred before or during this ancient radiation. If this is the case, differential gene
609 loss and retention during the process of diploidization is likely to complicate our ability to
610 resolve the order of lineage divergences. In addition, we cannot exclude the possibility of
611 hybridization and introgression between these ancient lineages at some point in the past, as
612 hybridization has been documented between plants that have been diverged for tens of millions
613 of years (Arias et al., 2014; Rothfels et al., 2015). It is even possible that some of the lineages
614 involved in such introgression have gone extinct in the intervening 100 million years.
615 Hybridization with such now-extinct "ghost lineages" could therefore represent a potentially
616 insurmountable obstacle to fully understanding the events that lead to the diversity of forms we
617 now find in Ericales. We chose not to test explicitly for evidence of hybridization here, due to
618 the seeming equivocal phylogenetic signal present in most gene trees for these contentious
619 relationships. We feel that to interpret the generally weak signal present in most conflicting gene
620 trees as anything other than a lack of reliable information, runs a high risk of overinterpreting

621 these data since network analyses generally treat tree topology as a state known without error.
622 However, future studies with more complete taxon sampling could find such an approach to be
623 appropriate for explaining the high levels of conflict among orthologs, but should carefully
624 consider alternative explanations for discordance.
625 ***Genome duplications***—Gene duplication analyses showed evidence for several whole genome
626 duplications in Ericales, including at least two, those in *Camellia* and *Actinidia*, that have been
627 verified using sequenced genomes (Huang et al., 2013; Xia et al., 2017; Shi et al., 2010).
628 Incomplete taxon sampling precludes resolving the precise location of these duplication events
629 within their respective families, but the 379 duplications shared by *Schima* and *Camellia*
630 suggests the genome duplication within Theaceae occurred before the divergence of these two
631 genera, though additional taxa from the family should be investigated in the future. Similarly,
632 our results show that all species of *Actinidia* included in our study share at least one whole
633 genome duplication. The corroboration of our results with those of previous work provides
634 support that the transcriptomic methods we employ here provide a viable means of assessing the
635 location of recent genome duplications in Ericales (Huang et al., 2013; Xia et al., 2017; Shi et al.,
636 2010).
637 The pattern of gene duplication observed along the backbone (i.e. hundreds to thousands
638 of inferred gene duplications at several contentious nodes) is difficult to interpret in any of our
639 resolved phylogenies for several reasons (Fig. 2; Appendix S12). The observed pattern could be
640 the result of biological processes, incomplete transcript sampling or assembly, systematic error in
641 phylogenetic reconstruction, or a combination of these factors (Sankoff and Zheng, 2018). If
642 these results were strictly due to biological processes, then the data would suggest that there were
643 repeated, nested, large-scale gene or whole genome duplications between successive divergence

644 events. It is also possible that there was a single genome duplication event followed by
645 cladogenesis and subsequent differential gene loss in descendent lineages. This would result in
646 the appearance of additional nested duplications because of the differential gene loss. It is
647 important to note that either of these biological processes produces evidence of shared ancestry
648 that is potentially useful in inferring evolutionary relationships, as we investigate here. In
649 addition to biological processes, failure to sequence or assemble transcripts in a given paralog, as
650 well as phylogenetic uncertainty in homolog trees, can result in a failure to correctly identify the
651 existence and/or location of duplications.

652 Our results strongly suggest that uncertainty should be considered when inferring
653 duplications with tree-based methods and/or from K_s plots (Zwaenepoel and Van de Peer, 2019).
654 In the case of the former, gene duplications are mapped based on a species tree and therefore the
655 location to which a given duplication appears can depend on which topologies are considered
656 (Fig. 2 A-H; Fig. 3; Appendix S12), while the latter may be affected by heterogeneity in
657 evolutionary rate with faster-evolving taxa accumulating synonymous substitutions at faster rates
658 than more slowly evolving lineages (Appendix S15; Smith and Donoghue, 2008). Many of our
659 transcriptomes appear to share two peaks, one around 2.0 that is well-known to be the result of a
660 genome duplication shared by all angiosperms (Jiao et al., 2011) and another that in most taxa
661 occurs at K_s values between 0.4 and 0.8 (Appendix S13-S14), though the location of the latter
662 varies quite substantially, possibly due to the rate heterogeneity in K_s accumulation we find
663 across ericalean taxa (Appendix S15). Several transcriptomes of outgroup taxa in Cornales also
664 contain a peak near these values, suggesting that this group also experienced a whole genome
665 duplication around the same time as Ericales.

666 Recent nuclear analyses (e.g. Zeng et al., 2017) have provided evidence that Ericales is
667 sister to Cornales, rather than the two as successive sister-groups to the other asterids, as several
668 earlier studies had suggested (Bremer et al., 2002, Wikström et al., 2015). Our use of Cornales as
669 outgroups prohibits us from addressing this topological question directly, but if the two clades
670 are sister, evidence of similar K_s peaks in both clades would be compatible with a shared
671 duplication event predating their divergence (Jiao et al., 2012; Wang et al., 2016, Landis et al.,
672 2018). The multispecies K_s plots do not appear to definitively resolve how many duplication
673 events took place, since some results suggest a whole genome duplication occurring before the
674 MRCA of Cornales and Ericales, while results from combinations of taxa directly conflict with
675 such an interpretation (Appendix S14). We show that the location of ortholog peaks can differ
676 quite dramatically among species pairs that share the same MRCA (Appendix S14-S15), which
677 has important implications for the interpretation of multispecies K_s plots. Given this and the
678 topological uncertainty that remains, we interpret our results to mean that at least one genome
679 duplication occurred around the time (i.e. before, during, or after) the divergence of Ericales
680 from other Asterids. Our sampling prevents us from explicitly testing the number of shared
681 duplications between Ericales and Cornales, but the tree-based methods we employ here detect
682 duplications in an ingroup (i.e. Ericales) relative to an outgroup (i.e. Cornales, *Helianthus*,
683 *Solanum*, *Arabadopsis*, and *Beta*). The fact that we observed thousands of gene duplications
684 along the backbone of Ericales not present in any outgroup taxon suggests that a whole genome
685 duplication occurred after Ericales diverged from other Asterids, since these duplicated genes are
686 not present in the seven cornalean or other outgroup transcriptomes based on reference genomes.
687 Additional data and expanded taxon sampling will likely shed light on whether a whole genome
688 duplication occurred before or after the MCRA of Ericales, but the evidence presented here

689 suggests that choice of taxa and methodology could have important effects on the inferred
690 locations of such events given that there are several lineage-specific whole genome duplications
691 towards the tips of the ericalean tree (Fig. 4; Appendix S13-S15), and that the backbone of the
692 phylogeny is highly sensitive to methodology (Fig. 3).

693 **CONCLUSIONS**

694 The first transcriptomic dataset broadly spanning Ericales and constructed from publicly
695 available data resolves many of the relationships within the clade and supports several
696 relationships that have been proposed previously. Our results confirm genome duplications in
697 Actinidiaceae and Theaceae and support at least one whole genome duplication close to the
698 origin of Ericales. We find evidence to suggest additional duplication events in Balsaminaceae,
699 Ericaceae, Polemoniaceae, and Primulaceae. As genomic sampling improves, we may discover
700 additional duplications and refine placement of others. While trees were largely concordant
701 within taxonomic families, our results concerning the topological resolution of the deep
702 divergences in Ericales are less decisive. We demonstrate that with the available data, there is
703 not enough information to strongly support any resolution. While these relationships have been
704 resolved in previous studies, our analyses strongly suggest that uncertainty should be thoroughly
705 investigated in phylotranscriptomic datasets, as strong support can be given by different methods
706 for conflicting topologies. Even in a dataset containing hundreds of genes and hundreds of
707 thousands of characters, the criteria used in dataset construction altered the inferred topology,
708 even without the presence of outlier genes. The high prevalence of conflict among datasets and
709 the lack of clear consensus in regards to the relationships among several major ericalean clades
710 led us to conclude that a single, fully resolved tree is not supported by these transcriptome data,
711 though we acknowledge that future improvements in sampling might justify their resolution. Our

results suggest that future work should employ a range of methodologies and support metrics so that topological uncertainty can be fully explored and reported.

Data Accessibility Statement

Data and custom software used in this study are available at Dryad [TBD].

Acknowledgements

The authors thank Christopher Dick, Greg Stull, Hannah Marx, Ning Wang, Caroline Edwards, and members of the Smith lab for their comments on the manuscript. Support came from the National Science Foundation; D.A.L and J.F.W. were supported by NSF DEB 1207915, O.M.V. was supported by NSF FESD 1338694, and S.A.S. was supported by NSF DBI 1458466.

Author Contributions

The study was conceived by D.A.L. and S.A.S. Dataset construction was led by D.A.L. with contributions by O.M.V. and input from J.F.W. Analyses were led by D.A.L. and J.F.W with contributions and input by S.A.S. D.A.L. wrote the manuscript with contributions and input from O.M.V, J.F.W., and S.A.S. All authors read and approved the final draft of the manuscript.

732 **LITERATURE CITED**

733 Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment
734 search tool. *Journal of molecular biology* 215: 403-410.

735 Anderberg, A. A., C. Rydin, and M. Källersjö. 2002. Phylogenetic relationships in the order
736 Ericales sl: analyses of molecular data from five genes from the plastid and mitochondrial
737 genomes. *American Journal of Botany* 89: 677-687.

738 Arias, T., M. A. Beilstein, M. Tang, M. R. McKain, and J. C. Pires. 2014. Diversification times
739 among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of
740 divergence. *American Journal of Botany* 101: 86-91.

741 Bremer, B., K. Bremer, N. Heidari, P. Erixon, R. G. Olmstead, A. A. Anderberg, M. Källersjö,
742 and E. Barkhordarian. 2002. Phylogenetics of asterids based on 3 coding and 3 noncoding
743 chloroplast DNA markers and the utility of noncoding DNA at higher taxonomic levels.
744 *Molecular Phylogenetics and Evolution* 24: 274–301.

745 Brown, J. M., and R. C. Thomson. 2016. Bayes factors unmask highly variable information
746 content, bias, and extreme influence in phylogenomic analyses. Systematic Biology 66:
747 517-530.

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

750 Brown, J. W., N. Wang, and S. A. Smith. 2017. The development of scientific consensus:
751 Analyzing conflict and concordance among avian phylogenies. Molecular Phylogenetics
752 and Evolution 116: 69-77.

753 Chase, M. W., M. Christenhusz, M. Fay, J. Byng, W. S. Judd, D. Soltis, D. Mabberley, et al.
754 2016. An update of the Angiosperm Phylogeny Group classification for the orders and

755 families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1-
756 20.
757 Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall,
758 et al. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the
759 plastid gene rbcL. *Annals of the Missouri Botanical Garden*: 528-580.
760 Eisen, J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by
761 evolutionary analysis. *Genome research* 8: 163-167.
762 Fu, L., B. Niu, Z. Zhu, S. Wu, and W. Li. 2012. CD-HIT: accelerated for clustering the next
763 generation sequencing data. *Bioinformatics* 28: 3150–3152.
764 Gabaldón, T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome*
765 *biology* 9: 235.
766 Geuten, K., E. Smets, P. Schols, Y.-M. Yuan, S. Janssens, P. Küpfer, and N. Pyck. 2004.
767 Conflicting phylogenies of balsaminoid families and the polytomy in Ericales: combining
768 data in a Bayesian framework. *Molecular Phylogenetics and Evolution* 31: 711-729.
769 Gitzendanner, M. A., P. S. Soltis, G. K. S. Wong, B. R. Ruhfel, and D. E. Soltis. 2018. Plastid
770 phylogenomic analysis of green plants: a billion years of evolutionary history. *American*
771 *Journal of Botany* 105: 291-301.
772 Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et
773 al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference
774 genome. *Nature biotechnology* 29: 644.
775 Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger,
776 et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity
777 platform for reference generation and analysis. *Nature Protocols* 8: 1494–1512.

778 He, Y., C. Kueffer, P. Shi, X. Zhang, M. Du, W. Yan, and W. Sun. (2014) Variation of biomass
779 and morphology of the cushion plant Androsace tapete along an elevational gradient in
780 the Tibetan Plateau. *Plant species biology* 29: e64-e71.
781 Hedwall, P. O., J. Brunet, A. Nordin, and J. Bergh. 2013. Changes in the abundance of keystone
782 forest floor species in response to changes of forest structure. *Journal of vegetation*
783 *science* 24: 296-306.
784 Jiao, Y., J. Leebens-Mack, S. Ayyampalayam, J. E. Bowers, M. R. McKain, J. McNeal, M. Rolf,
785 et al. 2012. A genome triplication associated with early diversification of the core
786 eudicots. *Genome biology* 13: R3.
787 Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P.
788 Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:
789 97.
790 Katoh, K., K. Misawa, K. i. Kuma, and T. Miyata. 2002. MAFFT: a novel method for rapid
791 multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:
792 3059-3066.
793 Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7:
794 improvements in performance and usability. *Molecular biology and evolution* 30: 772-
795 780.
796 Kozlov, A., D. Darriba, T. Flouri, B. Morel, and A. Stamatakis. 2019. RAxML-NG: A fast,
797 scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *bioRxiv*:
798 447110.

799 Landis, J. B., D. E. Soltis, Z. Li, H. E. Marx, M. S. Barker, D. C. Tank, and P. S. Soltis. 2018.
800 Impact of whole-genome duplication events on diversification rates in angiosperms.
801 *American Journal of Botany* 105: 348-363.
802 Löytynoja, A. 2014. Phylogeny-aware alignment with PRANK, Multiple sequence alignment
803 methods, 155-170. Springer.
804 Magallón, S., S. Gómez-Acevedo, L. L. Sánchez-Reyes, and T. Hernández-Hernández. 2015. A
805 metacalibrated time-tree documents the early rise of flowering plant phylogenetic
806 diversity. *New Phytologist* 207: 437-453.
807 Matasci, N., L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, et al.
808 2014. Data access for the 1,000 Plants (1KP) project. *GigaScience* 3: 17.
809 Memiaghe, H. R., J. A. Lutz, L. Korte, A. Alonso, and D. Kenfack (2016). Ecological
810 importance of small-diameter trees to the structure, diversity, and biomass of a tropical
811 evergreen forest at Rabi, Gabon. *PLoS One* 11: e0154988.
812 Kodama, Y., M. Shumway, and R. Leinonen. 2011. The sequence read archive. *Nucleic Acids*
813 *Research* 39: D19–D21.
814 Moquet, L., M. Vanderplanck, R. Moerman, M. Quinet, N. Roger, D. Michez, and A. L.
815 Jacquemart. 2017. Bumblebees depend on ericaceous species to survive in temperate
816 heathlands. *Insect Conservation and Diversity* 10: 78-93.
817 Pease, J. B., J. W. Brown, J. F. Walker, C. E. Hinchliff, and S. A. Smith. 2018. Quartet Sampling
818 distinguishes lack of support from conflicting support in the green plant tree of life.
819 *American Journal of Botany* 105: 385-403.
820 Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 -Approximately maximum-
821 likelihood trees for large alignments. *PLoS One* 5: e9490.

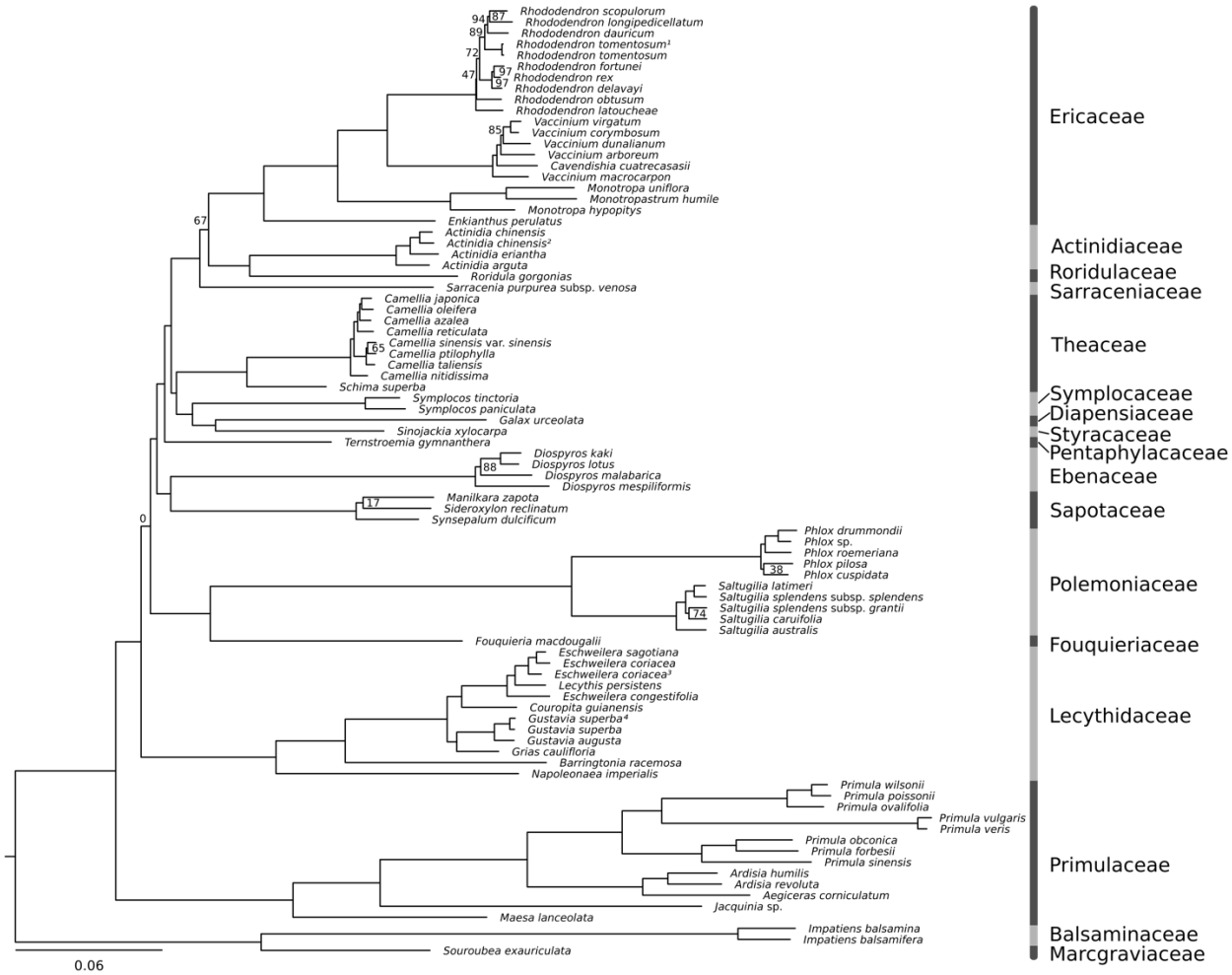
822 Reddy, S., R. T. Kimball, A. Pandey, P. A. Hosner, M. J. Braun, S. J. Hackett, K.-L. Han, et al.
823 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the
824 avian tree of life more than taxon sampling. *Systematic Biology* 66: 857-879.
825 Rose, J. P., T. J. Kleist, S. D. Löfstrand, B. T. Drew, J. Schönenberger, and K. J. Sytsma. 2018.
826 Phylogeny, historical biogeography, and diversification of angiosperm order Ericales
827 suggest ancient Neotropical and East Asian connections. *Molecular Phylogenetics and*
828 *Evolution* 122: 59-79.
829 Rothfels, C. J., A. K. Johnson, P. H. Hovenkamp, D. L. Swofford, H. C. Roskam, C. R. Fraser-
830 Jenkins, M. D. Windham, and K. M. Pryer. 2015. Natural hybridization between genera
831 that diverged from each other approximately 60 million years ago. *The American*
832 *Naturalist* 185: 433-442.
833 Sankoff, D., and C. Zheng. 2018. Whole Genome Duplication in Plants: Implications for
834 Evolutionary Analysis, Comparative Genomics, 291-315. Springer.
835 Shen, X.-X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic
836 studies can be driven by a handful of genes. *Nature ecology & evolution* 1: 0126.
837 Schönenberger, J., A. A. Anderberg, and K. J. Sytsma. 2005. Molecular phylogenetics and
838 patterns of floral evolution in the Ericales. *International Journal of Plant Sciences* 166.2:
839 265-288.
840 Shi, T., H. Huang, and M. S. Barker. 2010. Ancient genome duplications during the evolution of
841 kiwifruit (Actinidia) and related Ericales. *Annals of Botany* 106: 497-504.
842 Simon, S., A. Narechania, R. DeSalle, and H. Hadrys. 2012. Insect phylogenomics: exploring the
843 source of incongruence using new transcriptomic data. *Genome biology and evolution* 4:
844 1295-1309.

845 Smith, S.A., J. F. Walker, J. W. Brown, and N. Walker-Hale, 2018. Nested phylogenetic
846 conflicts, combinability, and deep phylogenomics in plants. *bioRxiv*, 371930.
847 Smith, S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history
848 in flowering plants. *Science* 322: 86-89.
849 Smith, S. A., M. J. Moore, J. W. Brown, and Y. Yang. 2015. Analysis of phylogenomic datasets
850 reveals conflict, concordance, and gene duplications with examples from animals and
851 plants. *BMC Evolutionary Biology* 15: 150.
852 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
853 large phylogenies. Bioinformatics 30: 1312-1313.
854 Stevens, P. F. 2001 onward. Angiosperm phylogeny website, version 14, July 2017 [more or less
855 continuously updated]. Website http://www.mobot.org/MOBOT/research/APweb/
856 [accessed 11 July 2019]
857 Ter Steege, H., N. C. Pitman, O. L. Phillips, J. Chave, D. Sabatier, A. Duque, J.-F. Molino, et al.
858 2006. Continental-scale patterns of canopy tree composition and function across
859 Amazonia. *Nature* 443: 444.
860 Vargas, O. M., E. M. Ortiz, and B. B. Simpson. 2017. Conflicting phylogenomic signals reveal a
861 pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae:
862 Astereae: Diplostephium). *New Phytologist* 214: 1736-1750.
863 Walker, J. F., J. W. Brown, and S. A. Smith. 2018. Analyzing contentious relationships and
864 outlier genes in phylogenomics. *Systematic Biology* 67: 916-924.
865 Walker, J. F., N. Walker-Hale, O. M. Vargas, and D. A. Larson, G. W. Stull. 2019.
866 Characterizing gene tree conflict in plastome-inferred phylogenies. *bioRxiv*: 512079.

867 Walker, J. F., Y. Yang, T. Feng, A. Timoneda, J. Mikenas, V. Hutchison, C. Edwards, et al.
868 2018. From cacti to carnivores: Improved phylotranscriptomic sampling and hierarchical
869 homology inference provide further insight into the evolution of Caryophyllales.
870 *American Journal of Botany* 105: 446-462.
871 Walker, J. F., Y. Yang, M. J. Moore, J. Mikenas, A. Timoneda, S. F. Brockington, and S. A.
872 Smith. 2017. Widespread paleopolyploidy, gene tree conflict, and recalcitrant
873 relationships among the carnivorous Caryophyllales. *American Journal of Botany* 104:
874 858-867.
875 Wang, N., Y. Yang, M. J. Moore, S. F. Brockington, J. F. Walker, J. W. Brown, B. Liang, et al.
876 2018. Evolution of Portulacineae marked by gene tree conflict and gene family expansion
877 associated with adaptation to harsh environments. *Molecular biology and evolution* 36:
878 112-126.
879 Wang, Y., S. P. Ficklin, X. Wang, F. A. Feltus, and A. H. Paterson. 2016. Large-scale gene
880 relocations following an ancient genome triplication associated with the diversification of
881 core eudicots. *PloS one* 11: e0155637.
882 Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam,
883 et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land
884 plants. *Proceedings of the National Academy of Sciences* 111: E4859-E4868.
885 Wikström, N., K. Kainulainen, S. G. Razafimandimbison, J. E. E. Smedmark, and B. Bremer.
886 2015. A revised time tree of the asterids: Establishing a temporal framework for
887 evolutionary studies of the coffee family (Rubiaceae). *PLoS One* 10: e0126690.

888 Xia, E.H., H.B. Zhang, J. Sheng, K. Li, Q.J. Zhang, C. Kim, Y. Zhang, et al. 2017. The tea tree
889 genome provides insights into tea flavor and independent evolution of caffeine
890 biosynthesis. *Molecular plant*, *10*: 866-877.
891 Yang, Y., M. J. Moore, S. F. Brockington, J. Mikenas, J. Olivieri, J. F. Walker, and S. A. Smith.
892 2018. Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy
893 events in Caryophyllales, including two allopolyploidy events. New Phytologist 217:
894 855-870.
895 Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K.-S. Wong, E. J. Carpenter, Y.
896 Zhang, et al. 2015. Dissecting molecular evolution in the highly diverse plant clade
897 Caryophyllales using transcriptome sequencing. Molecular biology and evolution 32:
898 2001-2014.
899 Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using
900 transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy
901 for phylogenomics. *Molecular biology and evolution* 31: 3081-3092.
902 Zeng, L., N. Zhang, Q. Zhang, P. K. Endress, J. Huang, and H. Ma (2017). Resolution of deep
903 eudicot phylogeny and their temporal diversification using nuclear genes from
904 transcriptomic and genomic datasets. *New Phytol.* 214, 1338–1354.
905 Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species
906 tree reconstruction from partially resolved gene trees. *BMC bioinformatics* 19: 153.
907 Zwaenepoel A., Y. Van de Peer. 2019. Inference of Ancient Whole-Genome Duplications and
908 the Evolution of Gene Duplication and Loss Rates. *Mol Biol Evol*. 36: 1384–1404.
909
910

911 **FIGURES**



912

913 **Figure 1.** Maximum likelihood topology recovered for a 387 ortholog dataset using RAxML.

914 Nodes receiving less than 100% rapid BS support are labeled. Branch lengths are substitutions

915 per base pair. The node that determined the placement of Lecythidaceae received zero support

916 (i.e. the ML topology was never recovered by a rapid bootstrap replicate).

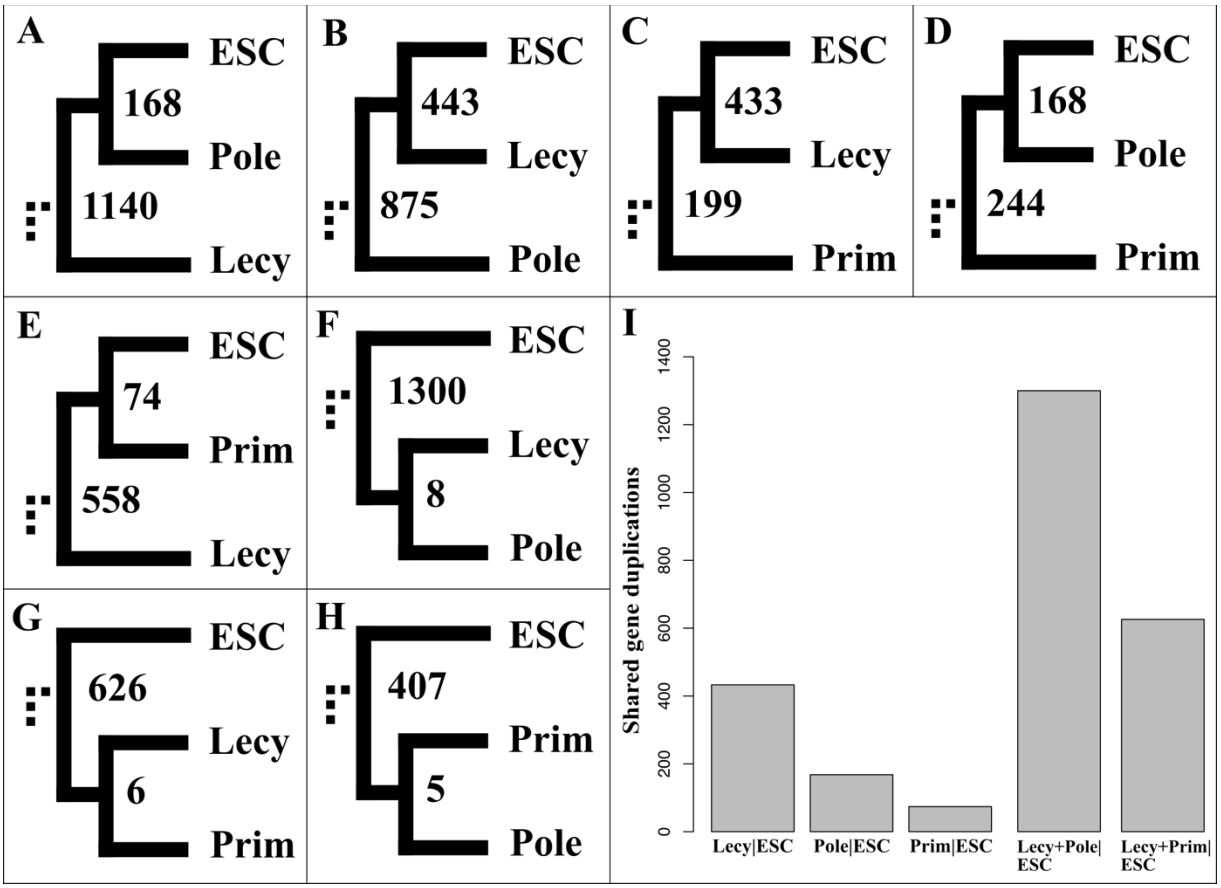


Figure 2. (A-H) Gene duplications obtained by mapping duplicated clades with at least 50% SH-aRLT support in homolog trees to several topologies recovered with a 387 ortholog dataset including the maximum likelihood topology (A) the rapid bootstrap topology (B) the ASTRAL topology (C) and several hypothetical topologies that demonstrate evidence for shared duplications in clades not recovered with species tree methods (D-H). Names of clades are abbreviated to four letters, ESC represents the clade Ebenaceae+Sapotaceae+Core Ericales, and "Pole" represents the clade Polemoniaceae+Fouquieriaceae in all cases. I) Bar chart showing the number of uniquely shared gene duplications between clades that can be considered a metric of support for distinguishing among conflicting topological relationships.

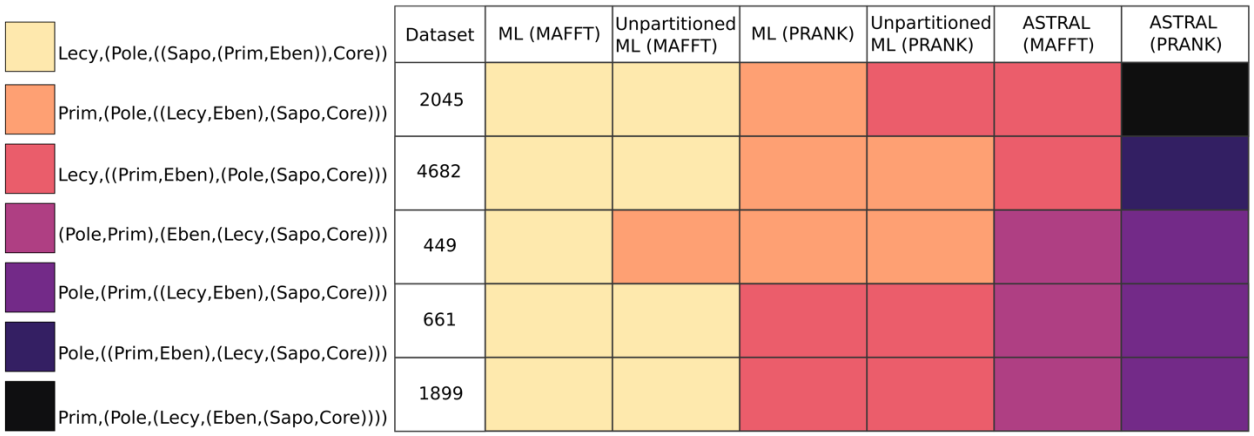
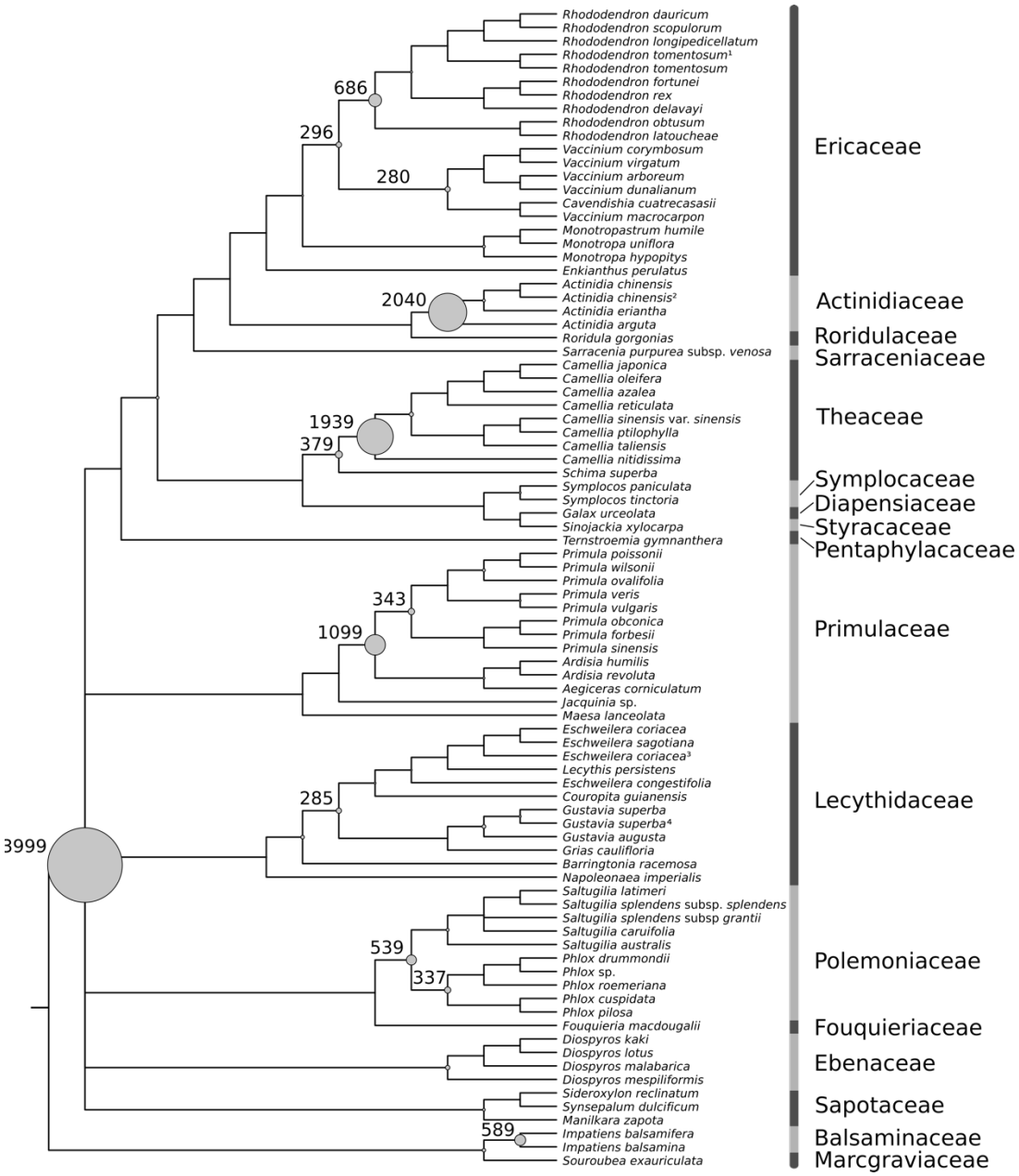


Figure 3. Topologies recovered from several combinations of ortholog datasets and species tree methods representing alternative approaches to those used for a 387 ortholog dataset. Each color corresponds to a unique backbone topology recovered in these analyses. Names of clades are abbreviated to four letters and "Pole" represents the clade Polemoniaceae+Fouquieriaceae in all cases.



933

934 **Figure 4.** Gene duplications mapped to a cladogram of the consensus topology. The diameter of
935 circles corresponds to the number of inferred duplications, and cases with at least 250 are
936 labelled along branches. The single largest number of duplications occurs on the branch leading
937 the non-Balsaminoid Ericales. Verified genome duplications in the Theaceae and the
938 Actinidiaceae appear as the second and third largest numbers of gene duplications respectively.

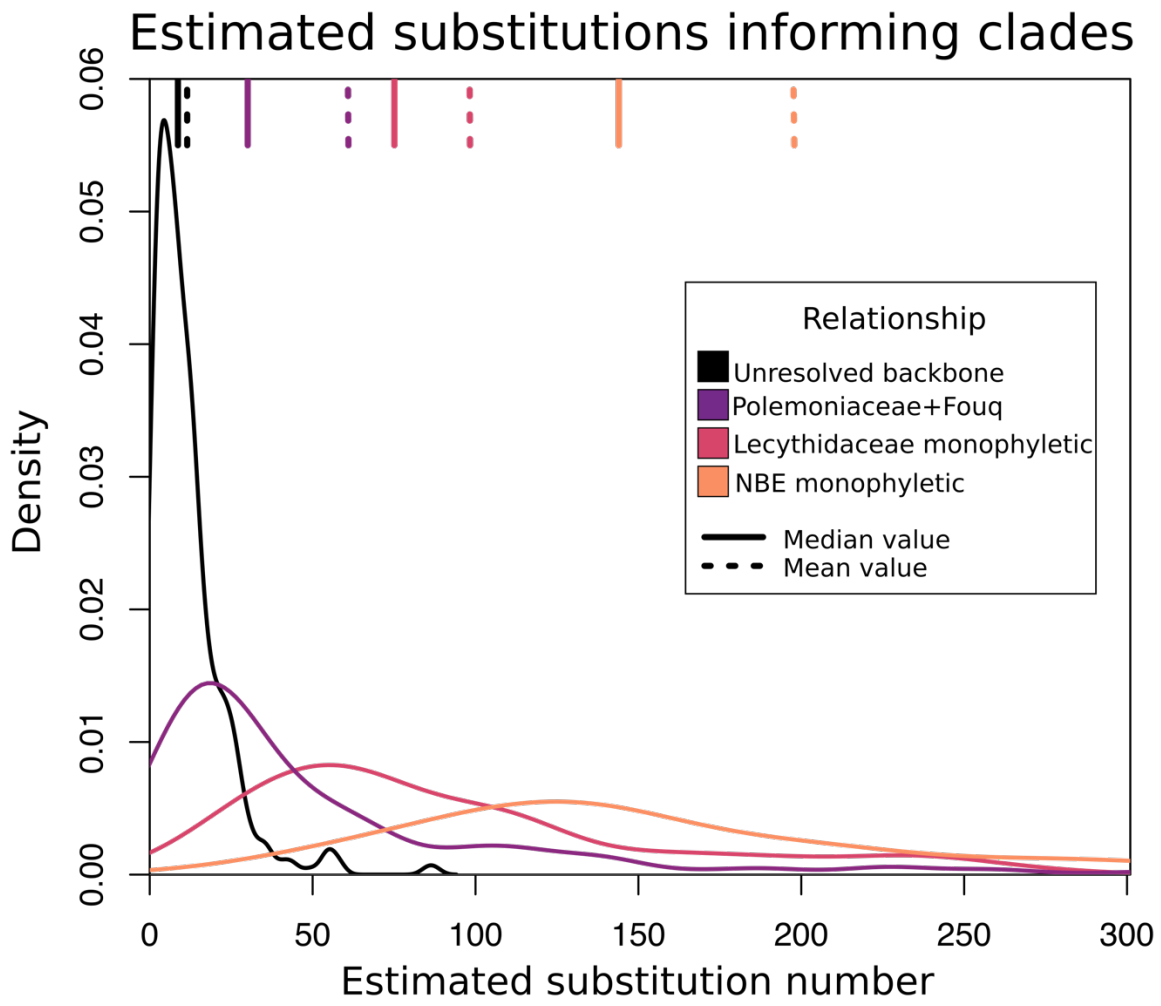


Figure 5. Estimated number of substitutions supporting various clades using rootable ortholog trees from the 449 ortholog MAFFT-aligned dataset. The number of substitutions supporting a clade was estimated by multiplying the relevant branch length by the length of the corresponding alignment for each ortholog tree in which that clade is present. Substitution support for clades that would resolve the polytomy in the consensus topology had a median value of 8.65 estimated substitutions informing them, compared to 30.08, 75.09, and 144.00 informing the monophyly of Polemoniaceae and Fouqieriaceae, Lecythidaceae, and the non-Balsaminoid Ericales respectively. In trees for the same orthologs aligned with PRANK, the corresponding values were 8.96, 30.73, 74.16 and 145.98.

949 **SUPPLEMENTAL MATERIALS**

950 **Supplemental Appendix 1.** The origins of transcriptome assemblies, reference genomes, and
951 raw reads used in homolog clustering. Citations associated with raw reads available on the
952 National Center for Biotechnology Information Sequence Read Archive are included where such
953 information was provided in the accession record or could be confidently identified through a
954 Google Scholar query of the relevant accession information.

955

956 **Supplemental Appendix 2.** Phylogram inferred using maximum likelihood on a supermatrix
957 consisting of the genes *rpoC2*, *rbcL*, *nhdF*, and *matK*. The topology of the tree was used to
958 assign taxa into eight groups for hierarchical homolog clustering, such that each clustering group
959 was monophyletic and not prohibitively large.

960

961 **Supplemental Appendix 3.** Results of a two topology test comparing gene-wise support in the
962 387 ortholog dataset for two alternative placements of Lecythidaceae recovered by the maximum
963 likelihood (ML) search. Delta gene-wise log-likelihood represents the extent to which an
964 ortholog supports the ML topology over that recovered unanimously in rapid bootstraps. The
965 dashed line represents the cumulative difference in log-likelihood between the two competing
966 topologies, such that removing any of the 27 orthologs with a score more positive than that
967 would cause the ML topology to change, even in the absence of obvious outlier genes.

968

969 **Supplemental Appendix 4.** Results of unguided, full bootstrapping with the 387 ortholog
970 supermatrix in RAxML. The number of bootstrap replicates in which various contentious
971 backbone relationships were recovered are reported.

Supplemental Appendix 5. The maximum quartet support species tree topology for the 387 ortholog set generated with ASTRAL. Node support values are ASTRAL local posterior probabilities and nodes receiving support less than 1.0 are labeled. Branch lengths are in coalescent units.

Supplemental Appendix 6. Results from the gene-wise comparisons of likelihood contributions for the three candidate topologies for the 387 ortholog dataset. A) The absolute log-likelihood for each ortholog, sorted by which topology they best support and organized in descending order of likelihood. B) The cumulative difference in log-likelihood relative to the ML topology as orthologs are added across the supermatrix. Delta gene-wise log-likelihood (ΔGWLL) represents the extent to which the topology has a worse score than the ML topology; values below zero indicate that a topology has a better likelihood than the ML topology. C) Schematic of the three candidate topologies in question.

Supplemental Appendix 7. Log-likelihood penalties incurred by constraining key edges consistent with each of the three candidate topologies for a 387 ortholog dataset using EdgeTest. Direct comparisons of scores between conflicting clades can be used as a metric of support.

Supplemental Appendix 8. The Phyckle results regarding the placement of Ebenaceae for the 387 ortholog dataset. Of the relationships examined, Ebenaceae+Sapotaceae was supported by the highest number of orthologs whether or not two log-likelihood support difference was required.

995

996 **Supplemental Appendix 9.** Cladogram showing results from quartet sampling on the rapid
997 bootstrap topology (RBT) from a 387 ortholog dataset. Branch labels show quartet concordance
998 (QC), quartet differential (QD), and quartet informativeness (QI) respectively for each
999 relationship. Quartet fidelity (QF) for each taxon is shown in parentheses after the relevant taxon
1000 label.

1001 **Supplemental Appendix 10.** A) Ortholog tree concordance and conflict for the 387 ortholog set
1002 mapped to the rapid bootstrap topology. B) The same analysis except requiring 70% ortholog
1003 tree bootstrap support for an ortholog to be considered informative. Blue indicates the proportion
1004 of informative orthologs that are concordant with the topology, green indicates the proportion of
1005 informative orthologs that support the single most common conflicting topology, red indicates all
1006 other informative ortholog conflict and grey indicates orthologs that are uninformative, either
1007 due to support requirements or lack of appropriate taxon sampling.

1008

1009 **Supplemental Appendix 11.** The Phyckle results regarding the placement of Ebenaceae for the
1010 2045 ortholog dataset. Ebenaceae+Primulaceae was supported by the highest number of
1011 orthologs whether or not two log-likelihood support difference was required.

1012

1013 **Supplemental Appendix 12.** Gene duplications mapped to a cladogram of the 449 ortholog
1014 MAFFT-aligned, partitioned, supermatrix ML tree. Number of inferred gene duplications in
1015 shown along branches. The diameter of circles at nodes are proportional to the number of
1016 duplications.

1017

1018 **Supplemental Appendix 13.** Single species K_s plots for pairs of paralogs within each taxa as a
1019 density plot. Density peaks indicate evidence for a large proportion of genes having been
1020 duplicated at approximately the same time, as would occur during a whole genome duplication.
1021 For each taxa, the x-axis ranges from 0.0 to 3.0 with each tick representing 0.5 synonymous
1022 substitutions between paralogs. The y-axis is scaled to the maximum density value for each
1023 transcriptome and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density
1024 respectively. Input transcriptomes have been filtered to remove short sequences and sample-
1025 specific duplications since these can represent transcript splice-site variants or errors in
1026 assembly.
1027
1028 **Supplemental Appendix 14.** Multispecies K_s plots representing a broad range of taxonomic
1029 pairings. Single species K_s density plots (i.e. pairs of paralogs) within each taxa are plotted on the
1030 same axis as a K_s density plot representing pairs of orthologs between the two taxa. The density
1031 peak for the orthologs corresponds to the time of divergence between the two taxa, since
1032 orthologs in both taxa begin accumulating synonymous substitutions after speciation. If
1033 evolutionary rate in the two taxa are equal and the accumulation of synonymous substitutions is
1034 clock-like, then the timing the duplication events relative to divergence of the two taxa can be
1035 compared, with older events occurring farther to the right. The x-axis ranges from 0.0 to 3.0 with
1036 each tick representing 0.5 synonymous substitutions. The y-axis is scaled to the maximum
1037 density value of any of the three density distributions (therefore the magnitudes of all peaks are
1038 relative) and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density respectively.
1039 Input transcriptomes have been filtered to remove short sequences and sample-specific
1040 duplications since these can represent transcript splice-site variants or errors in assembly.

1041

1042 **Supplemental Appendix 15.** Multispecies K_s plots representing each non-balsaminoid ericalean
1043 taxa paired with *Impatiens balsamifera* for which there is an unambiguous ortholog peak. Single
1044 species K_s density plots (i.e. pairs of paralogs) within each taxa are plotted on the same axis as a
1045 K_s density plot representing pairs of orthologs between the two taxa. The density peak for the
1046 orthologs corresponds to the time of divergence between the two taxa, since orthologs in both
1047 taxa begin accumulating synonymous substitutions after speciation. The dashed line represents
1048 the point along the x-axis where the ortholog peak achieves its maximum value and would be
1049 expected to occur at approximately the same point in all pairings under clock-like accumulation
1050 of synonymous substitutions since all pairs share the same MRCA. The x-axis ranges from 0.0 to
1051 3.0 with each tick representing 0.5 synonymous substitutions. The y-axis is scaled to the
1052 maximum density value of any of the three density distributions (therefore the magnitudes of all
1053 peaks are relative) and ticks correspond to 0.25, 0.5, 0.75, and 1.0 of the maximum density
1054 respectively. Input transcriptomes have been filtered to remove short sequences and sample-
1055 specific duplications since these can represent transcript splice-site variants or errors in
1056 assembly.