# Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in *Alchemilla* s.l. (Rosaceae).

Diego F. Morales-Briones[1,2,*], Berit Gehrke[3], Chien-Hsun Huang[4], Aaron Liston[5], Hong Ma[6], Hannah E. Marx[7], David C. Tank[2], Ya Yang[1]

[1]Department of Plant and Microbial Biology, University of Minnesota-Twin Cities, 1445 Gortner Avenue, St. Paul, MN 55108, USA

[2]Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, 875 Perimeter Drive MS 3051, Moscow, ID 83844, USA

[3]University Gardens, University Museum, University of Bergen, Mildeveien 240, 5259 Hjellestad, Norway

[4]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center of Genetics and Development, Ministry of Education Key Laboratory of Biodiversity and Ecological Engineering, Institute of Plant Biology, Center of Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200433, China

[5]Department of Botany and Plant Pathology, Oregon State University, 2082 Cordley Hall, Corvallis, OR 97331, USA

[6]Department of Biology, the Huck Institute of the Life Sciences, the Pennsylvania State University, 510D Mueller Laboratory, University Park, PA 16802 USA

[7]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109-1048, USA

24 * Correspondence to be sent to: Diego F. Morales-Briones. Department of Plant and Microbial
25 Biology, University of Minnesota, 1445 Gortner Avenue, St. Paul, MN 55108, USA, Email:
26 dfmoralesb@gmail.com

***Abstract.*** —Target enrichment is becoming increasingly popular for phylogenomic studies. Although baits for enrichment are typically designed to target single-copy genes, paralogs are often recovered with increased sequencing depth, sometimes from a significant proportion of loci. Common approaches for processing paralogs in target enrichment datasets include removal, random selection, and manual pruning of loci that show evidence of paralogy. These approaches can introduce errors in orthology inference, and sometimes significantly reduce the number of loci, especially in groups experiencing whole-genome duplication (WGD) events. Here we used an automated approach for paralog processing in a target enrichment dataset of 68 species of *Alchemilla* s.l. (Rosaceae), a widely distributed clade of plants primarily from temperate climate regions. Previous molecular phylogenetic studies and chromosome numbers both suggested the polyploid origin of the group. However, putative parental lineages remain unknown. By taking paralogs into consideration, we identified four nodes in the backbone of *Alchemilla* s.l. with an elevated proportion of gene duplication. Furthermore, using a gene-tree reconciliation approach we established the autopolyploidy origin of the entire *Alchemilla* s.l. and the nested allopolyploidy origin of four clades within the group. Here we showed the utility of automated orthology methods, commonly used in genomic or transcriptomic datasets, to study complex scenarios of polyploidy and reticulate evolution from target enrichment datasets.

**Keywords:** *Alchemilla*; allopolyploidy; autopolyploidy; gene tree discordance; orthology inference; paralogs; Rosaceae; target enrichment; whole genome duplication.

47 Polyploidy, or whole genome duplication (WGD), is prevalent throughout the evolutionary
48 history of plants (Cui et al. 2006; Jiao et al. 2011; Jiao et al. 2012; Leebens-Mack et al. 2019). As
49 a result, plant genomes often contain large numbers of paralogous genes from recurrent gene and
50 genome duplication events. With very few nuclear genes being truly single- or low-copy, careful
51 evaluation of orthology is critical for phylogenetic analyses (Fitch 1970). Orthology inference
52 has received much attention in the phylogenomic era with multiple pipelines available for this
53 task (e.g., Li et al. 2003; Dunn et al. 2013; Kocot et al. 2013; Yang and Smith 2014; Emms and
54 Kelly 2019, also see Glover et. al 2019 and Fernández et al. 2020 for recent reviews). All of
55 these approaches were developed for genome and transcriptome datasets. So far, few studies
56 have employed automated, phylogeny-aware orthology inference in target enrichment datasets.
57 The most common approach for resolving paralogy in target enrichment datasets is removing
58 loci that show evidence of potential paralogy (e.g., Nicholls et al. 2015; Jones et al. 2019;
59 Andermann et al. 2020; but see Moore et al. 2018). Removal of paralogous genes might be
60 appropriate in target enrichment datasets in which only a small number of loci show evidence of
61 paralogy (e.g., Larridon et al. 2020), but in some datasets this could result in a significant
62 reduction of the number of loci (e.g., Montes et al. 2019). Given the prevalence of WGD in
63 plants and the continuous increase of sequencing depth in typical target enrichment studies, the
64 latter scenario is becoming increasingly frequent, calling for automated, tree-based approaches
65 for processing paralogs in target enrichment data.

66 The ability to explicitly process paralogs opens the door for using target enrichment data
67 for inferring gene duplication events and pinpointing the phylogenetic locations of putative
68 WGDs. In the past, the phylogenetic placement of WGD events have most often been carried out
69 using genome and transcriptome sequencing data (e.g., Li et al. 2015; Huang et al. 2016; McKain

70 et al. 2016; Yang et al. 2018) using either the synonymous distance between of paralogs gene
71 pairs (Ks; Lynch and Conery 2000) or tree-based reconciliation methods (e.g., Jiao et al. 2011;
72 Li et al. 2015; Yang et al. 2015; Huang et al. 2016; Xiang et al. 2017; Leebens-Mack et al.
73 2019). Recently, Viruel et al. (2019) showed that allele frequency information from target
74 enrichment data can be informative in inferring recent WGDs. Target enrichment methods (e.g.,
75 Mandel et al. 2014; Weitemier et al. 2014; Buddenhagen et al. 2016) have been widely adopted
76 to collect hundreds to over a thousand nuclear loci for plant systematics, allowing studies at
77 different evolutionary scales (e.g., Villaverde et al. 2018), and the use of museum-preserved
78 collections (e.g., Forrest et al. 2019). This creates new opportunities to adopt tree-reconciliation
79 based methods to explore WGD patterns in groups for which genomic and transcriptomic
80 resources are not available or feasible.

81 With >1,100 species worldwide, *Alchemilla* in the broad sense has been a challenging
82 group to study due to the presence of reticulate evolution, polyploidy, and apomixis (Table S1).
83 Based on previous phylogenetic analyses, *Alchemilla* s.l. contains four clades: Afromilla,
84 *Aphanes,* Eualchemilla (or *Alchemilla* s.s.), and *Lachemilla*. Together they form a well-supported
85 clade nested in the subtribe Fragariinae, which also includes the cultivated strawberries (Gehrke
86 et al. 2008). Unlike the more commonly recognized members of the rose family (Rosaceae),
87 *Alchemilla* s.l. is characterized by small flowers with no petals, and a reduced number [1–4(–5)]
88 of stamens that have anthers with one elliptic theca on the ventral side of the connective that
89 opens by one transverse split (Perry 1929; Soják 2008). Gehrke et al. (2008) presented the first
90 phylogeny of *Alchemilla* s.l. and established the paraphyly of traditional *Alchemilla* s.s. into a
91 primarily African clade, Afromilla, and a Eurasian clade, Eualchemilla. Gehrke et al. (2008) also
92 suggested treating Afromilla and Eualchemilla, along with *Aphanes* and *Lachemilla* as a single

genus based on nomenclatural stability and the lack of morphological characters to distinguish between Afromilla and Eualchemilla. The four clades within *Alchemilla* s.l. are mainly defined by geographic distribution, as well as the number and insertion of the stamens on the disk lining the hypanthium (Table 1). Phylogenetic analyses using at least one nuclear and one chloroplast marker (Gehrke et al. 2008; 2016) found significant cytonuclear discordance regarding the relationships among the four major clades. Similar patterns, often attributed to hybridization and allopolyploidy, have been detected among most genera in Fragariinae (Lundberg 2009; Eriksson et al. 2015; Gehrke et al. 2016, Kamneva et al. 2017; Morales-Briones et al. 2018a), leaving the phylogenetic relationships of *Alchemilla* s.l. to the rest of Fragariinae unresolved. Recent phylogenomic analyses using target enrichment that included 32 species of *Lachemilla* detected a putative WGD event (or multiple events) shared by all *Alchemilla* s.l., based on high frequency of paralogs (Morales-Briones et al. 2018b). However, sampling was limited to one species each of Eualchemilla, Afromilla, and *Aphanes*, and the location of this shared WGD remains unresolved.

In this study we increased sampling to 68 species across the major clades of *Alchemilla* s.l., and included 11 additional closely related species in Fragariinae, which allowed us to 1) test for polyploid events in the origin of *Alchemilla* s.l., and 2) explore the reticulate evolution within the clade using a target enrichment dataset. Given the prevalence of polyploidy and reticulate history within *Alchemilla* s.l., this is an excellent group to explore the utility of tree-based methods for (1) processing paralogs, and (2) detecting and placing WGDs using target enrichment datasets.

## MATERIALS AND METHODS

### *Taxon sampling and data collection*

We sampled 68 species representing the four major clades of *Alchemilla* s.l. (sensu Gehrke et al. 2008), and 11 species to represent all other genera in Fragariinae (except *Chamaecallis*; sensu Morales-Briones and Tank 2019). Additionally, we sampled one species each of *Potentilla*, *Sanguisorba*, and *Rosa* as outgroups. Voucher information is provided in Table S2. We used a Hyb-Seq approach (Weitemier et al. 2014), that combines target enrichment and genome skimming, to target 257 putatively single-copy orthologous genes (in 1,419 exons) using baits designed for *Fragaria vesca* (strawberries, also member of Fragariinae; Kamneva et al. 2017).

Of the 82 total species, only sequences for *Fragaria vesca*, were from a reference genome (Shulaev et. al 2010). Twenty-two were from a previously published Hyb-Seq dataset using the same bait set as this study (Morales-Briones et al. 2018b; Table S2), including 19 species of *Lachemilla* that did not show evidence of hybridization within *Lachemilla,* and one species each of Eualchemilla, Afromilla, and *Aphanes*. Newly generated sequence data for 55 species (Table S2) were collected as follows. Total genomic DNA was isolated from silica-dried or herbarium material with a modified CTAB method (Doyle and Doyle 1987). Probe synthesis, library preparation, capture enrichment, and high-throughput sequencing (HiSeq2000 instrument, 2 × 101 bp) were carried out at Rapid Genomics LLC (Gainesville, FL, USA). Data for the remaining four species, *Drymocallis glandulosa*, *Potentilla indica*, *Rosa woodsii*, and *Sanguisorba menziesii* were collected as described in Kamneva et al. (2017).

139 *Read processing and assembly*

140 We removed sequencing adaptors and trimmed low-quality bases (Phred scores < 20) from raw
141 reads with SeqyClean v.1.10.07 (Zhbannikov et al. 2017) using default settings. Plastomes were
142 assembled using Alignreads v.2.5.2 (Straub et al. 2011) and 12 closely related plastome
143 references (with one Inverted Repeat removed; Table S3). Plastome assemblies were annotated
144 using *Fragaria vesca* as a reference in Geneious v.11.1.5 (Kearse et al. 2012). Assembly of
145 nuclear loci was carried out with HybPiper v.1.3.1 (Johnson et al. 2016) using exons of *F. vesca*
146 as references. Given the large number of paralogs detected in *Lachemilla,* Eualchemilla, and
147 Afromilla, multi-exon gene assemblies resulted in chimeric sequences of exons from distinct
148 paralogs (Morales-Briones et al. 2018b). To avoid chimeric sequences that can affect paralog
149 processing and phylogenetic analyses, assemblies were performed on each exon separately. Only
150 exons with a reference length of ≥ 150 bp were assembled (939 exons from 257 genes). Paralog
151 detection was carried out for all exons with the 'paralog_investigator' option in HybPiper. This
152 option flags loci with potential paralogs when multiple contigs cover at least 85% of the
153 reference sequence length. Exon assemblies that included flagged paralogs were extracted using
154 the 'paralog_retriever' command of HybPiper and used for orthology inference.

155

156 *Orthology inference for nuclear exons*

157 To infer orthologs for phylogenetic analyses, all exons were processed as follows (Fig. 1).
158 Individual exons were aligned using MACSE v.2.03 (Ranwez et al. 2018) with default
159 parameters. Codons with frameshifts (labeled with '!' by MACSE) were replaced with gaps, and
160 aligned columns with more than 90% missing data were removed using Phyx (Brown et al.
161 2017). Initial homolog trees were built using RAxML v.8.2.11 (Stamatakis 2014) with a

162 GTRCAT model and clade support assessed with 100 rapid bootstrap (BS) replicates.
163 Monophyletic and paraphyletic grades that belonged to the same taxon were masked by keeping
164 only the tip with the highest number of characters in the trimmed alignment following Yang and
165 Smith (2014). To obtain the final homolog trees, outlier tips with unusually long branches were
166 detected and removed by maximally reducing the tree diameter with TreeShrink v.1.3.2 (Mai and
167 Mirarab 2018). Orthology inference was carried out using two outgroup-aware strategies from
168 Yang and Smith (2014). First, we used the 'monophyletic outgroup' (MO) approach keeping
169 only ortholog groups with at least 25 ingroup taxa. The MO approach filters for homolog trees
170 with outgroup taxa being monophyletic and single-copy, and therefore filters for single- and low-
171 copy genes. The second approach used was the 'rooted tree' (RT), with at least 25 ingroup taxa.
172 The RT approach searches for the subtree with the highest number of ingroup taxa and cuts it out
173 as a rooted tree. Both approaches root the gene tree by the outgroups, traverse the rooted tree
174 from root to tip, and remove the side with fewer taxa (MO) or keep both sides (RT) when gene
175 duplication is detected at any given node. The RT approach maximizes the number of orthologs
176 compared to MO, but removes outgroups. To add outgroups back to the RT orthologs for
177 downstream analyses, we kept only RT orthologs from homologs that had a MO ortholog. Then
178 we used the outgroups of the MO ortholog for all the RT orthologs of the same homolog (Fig 1).
179 Scripts for orthology inference can be found at
180 https://bitbucket.org/dfmoralesb/target_enrichment_orthology.
181

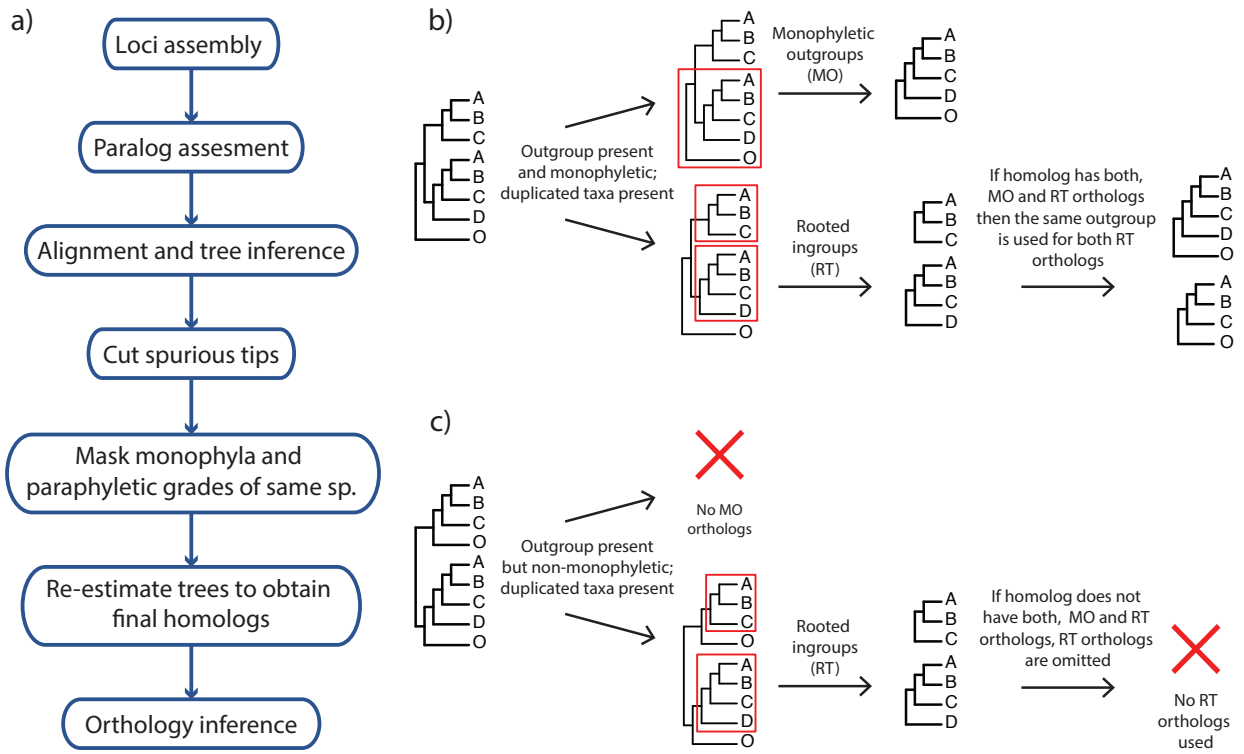


**Figure 1.** Paralog processing workflow and orthology inference methods used in *Alchemilla* s.l. homolog trees. a) Flow chart of paralog processing and homolog tree inference. b) Outgroup present and monophyletic. Monophyletic outgroups (MO) will prune single-copy genes keeping clades with at least a user-defined minimum number of ingroup taxa. Rooted ingroups (RT) will keep all subtrees with at least a user-defined minimum number of ingroups taxa. If the homolog trees can be pruned using both MO and RT, then RT orthologs are added to the same root. c) Outgroup present but non-monophyletic. MO orthologs cannot be obtained from this kind of homologs. RT orthologs can be pruned from this kind homologs but do not have a root. These RT orthologs can be used for phylogenetic inference (e.g., species tree inference), but were not used here as some downstream analyses require rooted orthologs.

### *Phylogenetic analyses*

We used concatenation and coalescent-based methods to reconstruct the phylogeny of *Alchemilla* s.l. Analyses were carried out in the two sets of final orthologs, MO and RT, separately. Each ortholog exon was aligned using MACSE v.2.03 with default parameters. Codons with frameshifts were replaced with gaps, and aligned columns with more than 90% missing data

198 were removed using Phyx, and alignments with at least 150 characters and 25 taxa were retained.
199 We first estimated a maximum likelihood (ML) tree from the concatenated matrices with
200 RAxML using a partition by gene scheme with a GTRGAMMA model for each partition. Clade
201 support was assessed with 100 rapid bootstrap (BS) replicates. To estimate a species tree that is
202 statistically consistent with the multi-species coalescent (MSC), we first inferred individual ML
203 gene trees using RAxML with a GTRGAMMA model, and 100 BS replicates to assess clade
204 support. Individual gene trees were then used to estimate a species tree using ASTRAL-III
205 v.5.6.3 (Zhang et al. 2018) using local posterior probabilities (LPP; Sayyari and Mirarab 2016) to
206 assess clade support.

207 To evaluate nuclear gene tree discordance, we calculated the internode certainty all (ICA)
208 value to quantify the degree of conflict on each node of the species tree given individual gene
209 trees (Salichos et al. 2014). Also, we calculated the number of conflicting and concordant
210 bipartitions on each node of the species trees. We calculated both the ICA scores and the number
211 of conflicting/concordant bipartitions with Phyparts (Smith et al. 2015) using individual gene
212 trees with BS support of at least 50% for the corresponding node. Additionally, to distinguish
213 strong conflict from weakly supported branches, we evaluated tree conflict and branch support
214 with Quartet Sampling (QS; Pease et al. 2018) using 1,000 replicates. Quartet Sampling
215 subsamples quartets from the input species tree and concatenated alignment to assess the
216 confidence, consistency, and informativeness of each internal branch by the relative frequency of
217 the three possible quartet topologies at each node (Pease et al. 2018).

218 In addition to species tree construction using inferred orthologs, we used a recently
219 developed quartet-based species tree method (ASTRAL-Pro; Zhang et al. 2019) to estimate the
220 phylogeny of *Alchemillla* s.l. ASTRAL-Pro directly uses multi-labeled gene trees while

221 accounting for gene duplications and losses to estimate a species tree that is statistically
222 consistent with the MSC and birth-death gene duplication and loss model. We used all 923 final
223 homolog trees as input for ASTRAL-Pro, ignoring trees with less than 20 taxa, and estimated
224 LPP to assess clade support. Additionally, we calculated ICA scores and the number of
225 conflicting/concordant bipartitions with Phyparts using homolog trees with BS support of at least
226 50% for the corresponding nodes.

227 For the plastome phylogenetic analyses, 74 partial plastome assemblies and eight
228 reference plastome sequences were included (Table S3). Plastome sequences were aligned using
229 the automatic alignment strategy in MAFFT v.7.307 (Katoh and Standley 2013), and aligned
230 columns with more than 70% missing data were removed with Phyx. We estimated an ML tree
231 of the plastome alignment with RAxML using a partition by coding (CDS) and noncoding
232 regions (introns and intergenic spacers) scheme, with a GTRGAMMA model for each partition
233 and clade support assessed with 100 rapid BS replicates and QS using 1,000 replicates, to detect
234 potential plastome conflict in the backbone of *Alchemilla* s.l. as recently reported in other groups
235 (e.g., Gonçalves et al. 2019; Walker et al. 2019; Morales-Briones et al. 2020; Zhang et al. 2020).
236

237 *Mapping whole genome duplications*

238 We took two alternative approaches for detecting WGD events by mapping gene duplication
239 events from gene trees onto the species tree. The first approach begins by extracting orthogroups
240 from the final homolog trees. Orthogroups are rooted ingroup lineages separated by outgroups
241 that include the complete set of genes in a lineage from a single copy in their common ancestor.
242 We extracted orthogroups requiring at least 50 out of 79 species in Fragariinae. Gene duplication
243 events were then recorded on the most recent common ancestor (MRCA) on the species tree

244 when two or more species overlapped between the two daughter clades (Yang et al. 2018). Each
245 node on a species tree can be counted only once from each gene tree to avoid nested gene
246 duplications inflating the number of recorded duplications (Yang et al. 2018). We mapped
247 duplication events onto both the MO and RT species trees using orthogroups from all 923 final
248 homologs while filtering orthogroups using an BS average of at least 50%. We carried out the
249 mapping using all exons and using only the longest exon per gene to avoid inflating the counts in
250 multi-exon genes.

251 For the second strategy we explicitly tested for polyploidy mode using GRAMPA
252 (Thomas et al. 2017). GRAMPA uses MRCA reconciliation with multi-labeled gene trees to
253 compare allo-or autopolyploid scenarios in singly- or multi-labeled species trees. To reduce the
254 computational burden of searching all possible reconciliations, we constrained searches to only
255 among crown nodes of major clades of *Alchemilla* s.l. (including the 'dissected' and 'lobed' clades
256 of Eualchemilla; see results) and genera within Fragariinae. We ran reconciliation searches using
257 all 923 final homologs, as well as using only the longest exon per gene (256), against either the
258 MO or RT species tree. We expected multiple WGD events within *Alchemilla* s.l. (see results),
259 but GRAMPA can only infer one WGD at a time. To disentangle nested duplication events, we
260 also carried out similar GRAMPA reconciliations using the MO species tree and sequentially
261 excluding major groups of *Alchemilla* s.l. that were identified as a polyploidy clade. We only
262 used the MO species tree as the only difference with the RT species trees is the location of the
263 'lobed' clade that was first clade identified as allopolyploid (see results). The removal of the
264 'lobed' clade from either the MO or RT species trees resulted in the same backbone for the
265 remaining *Alchemilla* s.l. Finally, to test for a polyploid origin of *Alchemilla* s.l., we carried out
266 searches among the constrained crown node of *Alchemilla* s.l. and the rest of the genera within

267 Fragariinae using the MO and cpDNA trees. The backbone of Fragariinae differed between the
268 MO (same as RT) species tree and the cpDNA tree. Thus, we tested how this affected the
269 inference of the polyploidy origin of *Alchemilla* s.l. We also carried out similar searches but
270 using each of the five major clades individually.
271

272 *Distribution of synonymous distance among gene pairs (Ks plots)*

273 To verify WGD events inferred from target enrichment, we analyzed the distribution of
274 synonymous distances (Ks) from RNA-seq data of four species of *Alchemilla* s.l. and nine
275 species of Fragariinae (Table S4). Read processing and transcriptome assembly followed
276 Morales-Briones et al. (2020). For each of the four species, a Ks plot of within-taxon paralog
277 pairs based on BLASTP hits was done following Yang et al. (2015). Significant Ks peaks were
278 identified using a mixture model as implemented in mixtools v.1.2.0 (Benaglia et al. 2009). The
279 optimal number of mixing components was estimated using parametric bootstrap replicates of
280 the likelihood ratio test statistic (McLachlan and Peel 2000). We tested up to five components
281 using 500 bootstrap replicates in mixtools.
282

283 **RESULTS**

284 *Assembly and orthology inference*

285 The number of assembled exons per species (with > 25% of the target length) ranged from 692
286 (*Alchemilla fissa*) to 939 (*Farinopsis salesoviana*) out of 939 single-copy exon references from
287 *F. vesca*, with an average of 902 exons (Table S5). The number of exons with paralog warnings
288 ranged from 10 in *Drymocallis glandulosa* to 746 in *Alchemilla mollis* (Table S5). The number
289 of exon alignments with ≥ 25 species was 923 from 256 genes. The orthology inference resulted

290 in 914 MO orthologs (Table S6), and 1,906 RT orthologs (Table S6). The trimmed alignments of
291 the MO orthologs ranged from 136 to 5,740 characters with a mean of 425 characters (median =
292 268). The concatenated alignment of the MO orthologs, with at least 150 aligned characters and
293 25 species for each exon, included 910 exons and 387,042 characters with a matrix occupancy of
294 66%. The trimmed alignments of the RT orthologs ranged from 136 to 5,740 characters with a
295 mean of 394 characters (median = 259). The concatenated alignment of the RT orthologs, with at
296 least 150 aligned characters and 25 species, included 1,894 exons and 746,562 characters with a
297 matrix occupancy of 54%. The chloroplast alignment included 124,079 characters.
298

299 *Nuclear phylogenetic analyses*

300 All nuclear analyses recovered the monophyly of *Alchemilla* s.l. with maximum support (i.e.,
301 bootstrap percentage [BS] = 100, local posterior probabilities [LPP] = 1.0; Fig. 2; Fig. S1), most
302 gene trees being concordant (858 out of 863 for MO; 977/984 for RT; 912/932 for ASTRAL-
303 Pro; ICA = 0.95), and full QS support (1.0/–/1.0; i.e., all sampled quartets supported that node).
304 Five major clades were identified within *Alchemilla* s.l.: Afromilla, *Aphanes*, Eualchemilla-
305 'dissected', Eualchemilla-'lobed,' and *Lachemilla*. Moreover, the relationships among these
306 clades showed high levels of discordance and varied among the MO and RT species trees.
307 Analyses of the MO ortholog exons using ASTRAL and concatenated ML approaches
308 resulted in similar topologies for the backbone of *Alchemilla* s.l. (Fig. 2). The monophyly of the
309 five major clades each received maximum support (BS = 100; LPP = 1.0) and had most gene
310 trees being concordant (except for the two clades of Eualchemilla). Eualchemilla was
311 paraphyletic and split into the 'dissected' and 'lobed' clades. The 'dissected' clade was supported
312 by 73 out of 420 informative gene trees (ICA = 0.06) and strong QS score (0.61/0.98/0.99),

313 while the ‘lobed’ clade was supported by 113 out of 316 informative gene trees (ICA = 0.06) and
314 strong QS score (0.87/0.34/1). In both cases, the ‘dissected’ and ‘lobed’ clades had a relatively
315 small percentage of supporting gene trees, but the conflict analysis and QS score did not reveal
316 any well-supported alternative topology. *Aphanes* was recovered as sister to the Eualchemilla-
317 ‘lobed’ clade with relatively low support (BS = 90, LPP = 0.62), 60 concordant gene trees (out of
318 430 informative gene trees; ICA = 0.08), and weak QS score (0.016/0.95/0.98) with similar
319 frequencies for all three possible alternative topologies. The Eualchemilla-‘dissected’ clade was
320 recovered as sister to Eualchemilla-‘lobed’ + *Aphanes* with maximum support, 279 concordant
321 gene trees (out of 482 informative gene trees; ICA = 0.29), and full QS score. Afromilla was
322 recovered as sister to the clade formed of Eualchemilla (‘dissected and ‘lobed’) and *Aphanes*
323 with high to low support (BS = 100, LPP = 0.88), only 146 concordant gene trees (out of 413
324 informative gene trees; ICA = 0.22), and weak QS support (0.2/0.44/0.99) with a skew in
325 discordance suggesting a possible alternative topology. Lastly, *Lachemilla* was recovered as the
326 sister to the rest of *Alchemilla* s.l.

327 Analysis of the RT ortholog exons using ASTRAL and concatenated ML approaches
328 both recovered the same major clades, but they differed in the relationship among these five
329 clades (Fig 2a; Fig. S1). In both analyses, *Lachemilla*, Afromilla, and *Aphanes* had maximum
330 support (BS = 100; LPP = 1.0) and had most gene trees being concordant. Eualchemilla was
331 recovered as monophyletic and composed of the 'dissected' and 'lobed' clades. Eualchemilla had
332 high to low support (BS = 99, LPP = 0.63), only 231 concordant gene trees (out of 819
333 informative gene trees; ICA = 0.07), and weak QS support (0.023/0.87/0.98) with similar
334 frequencies for all three possible alternative topologies. As in the MO analyses, the ‘dissected’
335 and ‘lobed’ clades had the maximum support, low number of concordant gene trees (218 out of

336 557 [ICA = 0.19] and 136 out of 707 [ICA = 0.08], respectively), and strong QS support
337 (0.98/0/1 and 0.62/0.17/0.99, respectively). Both the 'dissected' and 'lobed' clades had a reduced
338 number of supporting gene trees. Still, in the case of 'dissected' clade, the conflict analysis and
339 QS score did not reveal any well-supported alternative topology. In contrast, in the 'lobed' clade,
340 only the QS score showed the possible presence of a supported secondary topology.
341 Eualchemilla was recovered as sister of *Aphanes* with maximum support (BS = 100), 348
342 concordant gene trees (out of 728 informative gene trees; ICA = 0.29) and full QS support. The
343 ML concatenated tree (Fig 2a; Fig. S1) placed Afromilla as sister to the clade formed of
344 Eualchemilla and *Aphanes* with maximum support (BS = 100), 212 concordant gene trees (out of
345 771 informative gene trees; ICA = 0.27), and weak QS support (0.18/0.66/0.99) with a small
346 skew in discordance suggesting a possible alternative topology. *Lachemilla* was placed as sister
347 to the rest of *Alchemilla* s.l. The ASTRAL tree in turn (Fig S1a), retrieved *Lachemilla* as sister to
348 the clade formed of Eualchemilla and *Aphanes* with no support (LPP = 0.01), 247 concordant
349 gene trees (out of 953 informative gene trees; ICA = 0.19), and QS counter-support
350 (−0.21/0.29/0.99), showing that the majority of the quartets supported one of the alternatives
351 topologies. In this case, Afromilla was placed as sister to the rest of *Alchemilla* s.l.
352 The ASTRAL-Pro analysis using multi-labeled homolog trees recovered the same
353 backbone topology of *Alchemilla* s.l. as the concatenated ML analysis from the RT orthologs
354 (Fig. 2b; Fig. S3). All five major clades had the maximum support (LPP = 1.0). Eualchemilla,
355 composed of the 'dissected' and 'lobed' clades, had high to low support (LPP = 0.76) and only
356 502 concordant trees (out of 1,274 informative trees; ICA = 0.09). The 'dissected' and 'lobed'
357 clades had low numbers of concordant trees (402 out of 1,157 [ICA = 0.13] and 97 out of 975
358 [ICA = 0.05], respectively), but did not showed signal of any well-supported alternative

topology. *Aphanes* was placed as the sister of Eualchemilla with maximum support (LPP = 1.0), 577 concordant trees (out of 1,329 trees; ICA = 0.16), and no support for any well-supported alternative topology. Afromilla was recovered as sister to the clade formed of Eualchemilla and *Aphanes* with low support (LPP = 0.52), 649 concordant trees (out of 1,415 trees; ICA = 0.20), and no support for any well-supported alternative topology.

*Chloroplast phylogenetic analyses*

The chloroplast ML tree (Fig 2c; Fig S4) recovered a well-supported backbone *Alchemilla* s.l. where the monophyly of *Aphanes,* Afromilla, and *Lachemilla,* had maximum or near maximum support (i.e., bootstrap percentage [BS] = 100, QS support = [1.0/–/1.0]). Eualchemilla, composed of the 'dissected' and 'lobed' clades, also had the maximum support. The ‘dissected’ clade had strong support (BS = 75), but the QS score (0.8/0.43/0.88) showed a small skew in discordance suggesting a possible alternative topology. Similarly, the ‘lobed’ clade also had strong support (BS = 100) and a QS score (0.95/0.25/0.92) that suggested a possible alternative topology. *Aphanes* and Eualchemilla formed, with maximum support, a clade as in the nuclear analyses. In turn, Afromilla and *Lachemilla* were recovered as sister clades with maximum support, which differed from the nuclear analyses.

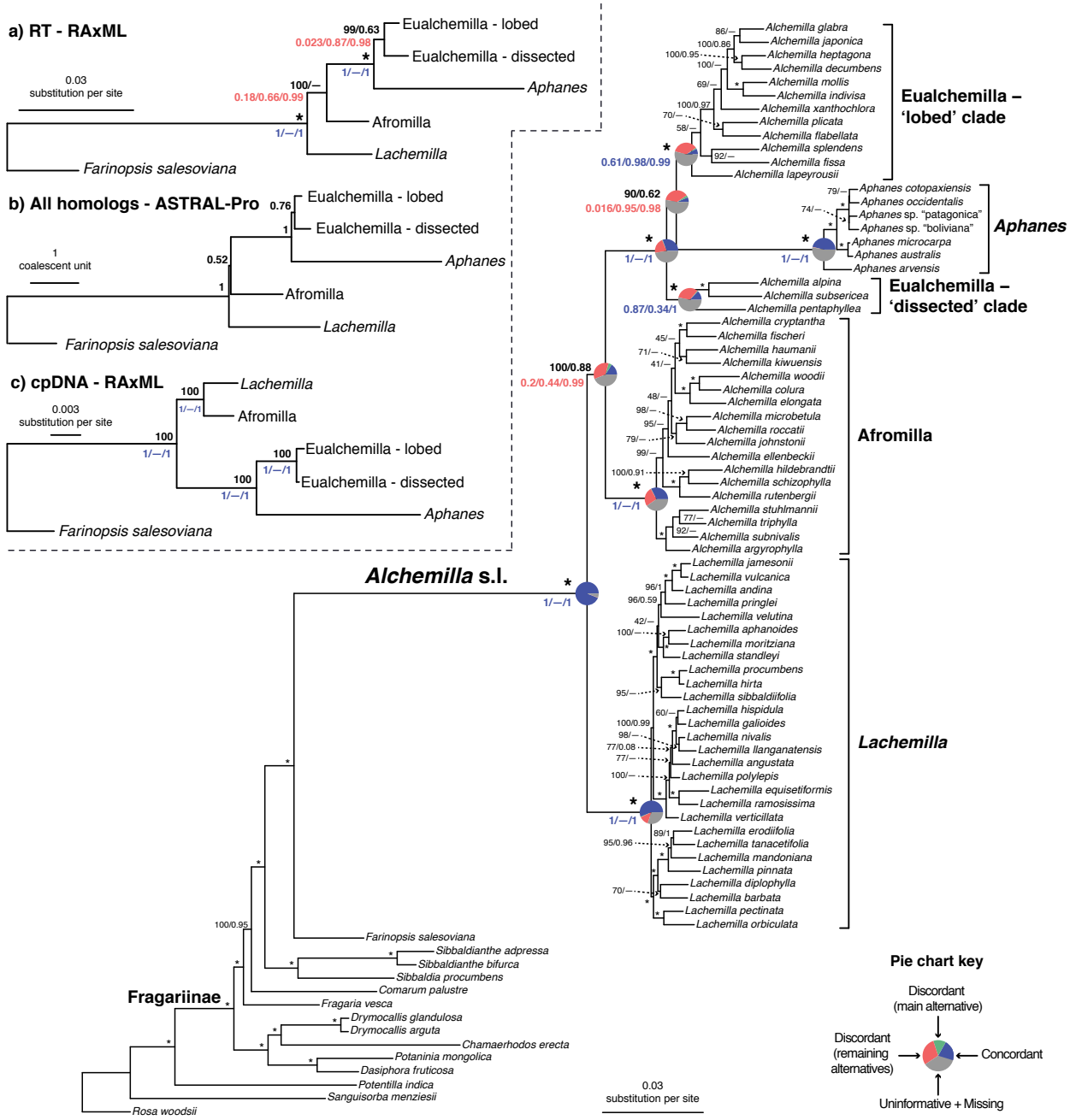


**Figure 2.** Maximum likelihood phylogeny of *Alchemilla* s.l. inferred from RAxML analysis of the concatenated 910-nuclear gene supermatrix from the 'monophyletic outgroup' (MO) orthologs. Bootstrap support (BS) and Local posterior probability (LLP) are shown above branches. Nodes with full support (BS= 100/LLP= 100) are noted with an asterisk (*). Em dashes (—) denoted alternative topology compared to the ASTRAL species tree. Quartet Sampling (QS) scores for major clades are shown below branches. QS scores in blue indicate strong support and red scores indicate weak support. QS scores: Quartet concordance/Quartet differential/Quartet informativeness. Pie charts for major clades represent

the proportion of exon ortholog trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support plus missing data (gray). Branch lengths are in number of substitutions per site (scale bar on the bottom). Inset: a) Summary Maximum likelihood phylogeny inferred from RAxML analysis of the concatenated 1,894-nuclear gene supermatrix from the 'rooted ingroup' orthologs (RT). BS and LLP are shown above branches and QS scores below the branches. Branch lengths are in number of substitutions per site; b) Summary ASTRAL-Pro species trees inferred from 923 multi-labeled homolog trees. LLP are shown next to nodes. Branch lengths are in coalescent units. c) Summary Maximum likelihood phylogeny inferred from RAxML analysis of partial plastomes. BS and LLP are shown above branches and QS scores below the branches. Branch lengths are in number of substitutions per site.

*Mapping whole genome duplications*

By mapping the most recent common ancestor (MRCA) of gene duplication events from rooted orthogroup trees onto the species tree, we found four nodes in *Alchemilla* s.l. that each had an elevated proportion of gene duplications (Fig. 3a–b). This trend was consistent regardless of using all 923 homologs exons (826 after orthogroup inference and BS filtering) or using only the 256 longest exons for each gene (250 after orthogroup inference and BS filtering; Fig. S5). Therefore, here we describe the results only for the latter. These four clades include (Fig. 3a; Fig. S5): 1) the MRCA of *Alchemilla* s.l. (86.0% of the 250 genes show evidence of duplication), 2) the MRCA of Eualchemilla , *Aphanes*, and Afromilla (34.4%), 3) the MRCA of Eualchemilla + *Aphanes* (MO: 18.4%; RT:15.6% ), and 4) the MRCA of *Lachemilla* (18.4%). These four nodes have an elevated proportion of gene duplications compared to all other nodes in Fragariinae (Fig. 3b) and it is consistent with the number of paralogs counted from the final homolog trees (Fig. 3c). Interestingly, *Aphanes* had a lower number of paralogs than the rest of *Alchemilla* s.l., resembling the other members of Fragariinae (Fig. 3c).

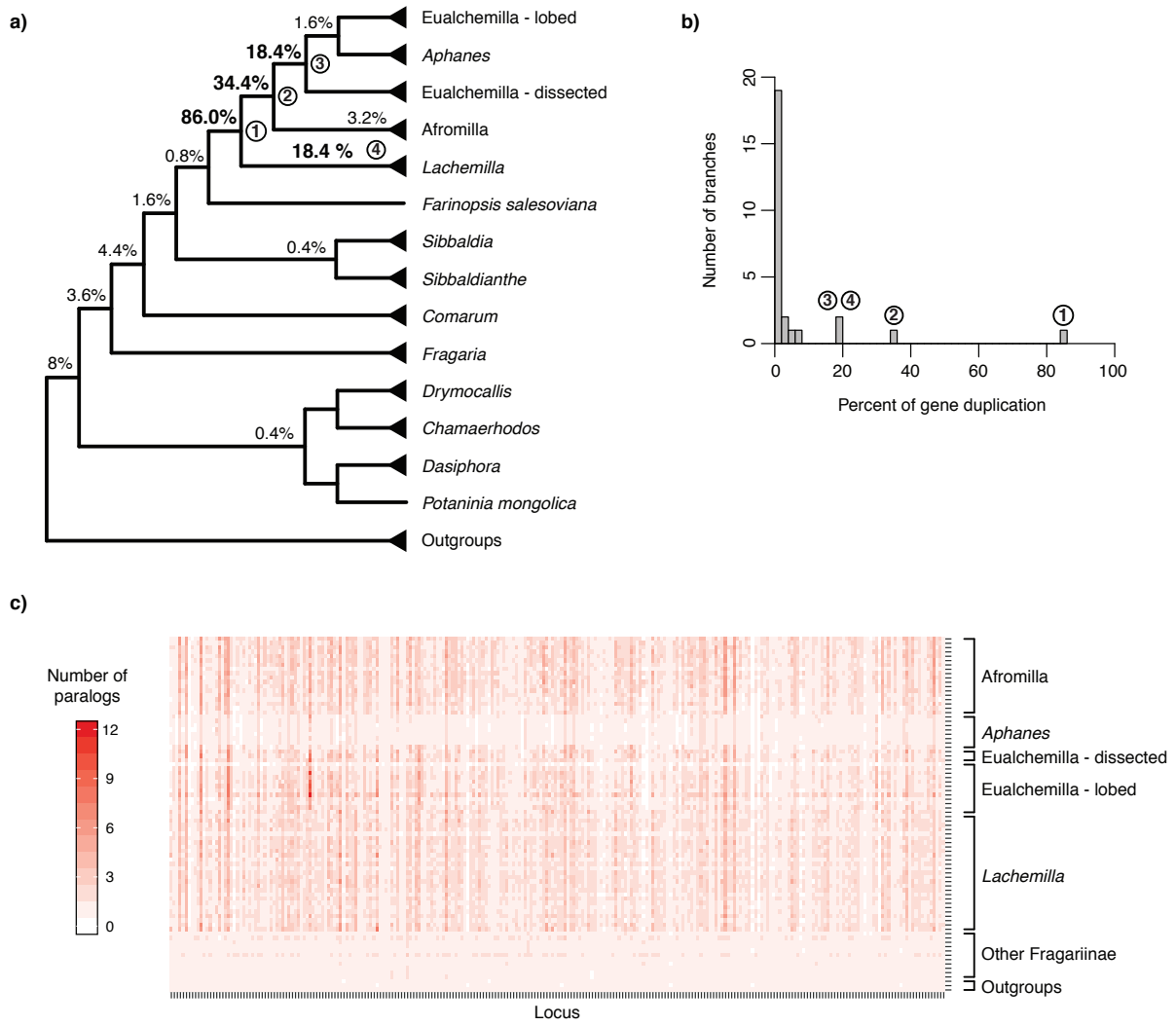


**Figure 3.** Orthogroup gene duplication mapping results. a) Summarized cladogram of *Alchemilla* s.l. from the 'monophyletic outgroup' (MO) orthologs species trees. Percentages next to nodes denote the proportion of duplicated genes when using orthogroups from only one exon per gene. Nodes with elevated proportions of gene duplications are numbered 1–4 as referenced in the main text. See Fig. S6 for the full tree. b) Histogram of percentages of gene duplication per branch. c) Number of paralogs per taxa in the final homolog trees. Each locus is represented by the longest exon per gene.

Similar to the results of MRCA mapping, the GRAMPA analyses recovered the same results when using all 923 homolog exons or only one exon per gene (256). GRAMPA reconciliations using all major clades of *Alchemilla* s.l. recovered multi-labeled trees with the

423 best score where the 'lobed' clade of Eualchemilla was of an allopolyploid origin, but the putative
424 parental lineages varied between the MO and RT ortholog species trees. The reconciliations
425 using the MO species tree (Fig. 4a; Fig. S6) showed that the 'lobed' clade was of allopolyploid
426 origin between an unsampled or extinct lineage sister to *Aphanes* and an unsampled or extinct
427 lineage (‘lineage’ for short hereafter) sister to 'dissected' + *Aphanes*. In turn, the reconciliations
428 using the RT species tree (Fig. S6) showed that the 'lobed' clade was of allopolyploid origin
429 between a ‘lineage’ sister to the 'dissected' clade, and also a ‘lineage’ sister to 'dissected' +
430 *Aphanes*.

431 The GRAMPA reconciliations performed on the MO species trees with removal of major
432 clades of *Alchemilla* s.l. inferred as allopolyploid resulted in the identification of additional
433 polyploidy events (Fig. 4b–d ). First, we removed the 'lobed' clade that resulted in the recovery
434 of Afromilla as an allopolyploid clade. Afromilla parental lineages were a ‘lineage’ sister to
435 *Aphanes* + the 'dissected' clade, and a ‘lineage’ sister to all remaining *Alchemilla* s.l. (Fig. 4b).
436 The further removal of Afromilla resulted in recovery of the 'dissected' clade as allopolyploid.
437 The ‘lobed’ clade had parental lineages to a ‘lineage’ sister to *Aphanes* and to a ‘lineage’ sister to
438 all remaining *Alchemilla* s.l. (Fig. 4c). Finally, the removal of the ‘dissected’ clade resulted in the
439 *Lachemilla* being recovered also as an allopolyploid clade. The parental lineages of *Lachemilla*
440 were a ‘lineage’ sister to *Aphanes* and a ‘lineage’ sister to all remaining *Alchemilla* s.l. (Fig. 4d).

441 The GRAMPA results from the analyses with constrained searches on the crown node of
442 *Alchemilla* s.l. recovered different modes of polyploidy when using the MO species tree versus
443 the cpDNA tree. The MO species tree had *Farinopsis*, *Sibbaldianthe* + *Sibbaldia*, *Comarum*, and
444 *Fragaria* forming a grade sister to *Alchemilla* s.l., while *Drymocallis*, *Chamaerhodos*, *Potaninia*,
445 and *Dasiphora* form a clade that is sister to all other Fragariinae (Fig. 2). The reconciliations on

446 the MO tree resulted in an allopolyploid event for the clade composed of *Alchemilla* s.l.,
447 *Farinopsis, Sibbaldianthe*, and *Sibbaldia* (Fig. S7). The parental lineages of this clade were a
448 'lineage' sister to *Comarum*, and a 'lineage' sister to the grade formed of *Comarum* and
449 *Fragaria* (Fig. S7). The reconciliations using individual major clades of *Alchemilla* s.l. resulted
450 in identical patterns as in the full constrained analysis (Fig. S8). The cpDNA tree had *Alchemilla*
451 s.l. as sister to the grade formed of *Farinopsis*, *Comarum*, and *Sibbaldianthe* + *Sibbaldia*, while
452 *Fragaria* was recovered as sister to the clade composed of *Drymocallis*, *Chamaerhodos*,
453 *Potaninia*, and *Dasiphora,* which is sister to all other Fragariinae (Fig. S6). The reconciliations
454 on the cpDNA tree recovered *Alchemilla* s.l. as of autopolyploid origin (Fig. 4e; Fig. S7). The
455 analyses using individual major clades of *Alchemilla* s.l. recovered identical patterns as in the
456 full constrained analysis, except for *Aphanes* that resulted in a singly-labeled tree (Fig. S8).
457 The distribution of synonymous distances in the four species of Eualchemilla (one
458 'dissected' and three 'lobed') shared three optimal mixing components with a Ks mean at
459 approximately 0.1, 0.34, and 1.67, respectively (Fig. S9). All nine species from other genera in
460 Fragariinae had two optimal mixing components with a shared Ks peak at 1.44–1.86. The only
461 two polyploid species from the other genera of Fragariinae, *Sibbaldianthe bifurca* (2*n*=28) and
462 *Comarum palustre* (2*n*=64), had KS peaks at 0.11 and 0.46, respectively (Fig. S9).

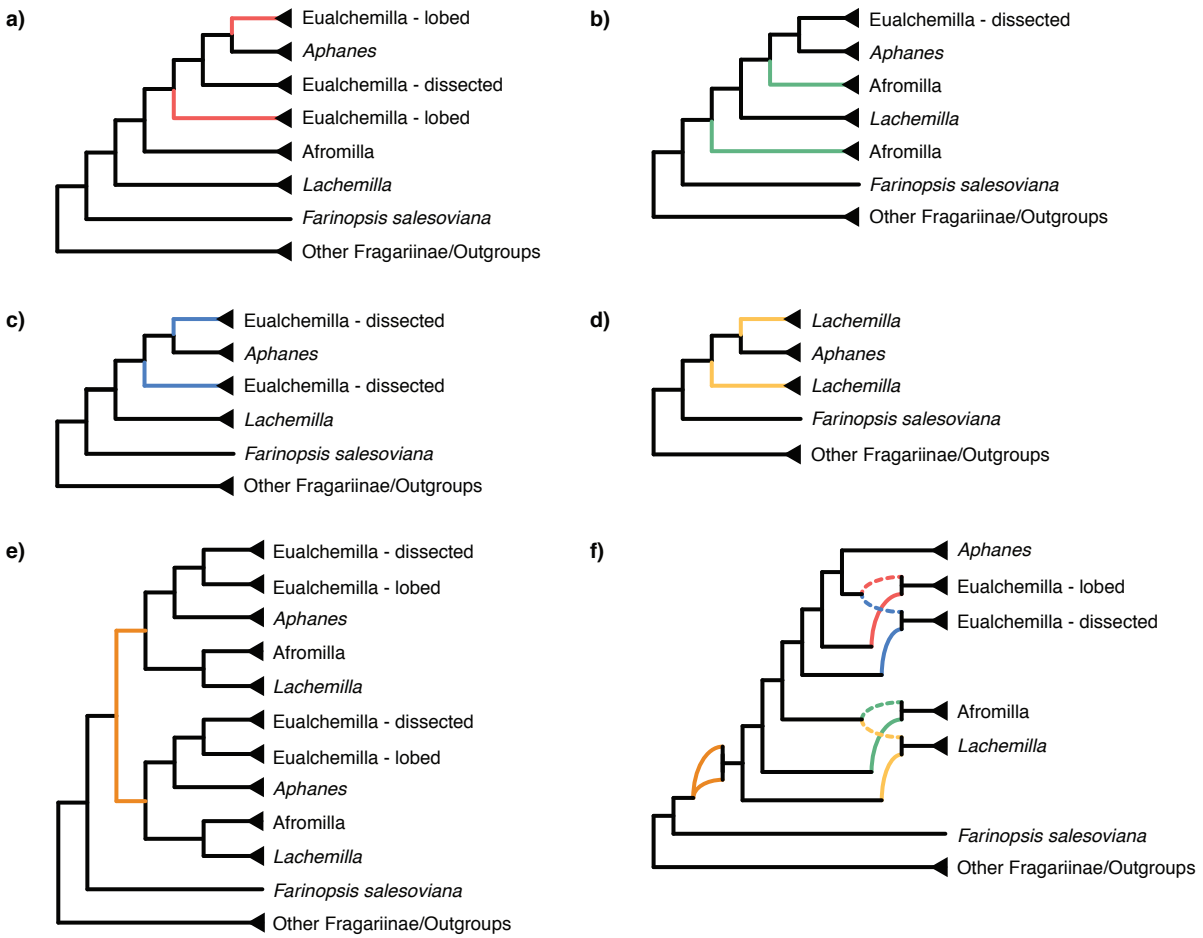


**Figure 4.** Summarized optimal multi-labeled tree (MUL-tree) inferred from GRAMPA analyses. a) MUL-tree for the complete 'monophyletic outgroup' (MO) orthologs species tree. Red branches denote the allopolyploid origin of the 'lobed' clade of Eualchemilla. b) MUL-tree for MO species trees after the removal of the 'lobed' clade of Eualchemilla. Green branches denote the allopolyploid origin of Afromilla. c) MUL-tree for MO species trees after the removal of the 'lobed' clade and Afromilla. Blue branches denote the allopolyploid origin of the 'dissected' clade of Eualchemilla. d). MUL-tree for MO species trees after the removal of the 'lobed' and 'dissected' clades, and Afromilla. Green lines denote the allopolyploid origin of *Lachemilla*. e) MUL-tree using constrained searches the crown node of *Alchemilla* s.l. on the cpDNA tree. Orange branches denote the autopolyploid origin *Alchemilla* s.l. f) Putative summary network of all reticulation events in *Alchemilla* s.l. Colored curved branches denote different polyploid events as in a–e. Dashed curved lines represent the maternal lineage (cpDNA) in allopolyploid events.

478 **DISCUSSION**

479 *Processing paralogs in target enrichment datasets*

480 The increased use of target enrichment methods in combination with reduced sequencing costs
481 and higher read coverage have facilitated the recovery of paralogs in such datasets. Paralogy is
482 viewed as a nuisance for phylogenetic reconstruction and is commonly aimed to be reduced in
483 early stages of experimental design, by targeting only single- or low-copy genes during the
484 selection of loci (e.g., Chamala et al. 2015; Nicholls et al. 2015; Gardner et al. 2016; Kamneva et
485 al. 2017). Still, the recovery of paralogs is inevitable when working with groups where WGD is
486 prevalent, especially in plants, leading to various strategies to remove them prior to phylogenetic
487 analyses. Commonly used target enrichment assembly pipelines (e.g., Faircloth 2016; Johnson et
488 al. 2016; Andermann et al. 2018) use different criteria to flag assembled loci with putative
489 paralogs that are later filtered or sorted prior to phylogenetic analysis. The most common
490 approach for processing paralogs in target enrichment datasets is removing the entire loci that
491 show any signal of potential paralogy (e.g., Crowl et al. 2017; Montes et al. 2019; Bagley et al.
492 2020). Other approaches either retain or remove contigs based on the distinction between
493 putative allelic variation (flagged sequences form monophyletic groups) or putative paralogs
494 (non-monophyletic) in combination with study-specific threshold or random selection (e.g.,
495 Villaverde et al; 2018; Liu et al. 2019; Stubbs et al. 2019), or manual processing (e.g., Garcia et
496 al. 2017; Karimi et al. 2019). Using these approaches, especially the removal of entire loci, can
497 significantly reduce the size of target enrichment datasets. As dataset size increases, manual
498 processing becomes prohibitive.

499 The presence of WGDs also poses some challenges for locus assembly. Target
500 enrichment design commonly includes multi-contig targets that assembly pipelines attempt to

501 assemble into single contigs (e.g. Faircloth 2016) or ‘supercontigs’ composed of multiple exons
502 and partially assembled introns (e.g., Johnson et al. 2016). In groups like *Alchemilla* s.l., where
503 multiple, nested WGD events led to a prevalence of closely-related paralogs, ‘supercontigs’
504 produces chimeric assemblies (Morales-Briones et al. 2018b). Instead, we assembled the exons
505 individually to minimize chimeric loci, at the cost of generating short exons that may produce
506 poorly resolved gene trees, which can affect orthology inference and downstream phylogenetic
507 analyses. Therefore, it is important to take this into consideration during target enrichment
508 experimental design, and to preferentially target long exons when possible. An alternative
509 strategy to avoid chimeric supercontig when gene duplications are frequent is to perform a
510 preliminary orthology inference in single exon-based trees and then use the inferred orthologs as
511 a reference to reassemble the loci into 'supercontigs' (e.g., Gardner et al. 2020; Karimi et al.
512 2020).

513 The utility of paralogs for phylogenetic reconstruction in target enrichment datasets is
514 gaining more attention (e.g., Johnson et al. 2016; Gardner et al. 2020). A few studies have
515 considered tree-based orthology inference to process affected loci (e.g., Garcia et al. 2017;
516 Moore et al. 2018, Morales-Briones et al. 2018b), but in some cases the orthology approaches
517 used cannot be applied to other groups. Here we demonstrated the utility of automated, tree-
518 based orthology inference methods (Yang and Smith 2014), originally designed for genomic or
519 transcriptomic datasets, to infer orthology from paralog-flagged loci in a target enrichment
520 dataset. Our approach facilitates the automated inference of orthologs while maximizing the
521 number of loci retained for downstream analyses. It is suitable for any type of target enrichment
522 dataset (e.g., anchored phylogenomics, exon capture; Hyb-Seq, ultraconserved elements) as long
523 as individual gene trees are informative.

524 Orthology inference methods used here (Yang and Smith et al. 2014) are a powerful tool
525 for target enrichment datasets. In the case of allopolyploidy, however, these methods can
526 introduce bias in the distribution of ortholog trees inferred. In the case of MO, each time a gene
527 duplication event is detected, the side with a smaller number of taxa is removed. When
528 allopolyploidy occurs, MO will likely result in preserving one subgenome instead of the other.
529 One could alternatively preserve a random side each time a gene duplication event is identified.
530 However, in practice, the side with a smaller number of taxa often contains misassembled or
531 misplaced sequences. The RT method of separating duplicated gene copies, on the other hand,
532 keeps any subtree with sufficient number of taxa, but removes outgroups, and worked best when
533 hierarchical outgroups were included in the taxon sampling. Therefore, both MO and RT lose
534 information, especially in cases with complex, nested polyploidy. Recently developed methods
535 based on quartet similarity (Zhang et al. 2019) or Robinson-Foulds distances (Molloy and
536 Warnow 2020) can directly estimate species trees from multi-labeled trees that are consistent
537 with the MSC and gene duplication and loss without inferring orthologs. However, their
538 behavior on complex datasets using archival materials is yet to be explored. For example, both
539 methods do not define ingroup-outgroup relationships a priori, and correctly inferring the root of
540 homolog trees can be challenging with missing data, or when WGD occurs near the root. In
541 addition, none of these above methods were designed to explicitly handle reticulate relationships.
542 This can result in species tree topology that is an “average” between subgenomes. Depending on
543 the topological distance of subgenomes, the resulting species tree may not represent any
544 subgenome history. Finally, most current methods for evaluating node support still require
545 orthologous gene trees as input. In such cases tools like Phyparts can still be used to visualize
546 gene tree discordance and calculate ICA scores using multi-labeled trees.

547

548 *Phylogenetic implications in Alchemilla s.l.*

549 Previous phylogenetic studies established the monophyly *Alchemilla* s.l. and four major clades of
550 the group (Gehrke et al. 2008; 2016), but the relationship among them and the placement of
551 *Alchemilla* s.l. within Fragariinae remain unresolved. Our nuclear and plastid analyses both
552 confirmed the monophyly of *Alchemilla* s.l. and its sister relationship to *Farinopsis*, as
553 previously shown by Morales-Briones and Tank (2019) based on plastome sequences only.
554 Gehrke et al. (2008) identified two well supported clades within Eualchemilla that were
555 distinguished by leaf shape, namely the 'dissected' and 'lobed' clades. Most species of
556 Eualchemilla have a leaf shape consistent with their clade placement, but some had different leaf
557 shapes that were attributed to their hybrid origin between the two clades (Gehrke et al. 2008).
558 More recently, Gehrke et al. (2016) and Morales-Briones and Tank (2019) found that
559 Eualchemilla is not monophyletic in analyses that included the external transcribed spacer (ETS)
560 of the nuclear ribosomal DNA (nrDNA) cistron. Both studies found *Aphanes* nested between the
561 'dissected' and 'lobed' clades of Eualchemilla. Our analyses of the nuclear loci supported the
562 monophyly of 'dissected' and 'lobed' clades, but the monophyly of Eualchemilla had low support
563 (Fig. 2; Fig. S1). The analysis using only the MO orthologs even weakly supported the 'lobed'
564 clade as sister of *Aphanes* (Fig. 2). In contrast, our plastome analysis recovered a well-supported,
565 monophyletic Eualchemilla, as well as well-supported 'dissected' and 'lobed' clades. Both nuclear
566 and plastid analyses strongly supported the clade composed of *Aphanes* and both clades of
567 Eualchemilla (Fig. 2; Fig. S4), a relationship that is consistent with previous nuclear and plastid
568 analyses (Gehrke et al. 2008; 2016; Morales-Briones and Tank 2019). Given the revealed
569 hybridization in the evolution and early divergence within *Alchemilla* s.l., the non-monophyly of

570 Eualchemilla could be explained by ancient gene flow or an allopolyploid origin of the
571 'dissected' and 'lobed' clades (Fig. 4a,c; see below). Besides the well supported relationship of
572 Eualchemilla + *Aphanes*, our nuclear analysis showed high levels of conflict among the other
573 clades (Fig. 2; Fig. S1) which could also be explained by additional ancient allopolyploid events
574 within *Alchemilla* s.l. (Fig. 4; see below).
575

576 *Ancient allopolyploidy in Alchemilla s.l.*

577 Whole-genome duplications are frequent across Rosaceae (Dickinson et al. 2007; Xiang et al.
578 2017), and allopolyploidy has been suggested as the primary source for the cytonuclear
579 discordance in Fragariinae (Lundberg et al. 2009; Gehrke et al. 2016; Morales-Briones and Tank
580 2019). We recovered four nodes in *Alchemilla* s.l. with a high percentage of gene duplications
581 (Fig. 3a; Fig. S5). One of the nodes showing a high percentage of gene duplication (18.4%) was
582 the MRCA of *Aphanes* and Eualchemilla (node 3 in Fig. 3a; Fig. S5). This duplication event
583 agreed with the MRCA of the ancestral lineages inferred with GRAMPA for the allopolyploid
584 origin of the 'lobed' clade of Eualchemilla (Fig. 4a). Moreover, the GRAMPA reconciliations
585 after the removal of the 'lobed' clade and Afromilla (Fig. 4c) inferred a scenario where the
586 'dissected' clade is of allopolyploid origin. Ks plots of all species of the 'dissected' and 'lobed'
587 clades shared two peaks that are missing from taxa outside of *Alchemilla* s.l. (Fig. S10)
588 suggested that at least two WGD events have happened between the stem lineage of *Alchemilla*
589 s.l. to the crown node of the dissected and the lobed clades. Given the elevated ploidy level of
590 the sampled species ($2n$=102–156) and that the sister relationship of the 'dissected' and 'lobed'
591 clades are not supported by nuclear genes, the two clades of Eualchemilla seem to have had

592 independent allopolyploid origins, while sharing the same or a closely related maternal lineage
593 (cpDNA; Fig. 4f).

594 The next node with high percentage of gene duplication (34.4%) was placed at the
595 MRCA of the clade composed of Afromilla, Eualchemilla, and *Aphanes* (node 2 in Fig. 3a). The
596 GRAMPA reconciliation (after the removal of the 'lobed' clades) recovered an allopolyploid
597 origin of Afromilla (Fig. 4b), with the MRCA of inferred parental lineages being the base of
598 *Alchemilla* s.l., consistent with the high percentage of gene duplication mapped to that node
599 (86.0% at node 1; Fig. 3a). After removing both Afromilla and Eualchemilla, *Lachemilla* was
600 recovered as allopolyploid, again with the MRCA of parental lineages mapped to the base of
601 *Alchemilla* s.l. (Fig. 4d)

602 Finally, the node with the highest percentage of duplicated genes (86%) was placed at the
603 MRCA of *Alchemilla* s.l. (node 1 in Fig. 3a). GRAMPA analysis using the MO species tree
604 showed an allopolyploid event for the clade that included *Alchemilla* s.l., *Farinopsis salesoviana*,
605 *Sibbaldia*, and *Sibbaldianthe* (Fig. S7). However, an allopolyploid origin of *Farinopsis*
606 *salesoviana*, *Sibbaldia*, and *Sibbaldianthe* is not supported by chromosome numbers, orthogroup
607 gene duplication counts, or Ks plots. All members in Fragariinae, with the exception of
608 *Alchemilla* s.l. mainly consists of diploid species and base chromosome number of seven ($n$ =7),
609 including *Sibbaldia* and *Sibbaldianthe.* On the other hand, *Alchemilla* s.l. has a base number of
610 eight ($n$ =8), and contains mostly species with high ploidy levels (octoploid to 24-ploid), with the
611 exception of most species of *Aphanes* (2$n$=16) and one species of *Lachemilla* (*L. mandoniana,*
612 2$n$=16). Also, our duplication counts show low percentages (1.6%) of gene duplication for the
613 MRCA of the GRAMPA-inferred allopolyploid clade or the MRCA (3.6%) of the inferred
614 parental lineages (Fig. 3). Furthermore, previous phylotranscriptomic analyses of Rosaceae

615 (Xiang et al. 2017) that included one species each of the ‘dissected’ and 'lobed' clades of
616 Eualchemilla, found 33.21% of duplicated genes for the MRCA of these two clades, but did not
617 recover any other node with elevated gene duplications within Fragariinae. The Ks plots of the
618 four species of *Alchemilla* s.l. all showed peaks with similar Ks means among them, but these
619 largely differed from Ks plots for diploid and polyploid species of *Sibbaldia* and *Sibadianthe*,
620 and do not share any peak. Instead, *Sibbaldia* and *Sibadianthe* showed Ks patterns similar to
621 other genera across Fragariinae (Fig. S9). Although the chromosome number and Ks data for
622 *Farinopsis salesoviana* is not available, all these suggest an unlikely allopolyploid origin of
623 *Farinopsis, Sibbaldia*, and *Sibadianthe* along with *Alchemilla* sl. In turn, the GRAMPA
624 reconciliations using the cpDNA tree resulted in an optimal multi-labeled tree where *Alchemilla.*
625 s.l. had an autopolyploid origin (Fig. 4e). This scenario is compatible with the high percentage of
626 gene duplication at the MRCA of *Alchemilla* s.l. and the low percentage of gene duplication in
627 the rest of Fragariinae. Another compatible scenario is an allopolyploid origin of *Alchemilla* s.l.
628 where both parental lineages are missing or extinct, but this scenario is indistinguishable from
629 autopolyploidy. The atypical high proportion of gene duplication in the crown of *Alchemilla* s.l.
630 can be explained by the inferred autopolyploid event. However, this high proportion of gene
631 duplication could also be the results of gene tree estimation error (e.g., uninformative genes),
632 incomplete lineage sorting (ILS), allopolyploid events among major clades of *Alchemilla* s.l.,
633 and/or homoeologous exchanges among subgenomes (Edger et al. 2018; McKain et al. 2018).
634 Altogether, our results showed evidence that the entire *Alchemilla* s.l. is the product of an ancient
635 autopolyploidy event, and that Afromilla, Eualchemilla (‘lobed’ and ‘dissected’ clades), and
636 *Lachemilla* originated from subsequent and nested ancient allopolyploid events (Fig. 4f).

While GRAMPA has been shown to be useful to identify multiple polyploidy events in the same tree (e.g., Thomas et al. 2017; Koenen et al. 2020), this tree-based approach can also be sensitive to gene tree estimation error or ILS (Thomas et al. 2017). Methods to infer species networks in the presence of ILS (e.g., Solís-Lemus and Ané 2016; Wen et al. 2018) could also be used to explore the prevalence of ancient hybridization in *Alchemilla* s.l. Although these methods are under constant development and improvement, they are still only tractable in simple scenarios with few reticulation events (Hejase and Liu 2016; Kamneva and Rosenberg 2017). Similarly, the signal of the *D*-Statistic (Green et al. 2010; Durand et al. 2011), commonly used to detect introgression, can be lost or distorted in presence of multiple reticulations (Elworth et al. 2018). Complex reticulate scenarios like *Alchemilla* s.l. are likely to face these problems and have phylogenetic network and *D*-statistic identifiability issues as seen in other groups (e.g., Morales-Briones et al. 2020).

## *Conclusions*

In this study, we have shown the potential of target enrichment datasets in combination with tree-based methods for paralog processing and WGD investigation. We have used *Alchemilla* s.l., a group showing nested WGDs, to highlight the potential of these methods in investigating cases of complex scenarios. Our study shows the need in selecting long exons that produce informative gene trees for target enrichment studies. As research continues to deepen in other clades across the Tree of Life using similar methods, we expect that complex patterns of duplications and reticulations, as those shown here in *Alchemilla* s.l. are more common than we previously thought.

660 **SUPPLEMENTARY MATERIAL**

661 Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/.[NNNN]

662

663 **ACKNOWLEDGMENTS**

664 This work was supported by the University of Minnesota, Graduate Student Research Grants
665 from the Botanical Society of America, American Society of Plant Taxonomists, International
666 Association of Plant Taxonomists, the University of Idaho Stillinger Herbarium Expedition
667 Funds to D.F.M.-B., and a National Science Foundation Doctoral Dissertation Improvement
668 Grant to D.C.T. for D.F.M.-B. (DEB-1502049).

669

670 **REFERENCES**

671 Andermann T., Cano Á., Zizka A., Bacon C., Antonelli A. 2018. SECAPR—a bioinformatics
672 pipeline for the rapid and user-friendly processing of targeted enriched Illumina
673 sequences, from raw reads to alignments. PeerJ. 6:e5175.

674 Andermann T., Torres Jiménez M.F., Matos-Maraví P., Batista R., Blanco-Pastor J.L.,
675 Gustafsson A.L.S., Kistler L., Liberal I.M., Oxelman B., Bacon C.D., Antonelli A. 2020.
676 A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. Front.
677 Genet. 10:1407.

678 Bagley J.C., Uribe-Convers S., Carlsen M.M., Muchhala N. 2020. Utility of targeted sequence
679 capture for phylogenomics in rapid, recent angiosperm radiations: Neotropical
680 Burmeistera bellflowers as a case study. Mol. Phylogenet. Evol.:106769.

681 Benaglia T., Chauveau D., Hunter D.R., Young D. 2009. mixtools : An R Package for Analyzing
682 Finite Mixture Models. J. Stat. Softw. 32:1–29.

683 Brown J.W., Walker J.F., Smith S.A. 2017. Phyx - phylogenetic tools for unix. Bioinformatics.
684 33:1886–1888.
685 Buddenhagen C., Lemmon A.R., Lemmon E.M., Bruhl J., Cappa J., Clement W.L., Donoghue
686 M.J., Edwards E.J., Hipp A.L., Kortyna M., Mitchell N., Moore A., Prychid C.J.,
687 Segovia-Salcedo M.C., Simmons M.P., Soltis P.S., Wanke S., Mast A. 2016. Anchored
688 Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates.
689 bioRxiv.:086298.
690 Chamala S., García N., Godden G.T., Krishnakumar V., Jordon-Thaden I.E., De Smet R.,
691 Barbazuk W.B., Soltis D.E., Soltis P.S. 2015. MarkerMiner 1.0: A new application for
692 phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci.
693 3:1400115.
694 Crowl A.A., Manos P.S., McVay J.D., Lemmon A.R., Lemmon E.M., Hipp A.L. 2019.
695 Uncovering the genomic signature of ancient introgression between white oak lineages
696 (Quercus). New Phytol.:nph.15842.
697 Dickinson T.A., Lo E., Talent N. 2007. Polyploidy, reproductive biology, and Rosaceae:
698 understanding evolution and making classifications. Plant Syst. Evol. 266:59–78.
699 Doyle J.J., Doyle J.L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf
700 tissue. Phytochem. Bull. 19:11–15.
701 Dunn C.W., Howison M., Zapata F. 2013. Agalma: an automated phylogenomics workflow.
702 BMC Bioinformatics. 14:330.
703 Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between
704 Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

705 Edger P.P., McKain M.R., Bird K.A., VanBuren R. 2018. Subgenome assignment in
706 allopolyploids: challenges and future directions. Curr. Opin. Plant Biol. 42:76–80.
707 Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L.K. 2018. DGEN: A Test Statistic
708 for Detection of General Introgression Scenarios. WABI.
709 Emms D.M., Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
710 genomics. Genome Biol. 20:238.
711 Eriksson T., Lundberg M., Töpel M., Östensson P., Smedmark J.E.E. 2015. Sibbaldia: a
712 molecular phylogenetic study of a remarkably polyphyletic genus in Rosaceae. Plant
713 Syst. Evol. 301:171–184.
714 Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic
715 loci. Bioinforma. Oxf. Engl. 32:786–788.
716 Fernández R., Gabaldon T., Dessimoz C. 2020. Orthology: Definitions, Prediction, and Impact
717 on Species Phylogeny Inference. In: Scornavacca C., Delsuc F., Galtier N., editors.
718 Phylogenetics in the Genomic Era. No commercial publisher | Authors open access
719 book. p. 2.4:1--2.4:14.
720 Fitch W.M. 1970. Distinguishing Homologous from Analogous Proteins. Syst. Biol. 19:99–113.
721 Forrest L.L., Hart M.L., Hughes M., Wilson H.P., Chung K.-F., Tseng Y.-H., Kidner C.A. 2019.
722 The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques.
723 Front. Ecol. Evol. 7:439.
724 García N., Folk R.A., Meerow A.W., Chamala S., Gitzendanner M.A., Oliveira R.S. de, Soltis
725 D.E., Soltis P.S. 2017. Deep reticulation and incomplete lineage sorting obscure the
726 diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). Mol.
727 Phylogenet. Evol. 111:231–247.

728 Gardner E.M., Johnson M.G., Pereira J.T., Ahmad Puad A.S., Arifiani D., Sahromi, Wickett
729 N.J., Zerega N.J.C. 2020. Paralogs and off-target sequences improve phylogenetic
730 resolution in a densely-sampled study of the breadfruit genus (Artocarpus, Moraceae).
731 bioRxiv.:854232.
732 Gardner E.M., Johnson M.G., Ragone D., Wickett N.J., Zerega N.J.C. 2016. Low-coverage,
733 whole-genome sequencing of Artocarpus camansi (Moraceae) for phylogenetic marker
734 development and gene discovery. Appl. Plant Sci. 4:1600017.
735 Gehrke B., Bräuchler C., Romoleroux K., Lundberg M., Heubl G., Eriksson T. 2008. Molecular
736 phylogenetics of Alchemilla, Aphanes and Lachemilla (Rosaceae) inferred from plastid
737 and nuclear intron and spacer DNA sequences, with comments on generic classification.
738 Mol. Phylogenet. Evol. 47:1030–1044.
739 Gehrke B., Kandziora M., Pirie M.D. 2016. The evolution of dwarf shrubs in alpine
740 environments: a case study of Alchemilla in Africa. Ann. Bot. 117:121–131.
741 Glover N., Dessimoz C., Ebersberger I., Forslund S.K., Gabaldón T., Huerta-Cepas J., Martin
742 M.-J., Muffato M., Patricio M., Pereira C., da Silva A.S., Wang Y., Sonnhammer E.,
743 Thomas P.D. 2019. Advances and Applications in the Quest for Orthologs. Mol. Biol.
744 Evol. 36:2157–2164.
745 Gonçalves D.J.P., Simpson B.B., Ortiz E.M., Shimizu G.H., Jansen R.K. 2019. Incongruence
746 between gene trees and species trees and phylogenetic signal variation in plastid genes.
747 Mol. Phylogenet. Evol. 138:219–232.
748 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
749 W., Fritz M.H.Y., Hansen N.F., Durand E.Y., Malaspinas A.S., Jensen J.D., Marques-
750 Bonet T., Alkan C., Prufer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-

751 Petri A., Butthof A., Hober B., Hoffner B., Siegemund M., Weihmann A., Nusbaum C.,
752 Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic
753 D., Kucan Z., Gusic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., de la
754 Rasilla M., Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D.,
755 Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D.,
756 Paabo S. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
757 Hejase H.A., Liu K.J. 2016. A scalability study of phylogenetic network inference methods using
758 empirical datasets and simulations involving a single reticulation. BMC Bioinformatics.
759 17:422.
760 Huang C.-H., Zhang C., Liu M., Hu Y., Gao T., Qi J., Ma H. 2016. Multiple Polyploidization
761 Events across Asteraceae with Two Nested Events in the Early History Revealed by
762 Nuclear Phylogenomics. Mol. Biol. Evol. 33:2820–2835.
763 Jiao Y., Leebens-Mack J., Ayyampalayam S., Bowers J.E., McKain M.R., McNeal J., Rolf M.,
764 Ruzicka D.R., Wafula E., Wickett N.J., Wu X., Zhang Y., Wang J., Zhang Y., Carpenter
765 E.J., Deyholos M.K., Kutchan T.M., Chanderbali A.S., Soltis P.S., Stevenson D.W.,
766 McCombie R., Pires J., Wong G., Soltis D.E., dePamphilis C.W. 2012. A genome
767 triplication associated with early diversification of the core eudicots. Genome Biol.
768 13:R3.
769 Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho
770 L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster
771 S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed
772 plants and angiosperms. Nature. 473:97–100.

773 Johnson M.G., Gardner E.M., Liu Y., Medina R., Goffinet B., Shaw A.J., Zerega N.J.C., Wickett
774 N.J. 2016. HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from
775 High-Throughput Sequencing Reads Using Target Enrichment. Appl. Plant Sci.
776 4:1600016.
777 Jones K.E., Fér T., Schmickl R.E., Dikow R.B., Funk V.A., Herrando-Moraira S., Johnston P.R.,
778 Kilian N., Siniscalchi C.M., Susanna A., Slovák M., Thapa R., Watson L.E., Mandel
779 J.R. 2019. An empirical assessment of a single family-wide hybrid capture locus set at
780 multiple evolutionary timescales in Asteraceae. Appl. Plant Sci. 7:e11295.
781 Kamneva O.K., Rosenberg N.A. 2017. Simulation-Based Evaluation of Hybridization Network
782 Reconstruction Methods in the Presence of Incomplete Lineage Sorting. Evol.
783 Bioinforma. 13:117693431769193.
784 Kamneva O.K., Syring J., Liston A., Rosenberg N.A. 2017. Evaluating allopolyploid origins in
785 strawberries (Fragaria) using haplotypes generated from target capture sequencing.
786 BMC Evol. Biol. 17:401.
787 Karimi N., Grover C.E., Gallagher J.P., Wendel J.F., Ané C., Baum D.A. 2020. Reticulate
788 Evolution Helps Explain Apparent Homoplasy in Floral Biology and Pollination in
789 Baobabs (Adansonia; Bombacoideae; Malvaceae). Syst. Biol. 69:462–478.
790 Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
791 Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.
792 Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A.,
793 Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. 2012.
794 Geneious Basic: An integrated and extendable desktop software platform for the
795 organization and analysis of sequence data. Bioinformatics. 28:1647–1649.

Kocot K.M., Citarella M.R., Moroz L.L., Halanych K.M. 2013. PhyloTreePruner: A Phylogenetic Tree-Based Approach for Selection of Orthologous Sequences for Phylogenomics. Evol. Bioinforma. Online. 9:429–435.

Koenen E.J.M., Ojeda D.I., Bakker F.T., Wieringa J.J., Kidner C., Hardy O.J., Pennington R.T., Herendeen P.S., Bruneau A., Hughes C.E. 2020. The Origin of the Legumes is a Complex Paleopolyploid Phylogenomic Tangle closely associated with the Cretaceous-Paleogene (K-Pg) Mass Extinction Event. Syst. Biol. syaa041.

Larridon I., Villaverde T., Zuntini A.R., Pokorny L., Brewer G.E., Epitawalage N., Fairlie I., Hahn M., Kim J., Maguilla E., Maurin O., Xanthos M., Hipp A.L., Forest F., Baker W.J. 2020. Tackling Rapid Radiations With Targeted Sequencing. Front. Plant Sci. 10:1655.

Leebens-Mack J.H., Barker M.S., Carpenter E.J., Deyholos M.K., Gitzendanner M.A., Graham S.W., Grosse I., Li Z., Melkonian M., Mirarab S., Porsch M., Quint M., Rensing S.A., Soltis D.E., Soltis P.S., Stevenson D.W., Ullrich K.K., Wickett N.J., DeGironimo L., Edger P.P., Jordon-Thaden I.E., Joya S., Liu T., Melkonian B., Miles N.W., Pokorny L., Quigley C., Thomas P., Villarreal J.C., Augustin M.M., Barrett M.D., Baucom R.S., Beerling D.J., Benstein R.M., Biffin E., Brockington S.F., Burge D.O., Burris J.N., Burris K.P., Burtet-Sarramegna V., Caicedo A.L., Cannon S.B., Çebi Z., Chang Y., Chater C., Cheeseman J.M., Chen T., Clarke N.D., Clayton H., Covshoff S., Crandall-Stotler B.J., Cross H., dePamphilis C.W., Der J.P., Determann R., Dickson R.C., Di Stilio V.S., Ellis S., Fast E., Feja N., Field K.J., Filatov D.A., Finnegan P.M., Floyd S.K., Fogliani B., García N., Gâteblé G., Godden G.T., Goh F. (Qi Y., Greiner S., Harkess A., Heaney J.M., Helliwell K.E., Heyduk K., Hibberd J.M., Hodel R.G.J., Hollingsworth P.M., Johnson M.T.J., Jost R., Joyce B., Kapralov M.V., Kazamia E.,

819 Kellogg E.A., Koch M.A., Von Konrat M., Könyves K., Kutchan T.M., Lam V., Larsson
820 A., Leitch A.R., Lentz R., Li F.-W., Lowe A.J., Ludwig M., Manos P.S., Mavrodiev E.,
821 McCormick M.K., McKain M., McLellan T., McNeal J.R., Miller R.E., Nelson M.N.,
822 Peng Y., Ralph P., Real D., Riggins C.W., Ruhsam M., Sage R.F., Sakai A.K.,
823 Scascitella M., Schilling E.E., Schlösser E.-M., Sederoff H., Servick S., Sessa E.B.,
824 Shaw A.J., Shaw S.W., Sigel E.M., Skema C., Smith A.G., Smithson A., Stewart C.N.,
825 Stinchcombe J.R., Szövényi P., Tate J.A., Tiebel H., Trapnell D., Villegente M., Wang
826 C.-N., Weller S.G., Wenzel M., Weststrand S., Westwood J.H., Whigham D.F., Wu S.,
827 Wulff A.S., Yang Y., Zhu D., Zhuang C., Zuidof J., Chase M.W., Pires J.C., Rothfels
828 C.J., Yu J., Chen C., Chen L., Cheng S., Li J., Li R., Li X., Lu H., Ou Y., Sun X., Tan
829 X., Tang J., Tian Z., Wang F., Wang J., Wei X., Xu X., Yan Z., Yang F., Zhong X.,
830 Zhou F., Zhu Y., Zhang Y., Ayyampalayam S., Barkman T.J., Nguyen N., Matasci N.,
831 Nelson D.R., Sayyari E., Wafula E.K., Walls R.L., Warnow T., An H., Arrigo N.,
832 Baniaga A.E., Galuska S., Jorgensen S.A., Kidder T.I., Kong H., Lu-Irving P., Marx
833 H.E., Qi X., Reardon C.R., Sutherland B.L., Tiley G.P., Welles S.R., Yu R., Zhan S.,
834 Gramzow L., Theißen G., Wong G.K.-S., One Thousand Plant Transcriptomes Initiative.
835 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature.
836 574:679–685.

837 Li L. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome
838 Res. 13:2178–2189.

839 Li Z., Baniaga A.E., Sessa E.B., Scascitelli M., Graham S.W., Rieseberg L.H., Barker M.S.
840 2015. Early genome duplications in conifers and other seed plants. Sci. Adv.
841 1:e1501084.

842 Liu Y., Johnson M.G., Cox C.J., Medina R., Devos N., Vanderpoorten A., Hedenäs L., Bell N.E.,
843 Shevock J.R., Aguero B., Quandt D., Wickett N.J., Shaw A.J., Goffinet B. 2019.
844 Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and
845 nuclear genomes. Nat. Commun. 10:1485.
846 Lundberg M., Töpel M., Eriksen B., Nylander J.A.A., Eriksson T. 2009. Allopolyploidy in
847 Fragariinae (Rosaceae): Comparing four DNA sequence regions, with comments on
848 classification. Mol. Phylogenet. Evol. 51:269–280.
849 Lynch M., Conery J.S. 2000. The Evolutionary Fate and Consequences of Duplicate Genes.
850 Science. 290:1151–1155.
851 Mai U., Mirarab S. 2018. TreeShrink: fast and accurate detection of outlier long branches in
852 collections of phylogenetic trees. BMC Genomics. 19:272.
853 Mandel J.R., Dikow R.B., Funk V.A., Masalia R.R., Staton S.E., Kozik A., Michelmore R.W.,
854 Rieseberg L.H., Burke J.M. 2014. A Target Enrichment Method for Gathering
855 Phylogenetic Information from Hundreds of Loci: An Example from the Compositae.
856 Appl. Plant Sci. 2:1300085.
857 McKain M.R., Estep M.C., Pasquet R., Layton D.J., Vela Díaz D.M., Zhong J., Hodge J.G.,
858 Malcomber S.T., Chipabika G., Pallangyo B., Kellogg E.A. 2018. Ancestry of the two
859 subgenomes of maize. bioRxiv.:352351.
860 McKain M.R., Tang H., McNeal J.R., Ayyampalayam S., Davis J.I., dePamphilis C.W., Givnish
861 T.J., Pires J.C., Stevenson D.W., Leebens-Mack J.H. 2016. A Phylogenomic Assessment
862 of Ancient Polyploidy and Genome Evolution across the Poales. Genome Biol. Evol.
863 8:1150–1164.
864 McLachlan G., Peel D. 2000. Finite Mixture Models. New York: Wiley.

865 Molloy E.K., Warnow T. 2020. FastMulRFS: fast and accurate species tree estimation under
866 generic gene duplication and loss models. Bioinformatics. 36:i57–i65.
867 Montes J.R., Peláez P., Willyard A., Moreno-Letelier A., Piñero D., Gernandt D.S. 2019.
868 Phylogenetics of Pinus Subsection Cembroides Engelm. (Pinaceae) Inferred from Low-
869 Copy Nuclear Gene Sequences. Syst. Bot. 44:501–518.
870 Moore A.J., Vos J.M.D., Hancock L.P., Goolsby E., Edwards E.J. 2018. Targeted Enrichment of
871 Large Gene Families for Phylogenetic Inference: Phylogeny and Molecular Evolution of
872 Photosynthesis Genes in the Portullugo Clade (Caryophyllales). Syst. Biol. 67:367–383.
873 Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F.,
874 Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2020. Disentangling Sources of Gene
875 Tree Discordance in Phylogenomic Datasets: Testing Ancient Hybridizations in
876 Amaranthaceae s.l. Syst. Biol. syaa066
877 Morales-Briones D.F., Romoleroux K., Kolář F., Tank D.C. 2018a. Phylogeny and Evolution of
878 the Neotropical Radiation of *Lachemilla* (Rosaceae): Uncovering a History of Reticulate
879 Evolution and Implications for Infrageneric Classification. Syst. Bot. 43:17–34.
880 Morales-Briones D.F., Liston A., Tank D.C. 2018b. Phylogenomic analyses reveal a deep history
881 of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New
882 Phytol. 218:1668–1684.
883 Morales-Briones D.F., Tank D.C. 2019. Extensive allopolyploidy in the neotropical genus
884 *Lachemilla* (Rosaceae) revealed by. Am. J. Bot. 106:415–437.
885 Nicholls J.A., Pennington R.T., Koenen E.J.M., Hughes C.E., Hearn J., Bunnefeld L., Dexter
886 K.G., Stone G.N., Kidner C.A. 2015. Using targeted enrichment of nuclear genes to

887 increase phylogenetic resolution in the neotropical rain forest genus Inga (Leguminosae:
888 Mimosoideae). Front. Plant Sci. 6:710.
889 Pease J.B., Brown J.W., Walker J.F., Hinchliff C.E., Smith S.A. 2018. Quartet Sampling
890 distinguishes lack of support from conflicting support in the green plant tree of life. Am.
891 J. Bot. 105:385–403.
892 Perry L.M. 1929. A Tentative Revision of Alchemilla § Lachemilla. Contrib. Gray Herb. Harv.
893 Univ.:1–57.
894 Ranwez V., Douzery E.J.P., Cambon C., Chantret N., Delsuc F. 2018. MACSE v2: Toolkit for
895 the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons. Mol.
896 Biol. Evol. 35:2582–2584.
897 Salichos L., Stamatakis A., Rokas A. 2014. Novel Information Theory-Based Measures for
898 Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31:1261–1271.
899 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
900 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
901 Shulaev V., Sargent D.J., Crowhurst R.N., Mockler T.C., Folkerts O., Delcher A.L., Jaiswal P.,
902 Mockaitis K., Liston A., Mane S.P., Burns P., Davis T.M., Slovin J.P., Bassil N.,
903 Hellens R.P., Evans C., Harkins T., Kodira C., Desany B., Crasta O.R., Jensen R.V.,
904 Allan A.C., Michael T.P., Setubal J.C., Celton J.-M., Rees D.J.G., Williams K.P., Holt
905 S.H., Rojas J.J.R., Chatterjee M., Liu B., Silva H., Meisel L., Adato A., Filichkin S.A.,
906 Troggio M., Viola R., Ashman T.-L., Wang H., Dharmawardhana P., Elser J., Raja R.,
907 Priest H.D., Bryant D.W., Fox S.E., Givan S.A., Wilhelm L.J., Naithani S., Christoffels
908 A., Salama D.Y., Carter J., Girona E.L., Zdepski A., Wang W., Kerstetter R.A., Schwab
909 W., Korban S.S., Davik J., Monfort A., Denoyes-Rothan B., Arus P., Mittler R., Flinn

910 B., Aharoni A., Bennetzen J.L., Salzberg S.L., Dickerman A.W., Velasco R.,
911 Borodovsky M., Veilleux R.E., Folta K.M. 2011. The genome of woodland strawberry
912 (Fragaria vesca). Nat. Genet. 43:109–116.
913 Smith S.A., Moore M.J., Brown J.W., Yang Y. 2015. Analysis of phylogenomic datasets reveals
914 conflict, concordance, and gene duplications with examples from animals and plants.
915 BMC Evol. Biol. 15:150.
916 Soják J. 2008. Notes on Potentilla XXI. A new division of the tribe Potentilleae (Rosaceae) and
917 notes on generic delimitations. Bot. Jahrb. Für Syst. Pflanzengesch. Pflanzengeogr.
918 127:349–358.
919 Solís-Lemus C., Ané C. 2016. Inferring Phylogenetic Networks with Maximum
920 Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet. 12:e1005896–21.
921 Stamatakis A. 2014. RAxML version 8 - a tool for phylogenetic analysis and post-analysis of
922 large phylogenies. Bioinformatics. 30:1312–1313.
923 Straub S.C., Fishbein M., Livshultz T., Foster Z., Parks M., Weitemier K., Cronn R.C., Liston A.
924 2011. Building a model: developing genomic resources for common milkweed
925 (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics. 12:211.
926 Stubbs R.L., Folk R.A., Xiang C.-L., Chen S., Soltis D.E., Cellinese N. 2020. A Phylogenomic
927 Perspective on Evolution and Discordance in the Alpine-Arctic Plant Clade Micranthes
928 (Saxifragaceae). Front. Plant Sci. 10:1773.
929 Thomas G.W.C., Ather S.H., Hahn M.W. 2017. Gene-Tree Reconciliation with MUL-Trees to
930 Resolve Polyploidy Events. Syst. Biol. 66:1007–1018.
931 Villaverde T., Pokorny L., Olsson S., Rincón-Barrado M., Johnson M.G., Gardner E.M., Wickett
932 N.J., Molero J., Riina R., Sanmartín I. 2018. Bridging the micro- and macroevolutionary

933 levels in phylogenomics: Hyb-Seq solves relationships from populations to species and
934 above. New Phytol. 220:636–650.
935 Viruel J., Conejero M., Hidalgo O., Pokorny L., Powell R.F., Forest F., Kantar M.B., Soto
936 Gomez M., Graham S.W., Gravendeel B., Wilkin P., Leitch I.J. 2019. A Target Capture-
937 Based Method to Estimate Ploidy From Herbarium Specimens. Front. Plant Sci. 10:937.
938 Walker J.F., Walker-Hale N., Vargas O.M., Larson D.A., Stull G.W. 2019. Characterizing gene
939 tree conflict in plastome-inferred phylogenies. PeerJ. 7:e7747.
940 Weitemier K., Straub S.C.K., Cronn R.C., Fishbein M., Schmickl R., McDonnell A., Liston A.
941 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant
942 Phylogenomics. Appl. Plant Sci. 2:1400042.
943 Wen D., Yu Y., Zhu J., Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet.
944 Syst. Biol. 67:735–740.
945 Xiang Y., Huang C.-H., Hu Y., Wen J., Li S., Yi T., Chen H., Xiang J., Ma H. 2017. Evolution
946 of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological
947 Times and Genome Duplication. Mol. Biol. Evol. 34:262–281.
948 Yang Y., Moore M.J., Brockington S.F., Mikenas J., Olivieri J., Walker J.F., Smith S.A. 2018.
949 Improved transcriptome sampling pinpoints 26 ancient and more recent polyploidy
950 events in Caryophyllales, including two allopolyploidy events. New Phytol. 217:855–
951 870.
952 Yang Y., Moore M.J., Brockington S.F., Soltis D.E., Wong G.K.-S., Carpenter E.J., Zhang Y.,
953 Chen L., Yan Z., Xie Y., Sage R.F., Covshoff S., Hibberd J.M., Nelson M.N., Smith
954 S.A. 2015. Dissecting Molecular Evolution in the Highly Diverse Plant Clade
955 Caryophyllales Using Transcriptome Sequencing. Mol. Biol. Evol. 32:2001–2014.

956 Yang Y., Smith S.A. 2014. Orthology Inference in Nonmodel Organisms Using Transcriptomes
957 and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for
958 Phylogenomics. Mol. Biol. Evol. 31:3081–3092.
959 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
960 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.
961 Zhang C., Scornavacca C., Molloy E.K., Mirarab S. 2019. ASTRAL-Pro: quartet-based species
962 tree inference despite paralogy. bioRxiv.:2019.12.12.874727.
963 Zhang R., Wang Y.-H., Jin J.-J., Stull G.W., Bruneau A., Cardoso D., De Queiroz L.P., Moore
964 M.J., Zhang S.-D., Chen S.-Y., Wang J., Li D.-Z., Yi T.-S. 2020. Exploration of Plastid
965 Phylogenomic Conflict Yields New Insights into the Deep Relationships of
966 Leguminosae. Syst. Biol. 69:613–622.
967 Zhbannikov I.Y., Hunter S.S., Foster J.A., Settles M.L. 2017. SeqyClean: A Pipeline for High-
968 throughput Sequence Data Preprocessing. Proc. 8th ACM Int. Conf. Bioinforma.
969 Comput. Biol. Health Inform.:407–416.
970